



*applied sciences*

# Machine Learning Methods with Noisy, Incomplete or Small Datasets

---

Edited by

Jordi Solé-Casals, Zhe Sun, Cesar F. Caiafa,  
Pere Marti-Puig and Toshihisa Tanaka

Printed Edition of the Special Issue Published in *Applied Sciences*

# **Machine Learning Methods with Noisy, Incomplete or Small Datasets**



# Machine Learning Methods with Noisy, Incomplete or Small Datasets

Editors

**Jordi Solé-Casals**

**Zhe Sun**

**Cesar F. Caiafa**

**Pere Marti-Puig**

**Toshihisa Tanaka**

MDPI • Basel • Beijing • Wuhan • Barcelona • Belgrade • Manchester • Tokyo • Cluj • Tianjin





*Editors*

Jordi Solé-Casals  
Department of Engineering  
University of Vic - Central  
University of Catalonia  
Vic  
Spain

Zhe Sun  
RIKEN National Science  
Institute  
RIKEN  
Wako  
Japan

Cesar F. Caiafa  
Instituto Argentino de  
Radioastronomía  
CONICET  
Villa Elisa  
Argentina

Pere Marti-Puig  
Department of Engineering  
University of Vic - Central  
University of Catalonia  
Vic  
Spain

Toshihisa Tanaka  
Department of Electrical and  
Electronic Engineering  
Tokyo University of Agriculture  
and Technology  
Tokyo  
Japan

*Editorial Office*

MDPI  
St. Alban-Anlage 66  
4052 Basel, Switzerland

This is a reprint of articles from the Special Issue published online in the open access journal *Applied Sciences* (ISSN 2076-3417) (available at: [www.mdpi.com/journal/applsci/special\\_issues/machine\\_learning\\_noisy\\_incomplete\\_datasets](http://www.mdpi.com/journal/applsci/special_issues/machine_learning_noisy_incomplete_datasets)).

For citation purposes, cite each article independently as indicated on the article page online and as indicated below:

LastName, A.A.; LastName, B.B.; LastName, C.C. Article Title. <i>Journal Name</i> <b>Year</b> , Volume Number, Page Range.
--

**ISBN 978-3-0365-1288-4 (Hbk)**

**ISBN 978-3-0365-1287-7 (PDF)**

© 2021 by the authors. Articles in this book are Open Access and distributed under the Creative Commons Attribution (CC BY) license, which allows users to download, copy and build upon published articles, as long as the author and publisher are properly credited, which ensures maximum dissemination and a wider impact of our publications.

The book as a whole is distributed by MDPI under the terms and conditions of the Creative Commons license CC BY-NC-ND.

# Contents

About the Editors . . . . .	vii
Preface to “Machine Learning Methods with Noisy, Incomplete or Small Datasets” . . . . .	ix
<b>Cesar F. Caiafa, Zhe Sun, Toshihisa Tanaka, Pere Marti-Puig and Jordi Solé-Casals</b> Machine Learning Methods with Noisy, Incomplete or Small Datasets Reprinted from: <i>Applied Sciences</i> <b>2021, 11</b> , 4132, doi:10.3390/app11094132 . . . . .	1
<b>Cesar Federico Caiafa, Jordi Solé-Casals, Pere Marti-Puig, Sun Zhe and Toshihisa Tanaka</b> Decomposition Methods for Machine Learning with Small, Incomplete or Noisy Datasets Reprinted from: <i>Applied Sciences</i> <b>2020, 10</b> , 8481, doi:10.3390/app10238481 . . . . .	5
<b>Jigang Tong, Jiachen Zhang, Enzeng Dong and Shengzhi Du</b> Severity Classification of Parkinson’s Disease Based on Permutation-Variable Importance and Persistent Entropy Reprinted from: <i>Applied Sciences</i> <b>2021, 11</b> , 1834, doi:10.3390/app11041834 . . . . .	25
<b>Shan Wang, Feng Duan and Mingxin Zhang</b> Convolution-GRU Based on Independent Component Analysis for fMRI Analysis with Small and Imbalanced Samples Reprinted from: <i>Applied Sciences</i> <b>2020, 10</b> , 7465, doi:10.3390/app10217465 . . . . .	45
<b>Suguru Yasutomi, Tatsuya Arakaki, Ryu Matsuoka, Akira Sakai, Reina Komatsu, Kanto Shozu, Ai Dozen, Hidenori Machino, Ken Asada, Syuzo Kaneko, Akihiko Sekizawa, Ryuji Hamamoto and Masaaki Komatsu</b> Shadow Estimation for Ultrasound Images Using Auto-Encoding Structures and Synthetic Shadows Reprinted from: <i>Applied Sciences</i> <b>2021, 11</b> , 1127, doi:10.3390/app11031127 . . . . .	63
<b>Hafiz Farooq Ahmad, Hamid Mukhtar, Hesham Alaqail, Mohamed Seliaman and Abdulaziz Alhumam</b> Investigating Health-Related Features and Their Impact on the Prediction of Diabetes Using Machine Learning Reprinted from: <i>Applied Sciences</i> <b>2021, 11</b> , 1173, doi:10.3390/app11031173 . . . . .	83
<b>Xiaoyue Qiao, Zheng Zhang and Xin Chen</b> Multifrequency Impedance Method Based on Neural Network for Root Canal Length Measurement Reprinted from: <i>Applied Sciences</i> <b>2020, 10</b> , 7430, doi:10.3390/app10217430 . . . . .	101
<b>Karina Gibert and Xavier Angerri</b> The INSESS-COVID19 Project. Evaluating the Impact of the COVID19 in Social Vulnerability While Preserving Privacy of Participants from Minority Subpopulations Reprinted from: <i>Applied Sciences</i> <b>2021, 11</b> , 3110, doi:10.3390/app11073110 . . . . .	115
<b>Permatasari Silitonga, Alhadi Bustamam, Hengki Muradi, Wibowo Mangunwardoyo and Beti E. Dewi</b> Comparison of Dengue Predictive Models Developed Using Artificial Neural Network and Discriminant Analysis with Small Dataset Reprinted from: <i>Applied Sciences</i> <b>2021, 11</b> , 943, doi:10.3390/app11030943 . . . . .	163

<b>Seokjin Lee, Minhan Kim, Seunghyeon Shin, Sooyoung Park and Youngho Jeong</b> Data-Dependent Feature Extraction Method Based on Non-Negative Matrix Factorization for Weakly Supervised Domestic Sound Event Detection Reprinted from: <i>Applied Sciences</i> <b>2021</b> , <i>11</i> , 1040, doi:10.3390/app11031040 . . . . .	<b>179</b>
<b>Amaia Gil, Marco Quartulli, Igor G. Olaizola and Basilio Sierra</b> Learning Optimal Time Series Combination and Pre-Processing by Smart Joins Reprinted from: <i>Applied Sciences</i> <b>2020</b> , <i>10</i> , 6346, doi:10.3390/app10186346 . . . . .	<b>195</b>
<b>Jun Wang, Yuanyuan Xu, Hengpeng Xu, Zhe Sun, Zhenglu Yang and Jinmao Wei</b> An Effective Multi-Label Feature Selection Model Towards Eliminating Noisy Features Reprinted from: <i>Applied Sciences</i> <b>2020</b> , <i>10</i> , 8093, doi:10.3390/app10228093 . . . . .	<b>213</b>
<b>Pere Marti-Puig, Amalia Manjabacas and Antoni Lombarte</b> Automatic Classification of Morphologically Similar Fish Species Using Their Head Contours Reprinted from: <i>Applied Sciences</i> <b>2020</b> , <i>10</i> , 3408, doi:10.3390/app10103408 . . . . .	<b>231</b>
<b>Hangli Ge, Xiaohui Peng and Noboru Koshizuka</b> Applying Knowledge Inference on Event-Conjunction for Automatic Control in Smart Building Reprinted from: <i>Applied Sciences</i> <b>2021</b> , <i>11</i> , 935, doi:10.3390/app11030935 . . . . .	<b>255</b>
<b>Yonggeol Lee and Sang-II Choi</b> Training Set Enlargement Using Binary Weighted Interpolation Maps for the Single Sample per Person Problem in Face Recognition Reprinted from: <i>Applied Sciences</i> <b>2020</b> , <i>10</i> , 6659, doi:10.3390/app10196659 . . . . .	<b>273</b>
<b>Despoina Mouratidis, Katia Lida Kermanidis and Vilemini Sосoni</b> Innovatively Fused Deep Learning with Limited Noisy Data for Evaluating Translations from Poor into Rich Morphology Reprinted from: <i>Applied Sciences</i> <b>2021</b> , <i>11</i> , 639, doi:10.3390/app11020639 . . . . .	<b>287</b>

## About the Editors

### **Jordi Solé-Casals**

Jordi Solé-Casals holds a permanent position as a Full Professor at the University of Vic–Central University of Catalonia (UVic-UCC) and is the head of the Data and Signal Processing Research Group (DSP). He has been Visiting Scientist (since 2016) at the Department of Psychiatry of the University of Cambridge (UK) and Visiting Scientist (since 2020) at the College of Artificial Intelligence, Nankai University (China). He obtained a PhD degree with a European label in 2000 and a B.Sc. degree in Telecommunications in 1995, from the Polytechnic University of Catalonia, and a B.Hum in 2010, from the Open University of Catalonia. He was Visiting Researcher at the GIPSA Lab. in Grenoble (France), the Lab. for Advanced Brain Signal Processing, BSI-RIKEN in Wako (Japan) and the Tensor Learning Team, RIKEN-AIP, Tokyo (Japan). His research interests include signal processing, machine learning/deep learning and statistical modelling for applied sciences.

### **Zhe Sun**

Zhe Sun is currently a Research Scientist with the Image Processing Research Team, Center for Advanced Photonics, RIKEN National Science Institute, Japan. He is a researcher in the Japanese AVATAR X Project and Fugaku Supercomputer Project. In Fugaku Supercomputer Project, he is focusing on human scale brain simulation. He joined RIKEN, in 2015, as a Research Support Assistant. He obtained a Ph.D. degree in 2017 from Yokohama City University, Japan. He has been a Research Scientist with RIKEN since 2017. Since 2014, his research topics have been the development of the spiking neuron model and spiking neural network to understand and elucidate brain functions. His current research interests include brain inspired vision systems, large-scale brain simulation, high-performance computing and neuromorphic engineering.

### **Cesar F. Caiafa**

Dr. Prof. Cesar F. Caiafa currently holds a permanent position as Independent Researcher (since 2010) at IAR—CONICET and Adjunct Professor (since 2015) at the Engineering Faculty—University of Buenos Aires, Argentina. He has also been Visiting Scientist (since 2018) at the Tensor Learning Team, RIKEN Center for Advanced Intelligence Project (AIP), Tokyo, Japan. He was Research Scientist (2016–2018) at the Psychology and Brain Sciences Department, Indiana University, Bloomington, Indiana, USA. Research Scientist (2008 – 2010) and Visiting Scientist (2011–2018) at Lab. for Advanced Brain Signal Processing, BSI-RIKEN, Wako, Japan. He currently works on the development of machine learning algorithms exploiting tensor decompositions and sparsity with diverse applications ranging from Neuroscience to Astronomy.

**Pere Marti-Puig**

Pere Marti-Puig received a Ph.D. and an M.Sc. degree in Telecommunications Engineering from the BarcelonaTECH Polytechnic University of Catalonia (UPC), Barcelona, in 2001 and 1993. His current research interests are in the area of signal and image processing with particular emphasis on multiresolution signal processing techniques. He also works on machine learning and statistical modeling for applied sciences. He joined the Engineering Department at the University of Vic–Central University of Catalonia (UVIC-UCC) in 1993 as Assistant Lecturer and was promoted to Lecturer in 1993. He has been teaching and researching the IT area, from signal processing to control theory. He has been involved in several research and development projects.

**Toshihisa Tanaka**

Toshihisa Tanaka received a B.E., an M.E., and a Ph.D. degrees from the Tokyo Institute of Technology in 1997, 2000, and 2002, respectively. From 2000 to 2002, he was a JSPS Research Fellow. From October 2002 to March 2004, he was a Research Scientist at RIKEN Brain Science Institute. In April 2004, he joined the Tokyo University of Agriculture and Technology, where he is currently a Professor. His research interests include a broad area of signal processing and machine learning, including brain and biomedical signal processing, brain-machine interfaces and adaptive systems. He is a leading co-editor of *Signal Processing and Machine Learning for Brain–Machine Interfaces* (with Arvaneh, IET, U.K.), 2018.

# **Preface to "Machine Learning Methods with Noisy, Incomplete or Small Datasets"**

In many machine learning applications, available datasets are sometimes incomplete, noisy or affected by artifacts. In supervised scenarios, it could happen that label information has low quality, which might include unbalanced training sets, noisy labels and other problems. Moreover, in practice, it is very common that available data samples are not enough to derive useful supervised or unsupervised classifiers. All these issues are commonly referred to as the low-quality data problem. This book collects novel contributions on machine learning methods for low-quality datasets, to contribute to the dissemination of new ideas to solve this challenging problem, and to provide clear examples of application in real scenarios.

**Jordi Solé-Casals, Zhe Sun, Cesar F. Caiafa, Pere Marti-Puig, Toshihisa Tanaka**

*Editors*



Editorial

# Machine Learning Methods with Noisy, Incomplete or Small Datasets

Cesar F. Caiafa <sup>1,\*</sup>, Zhe Sun <sup>2</sup>, Toshihisa Tanaka <sup>3</sup>, Pere Marti-Puig <sup>4</sup> and Jordi Solé-Casals <sup>4,\*</sup><sup>1</sup> Instituto Argentino de Radioastronomía—CCT La Plata, CONICET/CIC-PBA/UNLP, 1894 V. Elisa, Argentina<sup>2</sup> Computational Engineering Applications Unit, Head Office for Information Systems and Cybersecurity, RIKEN, Wako-Shi 351-0198, Japan; zhe.sun.vk@riken.jp<sup>3</sup> Department of Electrical and Electronic Engineering, Tokyo University of Agriculture and Technology, Tokyo 184-8588, Japan; tanakt@cc.tuat.ac.jp<sup>4</sup> Data and Signal Processing Research Group, University of Vic-Central University of Catalonia, 08500 Barcelona, Spain; pere.marti@uvic.cat

\* Correspondence: ccaiafa@gmail.com (C.F.C.); jordi.sole@uvic.cat (J.S.-C.)

**Abstract:** In this article, we present a collection of fifteen novel contributions on machine learning methods with low-quality or imperfect datasets, which were accepted for publication in the special issue “Machine Learning Methods with Noisy, Incomplete or Small Datasets”, Applied Sciences (ISSN 2076-3417). These papers provide a variety of novel approaches to real-world machine learning problems where available datasets suffer from imperfections such as missing values, noise or artefacts. Contributions in applied sciences include medical applications, epidemic management tools, methodological work, and industrial applications, among others. We believe that this special issue will bring new ideas for solving this challenging problem, and will provide clear examples of application in real-world scenarios.

**Keywords:** artificial intelligence; imperfect dataset; imperfect dataset; machine learning



**Citation:** Caiafa, C.F.; Sun, Z.; Tanaka, T.; Marti-Puig, P.; Solé-Casals, J. Machine Learning Methods with Noisy, Incomplete or Small Datasets. *Appl. Sci.* **2021**, *11*, 4132. <https://doi.org/10.3390/app11094132>

Received: 20 April 2021

Accepted: 28 April 2021

Published: 30 April 2021

**Publisher’s Note:** MDPI stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.



**Copyright:** © 2021 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

## 1. Introduction

In many machine learning applications, available datasets are sometimes incomplete, noisy or affected by artifacts. In supervised scenarios, it could happen that label information is of low quality, which includes unbalanced training sets, noisy labels and other problems. Moreover, in practice, it is very common that available data samples are not enough to derive useful supervised or unsupervised classifiers. All these issues are commonly referred as the *low-quality data problem*. Machine learning researchers and practitioners have been working on various strategies to correctly handle the low-quality problem in recent years. Far from being solved, this problem still represents a fundamental and classic challenge in the artificial intelligence community.

The aim of this Special Issue was to collect novel contributions on machine learning methods for low-quality datasets, to contribute to the dissemination of new ideas to solve this challenging problem, and to provide clear examples of application in real scenarios. Despite the COVID-19 crisis and lockdowns in most countries, this Special Issue attracted great attention among researchers worldwide. A total number of twenty-one papers were submitted and fifteen of them were accepted after appropriate revisions. We were pleasantly surprised by the diversity of nationalities of contributors and the variety of the addressed problems in applied sciences ranging from medical and health applications through specific industrial case study examples. The authors of the published papers are from nine countries located in Europe, America, Africa and Asia.

In the following sections, the accepted papers and their corresponding most relevant contributions are summarized, which are grouped in the following categories: medical applications, epidemics management tools, methodological papers, industrial applications, and others.



## 2. Medical Applications

Interestingly, the majority of the contributions are related to specific applications in medicine. Three papers addressed different problems or diseases in Neuroscience. For example, in [1], Caiafa et al. (Argentina–Spain–Japan) reviewed recent approaches to deal with incomplete or noisy measurements by applying signal decomposition methods and showed their usefulness in epileptic intracranial electroencephalogram (iEEG) signals classification, among other applications. Finding epileptic focus with iEEG is usually difficult mainly because available datasets labeled by expert medical doctors are scarce. In [2], Tong et al. (China–South Africa) proposed a few-shot learning method for the severity assessment of Parkinson’s disease based on a small gait dataset. The proposed algorithm solves the small-data problem by using permutation-variable importance (PVI) and persistent entropy of topological imprints; as well as applying a support vector machine (SVM) classifier to achieve the severity classification of Parkinson disease patients. In [3], Wang et al. (China) addressed the problem of small and unbalanced datasets in functional magnetic resonance imaging (fMRI) for neuroscience studies. Their technique combines Independent Component Analysis (ICA) for dimensionality reduction, data augmentation to balance data and a convolution-gated recurrent unit (GRU) network. Results on episodic memory evaluation are reported.

The other papers that addressed medical applications are described as follows. In [4], Yasutomi et al. (Japan) introduced a deep learning method based on an auto-encoder architecture to detect and remove shadow artifacts in ultrasound images. The model can be trained on unlabeled data (unsupervised) or with few pixel labels available (semi supervised). The method has been applied to fetal heart diagnosis. In [5], Ahmad et al. (Saudi Arabia) investigated a machine learning approach to predict diabetes mellitus based on a handful set of features obtained by simple laboratory tests, allowing a cost-effective and rapid screening tool. They compared different machine learning classifiers and provided a set of recommendations based on those analyses. In [6], Qiao et al. (China) proposed a method to measure the length of the root canal length, which is crucial for an effective treatment of endodontics and periapicalitis. The authors employed a neural network on multifrequency impedance measurements.

## 3. Epidemics Monitoring and Management Tools

Machine learning has been demonstrated to have an important role in dealing with infectious diseases and epidemics. In this collection, two contributions are devoted to the development of tools to deal with some aspects of COVID-19 and dengue epidemics. More specifically, in [7], Gibert Oliveras et al. (Spain) reported the results of a project developed in Catalonia, Spain, owing to help in the COVID-19 crisis. The project allowed for quick territory screening providing relevant information to support informed decision-making and strategy and policy design. The authors proposed a data-driven methodology in order to deal with small subgroups of the population for statistical secrecy preserving. In [8], Silitonga et al. (Indonesia) developed prediction models to estimate the severity level of dengue based on the laboratory test results of the corresponding patients using artificial neural network (ANN) and discriminant analysis (DA) applied to very small datasets.

## 4. Methodological Articles

Four contributions proposed general methods for machine learning with low-quality datasets. In [1], the authors provided a unified review of decomposition methods, which includes linear decomposition, low-rank matrix/tensor factorization, sparse matrix/tensor decomposition and empirical mode decomposition (EMD) models. This paper illustrates the ability of these decomposition models to impute missing features, denoising and to artificially generate additional data samples (data augmentation) with examples to the brain–computer interface (BCI) and epileptic EEG analysis, among others. In [9], Lee et al. (South Korea) developed feature extraction methods based on the non-negative matrix factorization (NMF) algorithm and it is applied in weakly supervised sound event detection.

The algorithm considers learning from strongly and weakly labeled data. On the other side, in [10], Gil et al. (Spain) investigated the use of optimization in the preprocessing step of time series joining. More specifically, the authors proposed an error function to measure the adequateness of the joining and demonstrated the effectiveness of the proposed method on the synthetical datasets and real industrial process scenario. Finally, in [11], Wang et al. (China–Japan) proposed a novel multi-label feature selection approach by embedding label correlations (dubbed ELCs) in order to eliminate irrelevant and redundant, features, also referred as noisy features.

## 5. Applications to the Industry

This Special Issue also includes two papers studying the application of machine learning to specific practical problems in different industries: the fishing and smart buildings industries. In [12], Marti-Puig et al. (Spain) addressed the problem of distinguishing between different Mediterranean demersal species of fish that share a remarkably similar form and that are also used for the evaluation of marine resources. The authors employed both a binary and a multi-class classification problem based on very small datasets with unreliable labels. In [13], Ge et al. (Japan–China) proposed a unified and practical framework for knowledge inference inside the smart building.

## 6. Other Applications

Two very important machine learning problems face recognition and natural language processing. These two problems were addressed in this Special Issue for cases with low-quality datasets. In [14], Lee et al. (Korea) studied the problem of training a facial recognition system provided that only one sample per identity is available. The authors proposed a data augmentation technique by introducing changes in pixels in face images associated with variations by extracting the binary weighted interpolation map (B-WIM) from neutral and variational images in the auxiliary set. In [1], the EMD method was applied to remove noise in face images, thus improving the classification accuracy of a machine learning classifier. Finally, in [15], Mouratidis et al. (Greece) provided an application to natural language processing. They developed a deep learning schema for machine translation evaluation (English–Greek and English–Italian), based on different categories of information (linguistic features, natural language processing metrics and embeddings), by using a model for machine learning based on noisy and small datasets.

## 7. Conclusions

The correct handling of noisy, incomplete or small datasets remains an open problem in the artificial intelligence community. However, this Special Issue collects fifteen research papers providing general approaches to some low-quality datasets problems and clear practical examples in different applied sciences disciplines. This collection of papers represents a good reference for the current *state-of-the-arts*, also providing an excellent starting point for developing new advanced methods in the future.

**Funding:** This research received no external funding.

**Institutional Review Board Statement:** Does not apply.

**Informed Consent Statement:** Does not apply.

**Data Availability Statement:** Does not apply.

**Conflicts of Interest:** The author declares no conflict of interests.

## References

1. Caiafa, C.F.; Solé-Casals, J.; Marti-Puig, P.; Zhe, S.; Tanaka, T. Decomposition Methods for Machine Learning with Small, Incomplete or Noisy Datasets. *Appl. Sci.* **2020**, *10*, 8481. [[CrossRef](#)]
2. Tong, J.; Zhang, J.; Dong, E.; Du, S. Severity Classification of Parkinson's Disease Based on Permutation-Variable Importance and Persistent Entropy. *Appl. Sci.* **2021**, *11*, 1834. [[CrossRef](#)]

3. Wang, S.; Duan, F.; Zhang, M. Convolution-GRU Based on Independent Component Analysis for fMRI Analysis with Small and Imbalanced Samples. *Appl. Sci.* **2020**, *10*, 7465. [[CrossRef](#)]
4. Yasutomi, S.; Arakaki, T.; Matsuoka, R.; Sakai, A.; Komatsu, R.; Shozu, K.; Dozen, A.; Machino, H.; Asada, K.; Kaneko, S.; et al. Shadow Estimation for Ultrasound Images Using Auto-Encoding Structures and Synthetic Shadows. *Appl. Sci.* **2021**, *11*, 1127. [[CrossRef](#)]
5. Ahmad, H.F.; Mukhtar, H.; Alaqail, H.; Seliaman, M.; Alhumam, A. Investigating Health-Related Features and Their Impact on the Prediction of Diabetes Using Machine Learning. *Appl. Sci.* **2021**, *11*, 1173. [[CrossRef](#)]
6. Qiao, X.; Zhang, Z.; Chen, X. Multifrequency Impedance Method Based on Neural Network for Root Canal Length Measurement. *Appl. Sci.* **2020**, *10*, 7430. [[CrossRef](#)]
7. Gibert, K.; Angerri, X. The INSESS-COVID19 Project. Evaluating the Impact of the COVID19 in Social Vulnerability While Preserving Privacy of Participants from Minority Subpopulations. *Appl. Sci.* **2021**, *11*, 3110. [[CrossRef](#)]
8. Silitonga, P.; Bustamam, A.; Muradi, H.; Mangunwardoyo, W.; Dewi, B.E. Comparison of Dengue Predictive Models Developed Using Artificial Neural Network and Discriminant Analysis with Small Dataset. *Appl. Sci.* **2021**, *11*, 943. [[CrossRef](#)]
9. Lee, S.; Kim, M.; Shin, S.; Park, S.; Jeong, Y. Data-Dependent Feature Extraction Method Based on Non-Negative Matrix Factorization for Weakly Supervised Domestic Sound Event Detection. *Appl. Sci.* **2021**, *11*, 1040. [[CrossRef](#)]
10. Gil, A.; Quartulli, M.; Olaizola, I.G.; Sierra, B. Learning Optimal Time Series Combination and Pre-Processing by Smart Joins. *Appl. Sci.* **2020**, *10*, 6346. [[CrossRef](#)]
11. Wang, J.; Xu, Y.; Xu, H.; Sun, Z.; Yang, Z.; Wei, J. An Effective Multi-Label Feature Selection Model Towards Eliminating Noisy Features. *Appl. Sci.* **2020**, *10*, 8093. [[CrossRef](#)]
12. Marti-Puig, P.; Manjabacas, A.; Lombarte, A. Automatic Classification of Morphologically Similar Fish Species Using Their Head Contours. *Appl. Sci.* **2020**, *10*, 3408. [[CrossRef](#)]
13. Ge, H.; Peng, X.; Koshizuka, N. Applying Knowledge Inference on Event-Conjunction for Automatic Control in Smart Building. *Appl. Sci.* **2021**, *11*, 935. [[CrossRef](#)]
14. Lee, Y.; Choi, S.-I. Training Set Enlargement Using Binary Weighted Interpolation Maps for the Single Sample per Person Problem in Face Recognition. *Appl. Sci.* **2020**, *10*, 6659. [[CrossRef](#)]
15. Mouratidis, D.; Kermanidis, K.L.; Sosoni, V. Innovatively Fused Deep Learning with Limited Noisy Data for Evaluating Translations from Poor into Rich Morphology. *Appl. Sci.* **2021**, *11*, 639. [[CrossRef](#)]

Review

# Decomposition Methods for Machine Learning with Small, Incomplete or Noisy Datasets

Cesar Federico Caiafa <sup>1,2,\*</sup>, Jordi Solé-Casals <sup>3,\*</sup>, Pere Marti-Puig <sup>3</sup>, Sun Zhe <sup>4</sup>  
and Toshihisa Tanaka <sup>5</sup>

<sup>1</sup> Instituto Argentino de Radioastronomía—CCT La Plata, CONICET/CIC-PBA/UNLP, 1894 V. Elisa, Argentina

<sup>2</sup> Tensor Learning Team—Center for Advanced Intelligence Project, RIKEN, Tokyo 103-0027, Japan

<sup>3</sup> Data and Signal Processing Research Group, University of Vic-Central University of Catalonia, 08500 Vic, Catalonia, Spain; pere.marti@uvic.cat

<sup>4</sup> Computational Engineering Applications Unit, Head Office for Information Systems and Cybersecurity, RIKEN, Wako-Shi 351-0198, Japan; zhe.sun.vk@riken.jp

<sup>5</sup> Department of Electrical and Electronic Engineering, Tokyo University of Agriculture and Technology, Tokyo 184-8588, Japan; tanakat@cc.tuat.ac.jp

\* Correspondence: ccaiafa@fi.uba.ar (C.F.C.); jordi.sole@uvic.cat (J.S.-C.)

Received: 31 October 2020; Accepted: 24 November 2020; Published: 27 November 2020

**Abstract:** In many machine learning applications, measurements are sometimes incomplete or noisy resulting in missing features. In other cases, and for different reasons, the datasets are originally small, and therefore, more data samples are required to derive useful supervised or unsupervised classification methods. Correct handling of incomplete, noisy or small datasets in machine learning is a fundamental and classic challenge. In this article, we provide a unified review of recently proposed methods based on signal decomposition for missing features imputation (data completion), classification of noisy samples and artificial generation of new data samples (data augmentation). We illustrate the application of these signal decomposition methods in diverse selected practical machine learning examples including: brain computer interface, epileptic intracranial electroencephalogram signals classification, face recognition/verification and water networks data analysis. We show that a signal decomposition approach can provide valuable tools to improve machine learning performance with low quality datasets.

**Keywords:** empirical mode decomposition; machine learning; sparse representations; tensor decomposition; tensor completion

---

## 1. Introduction

Machine learning (ML) has been developing without a break since its beginning in the middle of the 20th century with the introduction of the first computers. ML comprises the design and study of algorithms that can automatically learn from observations and take optimal decisions or provide valuable outputs. With recent accelerated improvements in computing power and the availability of massive datasets, ML methods based on deep neural networks, usually referred as deep learning [1], gave rise to an Artificial Intelligence (AI) revolution. AI continues changing our daily lives contributing with extraordinary advances in scientific data analysis and new technological applications [2].

ML algorithms are based on the mathematical modelling of variables and their interaction mechanisms that can explain the observations (dataset). Complex datasets, such as natural images, speech or brain signals, usually requires sophisticated ML models to capture the probability distribution of data. Most sophisticated ML models are built upon feed-forward deep neural networks having from dozens to hundreds layers leading to a very large set of model parameters. To train

such big deep learning models requires accessing to extremely large datasets, which are not always available or they are too expensive to obtain.

Standard training algorithms for modern deep learning models assume that datasets are “infinite”, i.e., they are large enough to allow successful training of very large models. In practice, and particularly when an image dataset is not large enough, it is common practice to artificially generate additional samples by applying a composition of random class-preserving transformations on available data samples such as crops, translations and rotations, which is widely known as “data augmentation”.

Moreover, available ML algorithms not only assume “infinite” datasets, they were also designed for perfect input data samples. Nevertheless, in practical applications data samples often suffer from imperfections, such as missing or noisy features. For example, when recording electroencephalographic (EEG) signals, corrupted data can be originated from impedance mismatching, electrode disconnection, body movements, etc. [3]. Other practical problems where data samples can be incomplete include: computer vision systems where objects in the view field can be partially occluded [4]; recommendation systems built from the information gathered by different users where not all the users have fully completed their forms [5]; or medical datasets where typically not all tests can be performed on all patients [6].

In this article, we review recent techniques to alleviate the serious consequences of having different types of low-quality datasets in ML applications. We demonstrate that, by using decomposition methods, we can model one-dimensional and multi-dimensional signals, which allow us to artificially generate class-preserving new signals or make inference on missing/corrupted features.

This article is organized as follows: in Section 1.1, a review of the state-of-the-art approaches for low-quality datasets in ML is given; in Section 1.2, the mathematical notation used throughout the paper is introduced; in Section 2, a unified view of signal decomposition methods is presented, which includes: subspace approximation, Empirical Mode Decomposition (EMD), sparse representations and tensor decompositions; Section 3 covers practical applications including ML problems in neurosciences, face detection/classification and analysis of water networks data; finally, the main conclusions and discussion are presented in Section 4.

### 1.1. ML with Low Quality Datasets: State-of-the-Art and Recent Progress

In this paper, we focus on the following types of low-quality datasets: (1) small, and (2) having incomplete or corrupted samples. The following subsections provide an overview of current approaches and recent progress in artificially generating new training samples (Section 1.1.1), and how to deal with incomplete datasets (Section 1.1.2).

#### 1.1.1. Classical Data Augmentation

Artificial generation of training samples for machine learning has been used for many years, for example, in the form of virtual examples for training support vector machines in supervised learning [7,8]. In these papers, training data is augmented so that the learned model will be invariant to known transformations or perturbations. By using this technique, in [8], it was reported the lowest test error (0.6%) until that moment in 2002 on the well-known MNIST digit recognition benchmark task. Since then, data augmentation has been considered essential for the efficient training of neural networks, specially on images where it is typically performed in an ad-hoc manner by using class preserving transformations such as random cropping and rotations. Data augmentation is crucial to achieve nearly all state-of-the-art results, for example, in 2010 a new record on the best test error on MNIST dataset was reported (0.35%) by using deep neural networks [9]. Data augmentation is also fundamental to attain very good performance results on discriminative unsupervised feature learning based on convolutional neural networks [10]. While it is usually easy for domain experts to specify the involved transformations, for example the cropping and rotations in images, applications in other domains may require a non trivial choice of transformations. Motivated by this, in [11], the authors proposed a method for automating data augmentation by learning a generative sequence model

over user-specified transformation functions using a generative adversarial approach, which was successfully applied to image and text datasets.

Data augmentation was also applied in machine applications other than image classification, for example, in music source separation [12]. In that paper, the authors augmented the training datasets by randomly swapping left/right channels for each instrument, chunking into sequences for each instrument and mixing them from different songs sources, which demonstrated to boost the performance of deep neural networks for this task. Other application domains where data augmentation improved the discriminative power of classifiers include: biometrics [13], fault diagnosis in industry [14], radio frequency fingerprint identification [15], synthetic aperture radar (SAR) target recognition [16], and others.

From the theoretical point of view only very recently some understanding of the underlying theoretical principles involved in data augmentation procedure was provided. In [17], the authors provide a general model of augmentation as a Markov process and show that, combined with a  $k$ -nearest neighbor ( $k$ -NN) rule, is asymptotically equivalent to a kernel classifier. This result provides novel connections between prior work on invariant kernels, tangent propagation and robust optimization, giving an illustrative view on how augmentation affect the learning model.

### 1.1.2. Classical Approaches to ML with Incomplete Data

The classical approach to supervised learning with missing or noisy data is to preprocess the available data in order to infer missing/corrupted values such that standard ML algorithms can be used on the corrected dataset [18,19]. This imputation approach can be based on statistical principles, such as computing the mean of available samples for missing features or more sophisticated estimators like the regression imputation, which has the advantage that it can take into consideration the correlations among various features. Other imputation methods are based on machine learning ideas by estimating missing entries through the  $k$ -nearest neighbor [20], Self Organization Maps (SOM) [21], multilayer or recurrent neural networks [22,23], and others.

A different approach is to avoid direct imputation of lost inputs and rely on a probabilistic model of input data, based for example on the Gaussian Mixture Model (GMM) and learning model parameters through the Expectation Maximization (EM) algorithm and building a Bayesian classification. The advantage of this approach is that class labels of input data is fully exploited which helps for a correct imputation of missing entries [19,24]. However, the latter “model-based” approach has the disadvantage that it requires a good probabilistic data model, which is usually not available, especially for real-world applications such as those involving natural images.

Recently, some approaches based on the low-rank property of the features data matrix were investigated and algorithms for data completion were proposed incorporating the label information [25–27]. Since the rank estimation of a matrix is a computationally expensive task, usually based on the Singular Value Decomposition (SVD), the obtained algorithms are prohibitive to solve modern machine learning problems with large datasets. Additionally, as in the case of the probabilistic generative models, none of these methods considered complex classifications functions. To overcome this drawback, more recently, a framework based on neural network architectures such as autoencoders, multilayer perceptrons and Radial Basis Function Networks (RBFNs), was proposed for handling missing input data by setting a probabilistic model, e.g., a GMM, for every missing value, which is trained together with the NN weights [28]. This method combined the great capability of NNs to approximate complex decision functions with the nice formulation of the GMM to model missing data. However, it inherited the drawbacks of GMMs, e.g., they are not well suited to higher-dimensional datasets.

### 1.2. Mathematical Notation and Definitions

Vectors and matrices are denoted using boldface lower- and upper-case letters, respectively. For example  $\mathbf{x} \in \mathbb{R}^I$  and  $\mathbf{A} \in \mathbb{R}^{I \times J}$  represent a vector and a matrix, respectively. Columns of a matrix  $\mathbf{A} \in \mathbb{R}^{I \times J}$  are referred as vectors  $\mathbf{a}_j \in \mathbb{R}^I$ .

A tensor is a multidimensional array generalizing vectors and matrices to higher number of dimensions. For example, a tensor  $\underline{\mathbf{X}} \in \mathbb{R}^{I \times J \times K}$  is a 3D array of real numbers whose elements  $(i, j, k)$  are referred to as  $x_{ijk}$ . The individual dimensions of a tensor are referred to as modes (1st mode, 2nd mode, and so on). By generalization of matrix multiplication, a tensor can be multiplied by a matrix in a specific mode, only if their size matches. Given a tensor  $\underline{\mathbf{X}} \in \mathbb{R}^{I_1 \times I_2 \times \dots \times I_N}$  and a matrix  $\mathbf{A} \in \mathbb{R}^{J \times I_n}$ , the mode- $n$  product  $\underline{\mathbf{Y}} = \underline{\mathbf{X}} \times_n \mathbf{A} \in \mathbb{R}^{I_1 \times \dots \times I_{n-1} \times J \times I_{n+1} \times \dots \times I_N}$  is defined by:  $y_{i_1 \dots i_{n-1} i_{n+1} \dots i_N} = \sum_{i_n=1}^{I_n} x_{i_1 \dots i_n \dots i_N} a_{i_n j}$ , with  $i_k = 1, 2, \dots, I_k$  ( $k \neq n$ ) and  $j = 1, 2, \dots, J$ .

The  $\ell_0$ -norm  $\|\mathbf{x}\|_0$  of a vector  $\mathbf{x} \in \mathbb{R}^N$  is defined as the number of non-zero entries of the vector. When the number of non-zero entries is much less than the dimension of the vector, i.e.,  $\|\mathbf{x}\|_0 \ll N$ , the vector is sparse.

Given two matrices:  $\mathbf{A} \in \mathbb{R}^{I_1 \times J_1}$  and  $\mathbf{B} \in \mathbb{R}^{I_2 \times J_2}$ , the Kronecker product is defined as follows:

$$\mathbf{A} \otimes \mathbf{B} = \begin{bmatrix} a_{1,1}\mathbf{B} & a_{1,2}\mathbf{B} & \dots & a_{1,J_1}\mathbf{B} \\ a_{2,1}\mathbf{B} & a_{2,2}\mathbf{B} & \dots & a_{2,J_1}\mathbf{B} \\ \vdots & \vdots & \ddots & \vdots \\ a_{I_1,1}\mathbf{B} & a_{I_1,2}\mathbf{B} & \dots & a_{I_1,J_1}\mathbf{B} \end{bmatrix}. \tag{1}$$

## 2. Methods

The idea of extracting valuable information from a dataset by decomposing each of the signals (data samples) as a sum of simpler components, is a very well established technique originated in branches of mathematics such as functional analysis and statistics. The general idea behind any decomposition method is to obtain a compact model that can capture the essential information of a signal or dataset. This compression capability will allow us, for example, to generate artificial data by adapting individual components, or recombining them in a different way, and using the decomposition as a generative model. On the other side, when data is incomplete, we can fit a decomposition model such that the available information is replicated as much as possible and we can use the model to estimate the values of missing data points.

The goal of this article is to present a unifying view of several useful decomposition methods and illustrate about its applications to practical ML problems. In the following, we present a mathematical formulation of the decomposition methods that will be used through the paper.

### 2.1. A unified View of Data Decomposition Models for ML

Given a linear subspace  $\mathcal{U} \subset \mathbb{R}^N$ , every vector data sample (one-dimensional signal)  $\mathbf{x} \in \mathcal{U}$  is decomposed into a sum of  $I$  components if it can be written as:

$$\mathbf{x} = \sum_{i=1}^I \alpha_i \boldsymbol{\phi}_i + \mathbf{r} = \boldsymbol{\Phi} \boldsymbol{\alpha} + \mathbf{r}, \tag{2}$$

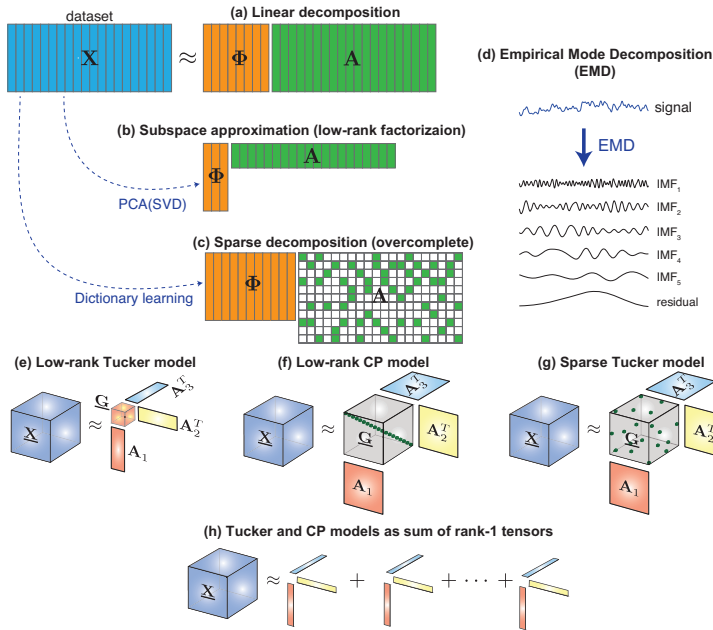
where  $\alpha_i \in \mathbb{R}$ , the set of vectors  $\{\boldsymbol{\phi}_i \in \mathbb{R}^N\}$  ( $i = 1, 2, \dots, I$ ) and  $\mathbf{r} \in \mathbb{R}^N$  are the coefficients, the generators of the linear subspace  $\mathcal{U}$ , and the residual or approximation error, respectively. A matrix notation is also shown in the rightmost part of Equation (2), where  $\boldsymbol{\Phi} \in \mathbb{R}^{N \times I}$  contains vectors  $\boldsymbol{\phi}_i$  as columns and  $\boldsymbol{\alpha} \in \mathbb{R}^I$  is the vector of coefficients. In some classical decomposition methods, vectors  $\boldsymbol{\phi}_i$  are constructed by theoretical principles as it is the case of the discrete Fourier, Cosine or Wavelet orthogonal bases, just to mention few, where the number of components equal to the space dimension ( $I = N$ ), meaning that those bases span the whole space  $\mathbb{R}^N$ .



Given a dataset composed of  $J$  vector samples  $\mathbf{x}_j \in \mathbb{R}^N, j = 1, 2, \dots, J$ , by arranging them as columns of matrix  $\mathbf{X} \in \mathbb{R}^{N \times J}$ , we can write the following matrix factorization equation (see Figure 1a):

$$\mathbf{X} \approx \Phi \mathbf{A}, \tag{3}$$

where  $\Phi \in \mathbb{R}^{N \times R}$  and  $\mathbf{A} \in \mathbb{R}^{R \times J}$  has entries  $\alpha_{i,j}$ . Such a compact representation of datasets can be accomplished in several ways. For example, by using a subspace of a lower dimension ( $R < N$ ) or using a sparse matrix  $\mathbf{A}$  meaning that each of the signals are approximated using few non-zero coefficients, compared to the size of the space  $N$ . In the following subsections we describes these possible models.



**Figure 1.** Decomposition models. (a) **General linear model:** a collection of vector data samples organized as columns of a matrix dataset  $\mathbf{X}$  is approximated by the product of matrices  $\Phi$  and  $\mathbf{A}$ . (b) **Subspace approximation:** all vectors in the dataset are approximated by linear combination of few vectors (principal components). (c) **Sparse coding:** each vector in the dataset is approximated by the linear combination of atoms (columns of a dictionary  $\Phi$ ). In both, (b,c), the optimal choice of matrix  $\Phi$  can be computed from the dataset itself by means of the SVD and a dictionary learning algorithm, respectively. (d) **EMD:** every single signal is decomposed as a sum of characteristic modes. Tensor decomposition models such as **Low-rank Tucker** (e), **Low-rank CP** (f) and **Sparse Tucker** (g) can be written as sum of rank-1 tensors (h).

### 2.2. Subspace Approximation (PCA)

If signals in a dataset can be well approximated within a subspace of lower dimension ( $R < N$ ), we can find the optimal basis by applying the celebrated Principal Component Analysis (PCA) algorithm. This method was originally introduced in statistics by Pearson in 1901 [29], but developed later independently by Karhunen [30] and Loève [31]. PCA basis vectors  $\phi_i$  and their associated coefficients are easily computed by means of a Singular Value Decomposition (SVD) of the data covariance matrix. Formally, given a set of normalized signals  $\{\mathbf{x}_j \in \mathbb{R}^N\}$  (zero-mean and unit-variance samples  $n = 1, 2, \dots, J$ ), the rank- $R$  PCA decomposition of any signal  $\mathbf{x}_j$  in the set is obtained by Equation (2), with orthonormal vectors  $\phi_i$  corresponding to the first  $R < N$  dominant singular vectors



of the data covariance matrix and coefficients are computed by  $\alpha_{i,j} = \mathbf{x}_j^T \boldsymbol{\phi}_i$ . In this case, the obtained decomposition model is also referred as low-rank matrix factorization as illustrated in Figure 1b.

### 2.3. Sparse Decomposition (SD)

More recently, in the signal processing field, it was discovered that a better way to capture the structure of natural images, speech audio and other types of signals is to have available a large collection of prototype atoms  $\{\boldsymbol{\phi}_i \in \mathbb{R}^N\}$  ( $i = 1, 2, \dots, I$ ) with  $I \geq N$  and use only few and distinctive coefficients to represent every signal  $\mathbf{x}$  in the space. This model is mathematically described by adding a sparsity constraint  $\|\boldsymbol{\alpha}\|_0 \leq K$  to the model of Equation (2) leading to the following equation:

$$\mathbf{x} \approx \boldsymbol{\Phi} \boldsymbol{\alpha}, \text{ with } \|\boldsymbol{\alpha}_i\|_0 \leq K \ll N, \tag{4}$$

This approach is usually referred as “sparse coding” or “sparse representation” of signals and matrix  $\boldsymbol{\Phi} \in \mathbb{R}^{N \times I}$  is called a “dictionary” [32,33] (see Figure 1c).

For general overcomplete ( $I \geq N$ ) dictionaries, the sparse vector of coefficients can be obtained by applying some of the available algorithms that were designed to solve the sparse coding problem, which includes greedy methods such as matching pursuit (MP) [34], orthogonal matching pursuit (OMP) [35], compressive sampling matching pursuit (CoSaMP) [36],  $\ell_1$  norm minimization methods such as basis pursuit [37] and many others (see [38] for a review of algorithms).

Some theoretically derived dictionaries, e.g., those based on the Discrete Cosine Transform (DCT) or Wavelet Transform (WT), are excellent candidates for sparse coding. However, in some ML applications when large datasets are available, sometimes it is good idea to learn an optimal dictionary for an specific dataset. To that end, some dictionary learning algorithms are proposed in the literature [39,40]. Sparse coding has to date a large record of successful applications in signal processing tasks like compressed sensing [41,42], blind source separation [43], denoising [39], inpainting [44] and others.

### 2.4. Empirical Mode Decomposition (EMD)

In the previously introduced methods (PCA and SD), one set of vectors  $\{\boldsymbol{\phi}_i\}$  is used to generate every signal  $\mathbf{x}$  in the dataset, meaning that the chosen generators are optimal for a particular dataset. Other methods have been proposed to find signal specific set of components. This was the case of the Empirical Mode Decomposition (EMD), which was first described by Huang N. et al. in [45] as a new method to analyze nonlinear and non-stationary signals. EMD is a data-based approach that decomposes any signal into a sum of so-called Intrinsic Mode Functions (IMF) plus a residual as illustrated in Figure 1d. Therefore, the original signal is modelled as a linear combination of amplitude and frequency modulation (AM-FM) functions. Every single function (IMF) is capturing specific information from a different frequency band present in the signal and is obtained in an iterative sifting procedure. Other variants have been defined since the introduction of EMD, but they all share the basic steps presented below. Let us suppose that we want to decompose a signal  $x(t)$  by means of EMD. Then its decomposition will be calculated as follows:

1. Determine the local maxima and minima of the signal  $x(t)$ .
2. Calculate the upper (lower) envelope by interpolating the local maxima (minima) points. The interpolation can be carried out in different ways (linear interpolation, spline interpolation, etc.), which could lead to slightly different results.
3. Calculate the local mean  $m(t)$  by averaging the upper and lower envelopes.
4. Calculate the first IMF candidate  $h_1(t) = x(t) - m(t)$ .
5. Checks whether candidate  $h_1(t)$  meets the criteria to be an IMF:
  - If  $h_1(t)$  meets the criteria, define the first IMF as  $c_1(t) = h_1(t)$ .
  - If  $h_1(t)$  does not meet the criteria, set  $x(t) = h_1(t)$  and repeat from step 1

The next IMF will be extracted using the same procedure on the signal  $r_1(t)$  that remains after subtracting the first IMF from the signal:  $r_1(t) = x(t) - c_1(t)$ . The process stops when two consecutive IMFs are (almost) identical and the empirical mode decomposition of the signal  $x(t)$  is written as:

$$\mathbf{x}(t) = \sum_{i=1}^n c_i(t) + r_n(t), \tag{5}$$

indicating that the original signal  $x(t)$  has been decomposed in  $n$  IMFs plus a residual signal. This residual signal captures the trend (or the mean) of the original signal.

### 2.5. Tensor Decomposition (TD)

Sometimes input data samples have a multidimensional structure or it is useful to arrange one-dimensional signals into multidimensional arrays or tensors. For example, EEG signals are simultaneously recorded with multiple sensors (electrodes) thus, for each subject, a (time  $\times$  channel) matrix is recorded. A natural way to construct a tensor for an EEG experiment is to use a third dimension to index subject, which results in a three dimensional data tensor  $\mathbf{X} \in \mathbb{R}^{I_1 \times I_2 \times I_3}$ , where  $I_1, I_2, I_3$  correspond to numbers of time samples, channels (sensors) and subjects, respectively.

Matrix factorization models, such as PCA in Equation (3), can be generalized to tensors by means of the Tucker( $R_1, R_2, R_3$ ) decomposition [46]. Given a data tensor  $\mathbf{X} \in \mathbb{R}^{I_1 \times I_2 \times I_3}$ , it can be decomposed as:

$$\mathbf{X} = \mathbf{G} \times_1 \mathbf{A}_1 \times_2 \mathbf{A}_2 \times_3 \mathbf{A}_3 + \mathbf{R}, \tag{6}$$

where  $\times_n$  is the mode- $n$  tensor-by-matrix product.  $\mathbf{G} \in \mathbb{R}^{R_1 \times R_2 \times R_3}$  is the core tensor and  $\mathbf{A}_n \in \mathbb{R}^{I_n \times R_n}$  are factor matrices (Figure 1e). As a particular case, when core tensor is diagonal with  $R = R_1 = R_2 = R_3$  this model is reduced to the CANDECOMP/PARAFAC or CP( $R$ ) decomposition model (see Figure 1f), which has demonstrated to be very useful in a wide range of applications [47,48]. It is interesting to note that Equation (6) can be also written as a sum of rank-1 tensors as shown in Figure 1h which, if vectorized, it is then reduced to our general decomposition model of Equation (2) as follows:

$$\mathbf{x} = \sum_{i_1, i_2, i_3} g_{i_1 i_2 i_3} \phi_{i_1 i_2 i_3} + \mathbf{r}, \tag{7}$$

where  $\mathbf{x} = \text{vec}(\mathbf{X})$ ,  $\mathbf{r} = \text{vec}(\mathbf{R})$ ,  $g_{i_1 i_2 i_3}$  are coefficients and  $\phi_{i_1 i_2 i_3} = \mathbf{a}_3^{i_3} \otimes \mathbf{a}_2^{i_2} \otimes \mathbf{a}_1^{i_1}$  with  $\mathbf{a}_n^i$  denoting the  $i$ -column of matrix  $\mathbf{A}_n$ . Tucker decomposition can provide data compression under two very different model assumptions leading to the following cases:

**Low-rank Tucker decomposition:** when the core tensor is much smaller than the original, i.e.,  $R_n \ll I_n$  [47,48] (see Figure 1e)

**Sparse Tucker decomposition:** when core tensor is of the same size or larger than tensor  $\mathbf{X}$  but it is sparse as illustrated in Figure 1g. In this case, by looking at Equation (7), we conclude that the Sparse Tucker model corresponds to the classical Sparse Coding model of (4) with a dictionary that is obtained as the Kronecker product of three mode dictionaries, i.e.,  $\Phi = \mathbf{A}_3 \otimes \mathbf{A}_2 \otimes \mathbf{A}_1$  [49,50]. Mode dictionaries can be chosen from classical sparsifying transforms such as wavelets, cosine transform and others or, if enough data is available, they can be learned from a dataset, which usually provides higher levels of sparsity and compression. A Kronecker dictionary learning algorithm was introduced in [50] and later a variant with orthogonality constraints was proposed in [51].

### 2.6. Comparison of Methods for ML with Low-Quality Datasets

In Table 1, we summarize all the methods discussed in this paper and compare them in terms of their main characteristics, shortcomings, advantages and main applications. A detailed reference to the sections of this article where each of the methods is presented and discussed is given. Furthermore, main bibliographic references are indicated for each of the approaches.

**Table 1.** Comparison of methods for Machine Learning (ML) problems with low-quality datasets. Article sections in which these methods are discussed are noted in the first column and relevant references are included in the last column.

Method	Characteristics	Shortcomings	Advantages	Application	References
<b>Class preserving transforms</b> (Section 1.1.1)	Ad-hoc; mostly images oriented but some extensions to other types of data were explored	Limited theory available; difficult to apply to arbitrary type datasets	Easy to use; widely available in deep learning platforms	Data augmentation	[7–17]
<b>Empirical Mode Decomposition (EMD) based data generation</b> (Sections 2.4, 3.1.2, 3.1.3 and 3.2)	Ad-hoc; based on the manipulation and recombination of Intrinsic Mode Functions (IMFs);	Lack of theoretical ground	Easy to use; capture dataset discriminative features; denoising power	Electroencephalography (EEG)/invasive EEG (iEEG) data augmentation and denoising	[45,52–55]
<b>Transform domain based data generation</b> (Section 3.1.3)	Ad-hoc; based on the manipulation and recombination of spectrum domain components obtained by Discrete Cosine Transform (DCT), Wavelets, etc.	Lack of theoretical ground	Easy to use; capture dataset discriminative features	iEEG data augmentation	[45,52,53,56]
<b>Statistical imputation</b> (Section 1.1.2)	Preprocessing step in ML; exploit statistical properties of datasamples; wide variety of methods, from simple ones (mean) to more sophisticated (regression, k-Nearest Neighbor (KNN), Self Organization Map (SOM), etc.)	Does not use the class label information of data samples	Computationally efficient	ML with incomplete or corrupted data	[18–23]
<b>Probabilistic modelling</b> (Section 1.1.2)	Gaussian Mixture Model (GMM) as data model; Bayesian classification; fitting model and classifiers in an Expectation-Maximization (EM) fashion; can be adapted to deep neural networks	Computationally expensive	Incorporates class label information of data samples; elegant theoretical approach	ML with incomplete or corrupted data	[19,24,28]
<b>Low-rank matrix completion</b> (Section 1.1.2)	Based on Singular Value Decomposition (SVD)	Computationally very expensive; not suitable for complex boundary functions	Incorporates class label information of data samples	ML with incomplete or corrupted data	[25–27]
<b>Tensor decomposition (TD) based imputation</b> (Sections 2.5, 3.1.1 and 3.3)	Preprocessing step in ML; based on low-rank TDs (e.g., Tucker, CANDECOMP/PARAFAC (CP), etc.) or sparse TDs	Does not use the label information of data samples	Exploits intricate relationship among modes in multidimensional data	ML with incomplete or corrupted data	[50,57–62]

### 3. Results

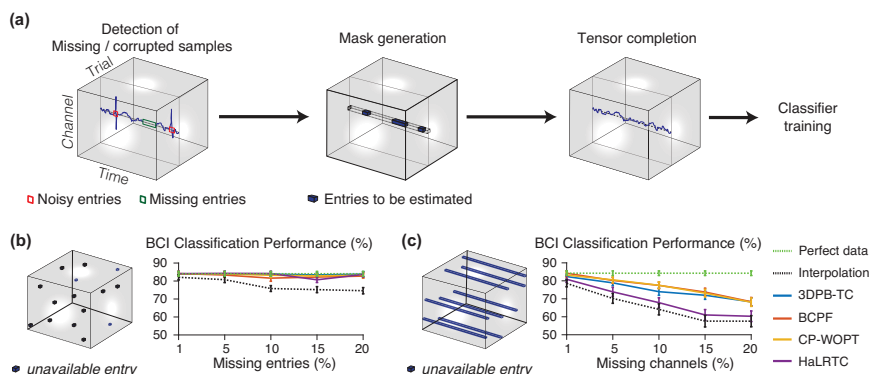
This section illustrates the application of the methods discussed above and presents the results obtained in the following selected case studies: brain signal classification with missing, corrupted or small datasets (Section 3.1); classification of noisy faces (Section 3.2) and analysis of water network data (Section 3.3).

#### 3.1. Brain Signal Classification

Brain signal activity can be recorded non-invasively by using, for example, electroencephalography (EEG) or invasively through electrodes located in the brain (iEEG). Decoding brain activity have found many applications in medicine and has great potential for the next generation of human-machine communication technologies. In a Brain Computer Interface (BCI) application, a user generates specific brain activity patterns that can be decoded by a machine learning algorithm. One popular paradigm in BCI is Motor Imagery (MI), which states that the brain activity generated by a subject imagining movement of one of their limbs for a few seconds activate areas in the motor cortex, which are similar to those that are activated with real movements, thus this particular neural activity can be detected by a machine learning algorithm (classifier).

##### 3.1.1. BCI with Missing/Corrupted Measurements

In BCI applications, noisy or missing data can arise. This can occur due to the lack of connection between wireless EEG headset and the computer, or because artifacts appear due to muscle movements, eye movements or electromagnetic interference, among others. Since EEG measurements can be organized as multidimensional datasets, in [57] a tensor completion approach was proposed, which consists in fitting a tensor decomposition model to the available clean measurements and then infer the noisy or missing parts based on those models (see Figure 2a). The advantage of using tensor methods, compared to classical interpolation algorithms, lies in the ability of these models to handle multidimensional information, in other words, they can capture the intricate relationship among entries in a multidimensional signal. For example, to infer a missing entry in an EEG data tensor, these methods can efficiently exploit the available information in other channels, time samples and trials.



**Figure 2.** Training a BCI classifier (LDA) with noisy/missing EEG measurements. (a) Preprocessing steps: first, the positions in which the data is missed/corrupted are identified; then, a mask is created to ignore values in those positions; and finally, the tensor model reconstructs the missing data. (b) Results with randomly missing entries. (c) Results with random missing channels. (Figure adapted from [57]).

Several tensor decomposition models and tensor completion algorithms were compared on a freely available dataset (<http://mon.uvic.cat/data-signal-processing/software/>) in [57], which are

based on the CP model of the whole tensor, such as the CP Weighted Optimization (CP-WOPT) [58], the High accuracy Low Rank Tensor Completion (HaLRTC) [59] and the Bayesian CP factorization (BCPF) for tensor completion [60]; and one method that uses the Sparse Tucker decomposition of every  $6 \times 6 \times 6$  tensor patch (subtensors), the 3D Patch-based Tensor Completion (3DPB-TC) [50]. In the latter case, a Kronecker dictionary was first learned from a clean EEG training dataset. For the experiments with incomplete measurements, two different patterns of missing data were considered: random missing entries, and random missing channels, as shown in Figure 2b,c, respectively. The latter case, represent a realistic situation in which some electrodes are simultaneously disconnected during a complete trial.

The performance of the tensor completion algorithms was compared, together with a simple interpolation strategy as shown in Figure 2b,c. The classification accuracy of imagined movement in a BCI experiment, using a Linear Discriminant Analysis (LDA) classifier, was evaluated in the perfect case (no missing data) and with each one of the recovered missing data through the tensor completion algorithms. Experimental results demonstrated that all tensor completion algorithms were able to recover missing samples increasing the classification performance compared to a simple interpolation approach, because tensor methods are able to exploit the multidimensional correlation of data. As expected, the random missing samples was easier to reconstruct using the neighbour points, while the random missing channels was more difficult because the amount information was missed in the same neighbourhood (the same channel, in that case). Therefore, as experimentally demonstrated, tensor completion algorithms could be used in real BCI data to avoid discarding noisy frames or frames with missing data, and instead recover the missing data using that technique.

### 3.1.2. Efficient Data Augmentation for BCI

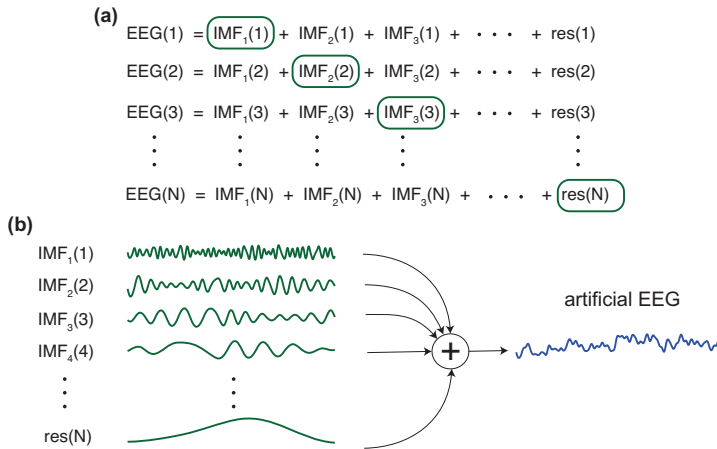
Small datasets are common in many EEG applications. This is especially the case when developing systems for automatically detect some brain disease or brain injuries. Basically, because sometimes it is not easy or just impossible to have enough patients from which to record EEG or iEEG signals. MI-BCI systems for neurorehabilitation require a calibration step before being used. This is due to the fact that the system needs a classifier which is particular for each subject and session because, for example, the location of the electrodes in each session will never be exactly the same. The Common Spatial Pattern (CSP) algorithm [63] is habitually used to extract features. This calibration step implies recording several MI frames which will be used to extract the CSP filters and train the classifier. Each frame is composed by the EEG recordings of a particular trial.

Since the quality of the classifier can be greatly improved by using a large number of frames from each type of MI [63], it is habitual to record 100 or more frames, in total. Taking into account that the MI paradigm can last about 10 s per frame, approximately a minimum of 16 min will be employed in the recording session. Over that, the time needed to set up the EEG montage has to be added.

A way to shorten the calibration time is by reducing the number of registered frames but generating artificial ones to keep high the total number of frames. While available data augmentation techniques are proved to be efficient to boost the training of neural networks and support vector machines, they were developed mainly for image datasets, where natural transformations are cropping and rotations, for example. These type of transforms have no sense for EEG data and new methods for data augmentation need to be developed. Methods based on signal decomposition and recombination of its main components are a natural way to solve that issue. In [52], the authors developed, for the first time, a method to generate artificial frames based on an EMD decomposition/combination strategy. Starting from a real frame collection, a new artificial frame of a specific class is built as described in Figure 3, comprising the following steps:

1. Randomly select  $N$  frames from the set of frames belonging to the selected class.
2. Decompose, using EMD, each one of the  $N$  frames, generating a set of IMFs per channel and frame.

3. Then, select the first IMF from the first selected frame (one per channel and keeping the same position for each channel), the second IMF from the second selected frame, and successively until the  $N$ th frame, which contributes with its  $N$ th IMF.
4. Add up all the IMFs corresponding to the same channel to build each new EEG channel of the new artificial frame.



**Figure 3.** EEG data augmentation: (a) For each new EEG signal to be generated,  $N$  available EEG signals are randomly selected and their EMDs are computed. (b) To generate an artificial EEG signal, IMFs from different signals are combined.

Using this method, authors in [52] were able to diminish the amount of acquired data for the calibration step in a BCI scenario while maintaining the performance. Specifically, they replaced original frames with artificial frames and tested the behaviour of the classifier derived from the data. Depending on the percentage of artificial frames in the data, they concluded that up to 50% of the original frames could be replaced without affecting the classifier’s performance as it is presented in Table 2. The performance of each classifier was validated through the median absolute deviation (MAD) method to detect outliers [64], and the dispersion ratio  $R$  was calculated as

$$R = |(x - \bar{x}) / \text{MAD}|, \tag{8}$$

where, for a set of measures,  $\bar{x}$  is the median value and  $x$  is the measure to be tested. In the experiments, the error rate of the classifier was tested. Usually, if  $R < 3$  the measure  $x$  is not considered to be an outlier, i.e., this classifier had a similar behaviour even if a specific percentage of real frames was replaced by artificial frames. We can see in Table 2 that four subjects (S01, S04, S05 and S07) have a value of  $R < 2.0$  for both left and right sides (motor imagery of left and right arm movement, respectively), and only one subject, S02, had a value of  $R > 3.0$  for the right side movement imagination.

**Table 2.** Dispersion ratio  $R$  computed in seven subjects (S01-S07) with Equation (8) for right (R) and left (L) classes at different levels of used artificial frames (AF). Results with  $R > 3$  are highlighted in red and with  $2 < R < 3$  in orange.

AF(%)	S01		S02		S03		S04		S05		S06		S07	
	R	L	R	L	R	L	R	L	R	L	R	L	R	L
2.5	0.12	0.67	0.22	0.64	0.58	1.27	0.32	0.31	0.18	0.27	0.33	0.64	0.34	0.69
5.0	0.05	1.03	0.82	0.56	1.11	1.02	0.46	0.45	0.18	0.35	0.47	0.83	0.01	0.63
7.5	0.29	0.88	1.03	0.07	1.06	1.51	0.51	0.51	0.00	0.02	1.17	1.49	0.46	0.62
10.0	0.37	1.13	0.99	0.11	1.19	1.75	0.80	0.46	0.38	0.08	1.04	1.66	0.49	0.84
12.5	0.24	0.94	1.42	0.04	1.89	1.86	1.00	0.44	0.46	0.27	0.87	1.52	0.40	0.85
25.0	0.09	1.44	2.79	0.44	2.13	1.94	1.28	0.61	0.96	0.78	0.71	2.09	0.51	1.28
37.5	0.11	1.55	3.12	0.41	1.97	2.01	1.20	0.69	1.07	1.18	0.57	2.66	0.73	1.92
50.0	0.15	1.45	2.86	1.00	2.18	2.68	1.27	1.06	1.42	1.23	0.62	2.76	0.73	1.86

Another application based on the same idea, but for deep neural network classifiers purposes, was proposed in [53]. Here, the authors also used a BCI scenario to exemplify how the EMD decomposition/recombination technique could be useful to create enough artificial data to train a deep learning structure while avoiding overfitting. This is a very important application when dealing with small datasets. If the classifier has many parameters, which is what happens in a deep learning structure, the system can easily suffers from overfitting. The only way to reduce it is by simplifying the classifier, hence, for example changing the deep learning structure for a more traditional machine learning structure with few parameters, or increasing the amount of data. When data is difficult to acquire or impossible for any reason (economical, practical, availability, etc.), the EMD decomposition/recombination technique can be used to generate artificial data. This is what was explored in [53]. In this work, a convolutional neural network and a wavelet neural network were proposed to classify BCI data. To be able to train deep learning structures with few data, the method described above was used as a data augmentation strategy. The authors showed experimentally that the artificial EEG frames were useful to improve the training of neural networks.

### 3.1.3. Epileptic Focal Detection with Limited Data

Epilepsy is a general term for a condition that causes repetitive seizures caused by excessive activity of neurons in the cerebrum, and such excessive activity is detected on EEG tests. Focal epilepsy, in which a part of the brain becomes abnormally excited and causes seizures, may be reduced or cured by removing the abnormally excited part of the brain (epileptogenic zone). Identification of the epileptogenic area requires intracranial electroencephalography (iEEG), in which electrodes are placed intracranially and measured, as well as brain imaging evaluation such as MRI.

The identification of the epileptogenic zone requires a long-term iEEG recording. Besides, the number of clinical experts (epileptologist) is limited. Therefore, the automated identification of the brain area of seizure onset of focal epilepsy (a.k.a automated focal identification) using interictal (non-seizure) iEEG signals is in strong demand. However, clinical iEEG data usually stay in each medical facility and cannot be in public, so that the amount of data available for training a machine learning model is also limited. This situation brings the necessity of appropriate processing for a small dataset.

A possible and straightforward way to cope with this problem is data augmentation. Two types of approaches to the iEEG data augmentation have been proposed recently. The first method is to augment data in the signal domain [54]. The second method is the data augmentation in the feature domain [56]. Both methods work efficiently in identifying focal locations from interictal iEEG.

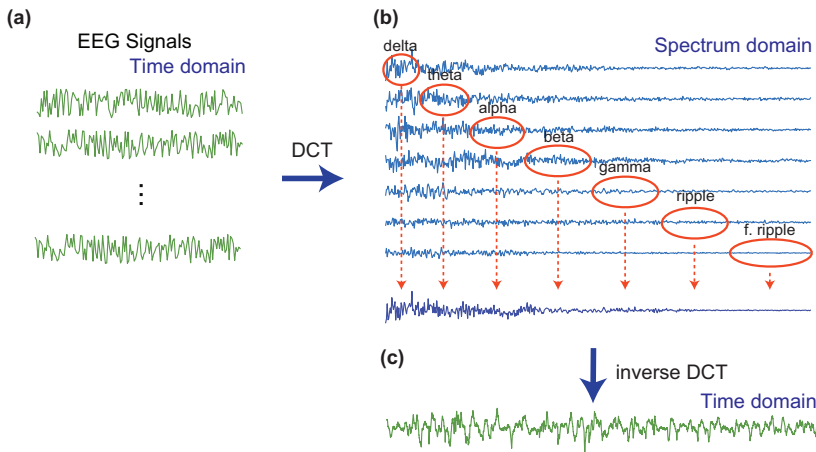
The data augmentation in the signal domain applies orthogonal transforms such as the discrete Fourier transform, the discrete cosine transform (DCT), and the discrete wavelet transform (DFT) to the raw iEEG signal, and then the transform coefficients are shuffled across multiple epochs (or segments). Zhao et al. constructed an efficient and sophisticated method for the data augmentation based on the DCT [54] and applied to the Bern-Barcelona dataset [65], which is a well-known dataset



consisting of epileptic iEEG signal at focal and non-focal locations. The steps to increase the number of samples are summarized as follows:

1. Randomly choose seven iEEG signals from the dataset and apply the DCT to obtain the spectrum.
2. Segment the spectrum into the seven physiological frequency bands (Delta: 0–4 Hz, Theta: 4–8 Hz, Alpha: 8–13 Hz, Beta: 13–30 Hz, Gamma: 30–80 Hz, Ripple: 80–150 Hz, and Fast Ripple: 150–Nyquist Hz), extract one frequency band of each of the decompositions, from lowest to highest frequencies, and merge the seven extracted components (frequency bands) to create a new artificial spectrum. For example, we can extract the delta, the theta, the alpha, the gamma, the ripple, and the fast ripple from the first, the second, the third, the fourth, the fifth, the sixth, and the seventh signal, respectively.
3. Apply the inverse DCT to the artificial spectrum in the frequency domain to obtain an artificial signal in the time-domain.

This is illustrated in Figure 4. The above procedure should be applied to focal and non-focal signals separately. Since this approach to the data augmentation increases the number of samples in the signal-domain, it may be suitable for deep learning-based techniques. In the work [54], the authors successfully applied this data augmentation for a convolutional neural network to identify the focal signals. The data augmentation strategy demonstrated to be useful: using 20% of available real data plus artificial data, the classifier was able to improve its accuracy results to 83.91%, compared to 81.52% obtained using only real data.



**Figure 4.** An artificial signal generation with the DCT. (a) Seven intracranial iEEG signals at either focal or non-focal area. (b) DCT coefficients in the spectrum-domain. The spectra are segmented into seven physiological sub-bands, and the sub-band components extracted from each spectrum are merged to create an artificial spectrum. (c) The inverse DCT leads to the resulting artificial signal.

The other approach to data augmentation is a method in the feature-domain. Most of the classification methods for epileptic EEG signals typically extract features from the raw signals. In general, a conventional yet effective classifier, such as an SVM, requires fewer samples than a deep neural network. Thus, increasing the number of samples in the feature-domain is a strategy to balance positive and negative samples. In the case of epileptic iEEG data, positive (focal) samples are much fewer than negative samples, and this imbalance data can deteriorate the performance of classification accuracy. Akter et al. [56] successfully applied data augmentation based on an adaptive synthetic oversampling approach (ADASYN) [66] to balance an in-hospital dataset of epileptic interictal iEEG signals to identify the seizure onset zones automatically.

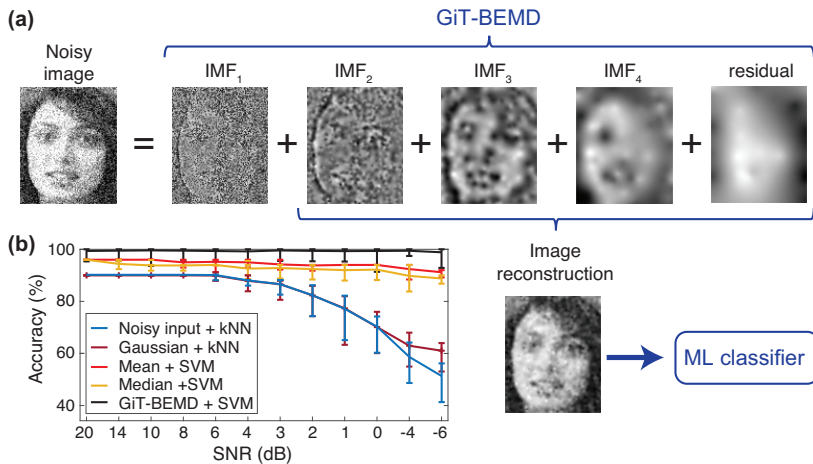


### 3.2. Classification of Noisy Faces

Working with noisy data is a challenge and generates many problems when developing classification systems. Denoising algorithms have to be applied before deriving the classifier and /or when using it. This can be the case in image processing, in which noise can corrupt the image. In an image recognition or verification system, noise could make the task difficult inducing more errors. This is why robust to noise systems are needed. To this end, several strategies can be implemented, for example using classical filters such as a Gaussian [67], bilateral [68], arithmetic [69], median [70] or Wiener filters [71]. However, all of these filters are good in some situations or type of noise, but not good enough in others.

To overcome that issue, a denoising technique based on an empirical mode decomposition with Green’s functions was proposed in [55]. The system uses the capability of the bi-dimensional EMD decomposition to capture (almost all of) the noise in the first IMFs. Therefore, the noisy images are decomposed using a bi-dimensional EMD algorithm, then the first modes are eliminated and the rest of the modes are summed up together to recover the remaining image, almost without noise. In this specific work, the bi-dimensional EMD algorithm was the Green’s function in tension BEMD (GiT-BEMD) which uses Green’s functions to interpolate the surface of the images [72].

The method used in [55] is depicted in Figure 5, in which we can see that the image is decomposed by means of the GiT-BEMD algorithm, the first IMF is discarded, the image is later reconstructed without the contribution of the noise and used to feed a classifier.



**Figure 5.** (a) The new proposed approach to eliminate the noise and improve the classification accuracy is based on the GiT-BEMD decomposition. The high frequency IMFs are discarded and the (noiseless) image is reconstructed by summing up the rest of the modes. This is the image that will feed the classifier. (b) Comparison of classification results using a Support Vector Machine (SVM) and K-Nearest Neighbor (kNN) classifiers applied to noisy, filtered faces (Gaussian, Mean, Median) and GiT-BEMD processed faces.

Experiments were carried out using several type of noise (Gaussian, Uniform, Laplacian and Speckle) and at several levels (SNR from 20<sub>dB</sub> to -6<sub>dB</sub>). The classification accuracy was compared to the ones obtained with classical filters (the ones named before). Classical filters were able to keep good performance when  $SNR > 3_{dB}$ , but they started to fail in much noisy scenarios, corresponding to  $3_{dB} > SNR > -6_{dB}$ . The proposed method was able to maintain the same level of performance in all the ranges, from almost noise-free images ( $SNR = 20_{dB}$ ) to highly noisy images ( $SNR = -6_{dB}$ ). Similar results occurred in the verification case, in which the Equal Error Rate (EER) were reported to

evaluate the results. Independently of the type of classifier used (SVM or k-NN), GiT-BEMD method obtained the lowest EER.

Finally, it is interesting to note that the method is efficient for any type of noise and under high levels of it, and is transparent to the user, hence simple to apply. There are no parameters to tune because the GiT-BEMD algorithm is data-driven, and the IMFs of the images are obtained automatically. Bidimensional EMD algorithms, and specially GiT-BEMD algorithm, are able to decompose images in several IMFs, capturing the noise in the first IMF, which allows for training a classification system that is robust to noisy scenarios.

### 3.3. Scada Data Completion in Water Networks

In industrial applications, usually having a Supervisory Control And Data Acquisition (SCADA) system collecting data, is habitual to have missing values due to several problems (sensor failures, communication loss, etc.). In this scenario, tensor completion algorithms can be used to reconstruct missing data providing a better alternative compared to classical interpolation methods. Because of the cyclic behaviour of data consumption, a tensor structure can be defined considering time scales (days and weeks). The redundancy present into the data is exploited by the tensor completion algorithms, allowing to recover missing data with greater accuracy outperforming classical interpolation or filtering methods.

Drinking water network distribution enterprises usually have a SCADA system which manages the information collected by flow-meters, manometers, level sensors, valves, pumps, etc. By accessing to this centralized dataset, they are able to manage and optimize the operation of the water network.

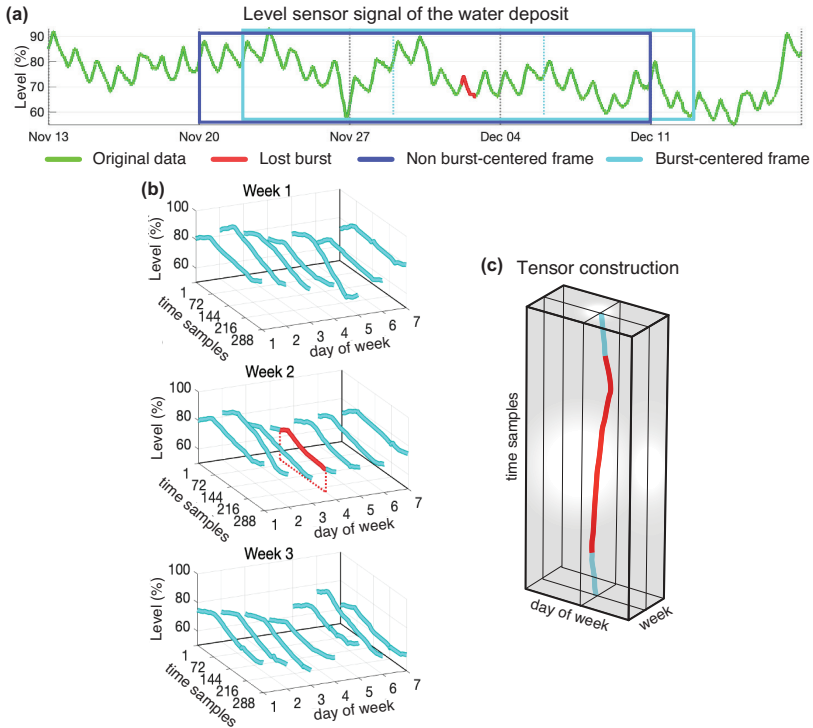
One of the major problems involved in managing long term SCADA data is to deal with the loss of bursts of data due to sensor failures, sensor re-calibrations, or communication failures that take time to be repaired and therefore cause the loss of entire bursts of data. Completing the data lost in bursts remains a difficult task, and most data completion methods that work fairly well when data are lost more or less evenly distributed over time, collapse in that situation.

Water network data completion was investigated in [61,62]. For this application, authors used data coming from the drinking water network distribution enterprise named Aigües de Vic S.A. (AVSA) which is the responsible for the water supply of the city of Vic, in Catalonia, where the heterogeneous data from the SCADA system are stored every 5 min in Structured Query Language (SQL) databases. To explore how tensor completion methods would behave in this scenario, a block of 77 consecutive weeks of data was selected. Then, some parts of the data were deleted to simulate missing burst. The Mean Square Error (MSE) per sample was used as a measure to check the accuracy of the reconstructed data.

In [61], besides of taking advantage of the classical methods, an additional improvement was achieved by performing a tensorization of the data and applying tensor decomposition to recover missing data. This tensorization operation is illustrated in Figure 6.

In [62], a new approach was developed that improved the performance of the previous method by performing two concatenated tensor decompositions. The new approach has several steps. The first step, consisted of applying a smoothing process to the signal to avoid oscillations between adjacent discrete values eventually produced around the point of quantification. The second step consisted of using a very rough imputation method (linear interpolation). After this operation, no empty values were present. Then tensorization that places the original burst positions precisely in the central positions, the burst centered tensorization, was performed (Figure 6a). The tensor obtained was used to make a low-rank tensor model to capture the background trend of the missing burst. The data of the low-rank model in the burst positions obtained was offset corrected and burst-centred again in a new tensor which was used to find a most refined model by employing more modes in a second tensor approximation. The samples  $x_i$  occupying the positions of the burst were retrieved, offset corrected and became the output of the algorithm. The concatenated decompositions that optimize the results independently of the length of the burst to recover were the Tucker(4,6,1) followed by the

Tucker(4,7,7) when using the Tucker model, and the CP(1) followed by the CP(15) when using the CANDECOMP/PARAFAC model. Table 3 shows a comparison between the best classical algorithm tested, the one based on the combination of forward and backward predictors, the singleDecomp [61] and the doubleDecomp [62] algorithms. TK (Tucker) and CP (CANDECOMP/PARAFAC) indicate the decomposition model used by the algorithms. The comparison is carried out considering two tensor dimensions and two lengths of lost bursts.



**Figure 6.** Data tensorization of a 3 week tensor with 200 samples of lost data bursts. In (a) the green line shows the original data, and the red line shows the lost burst. The soft blue window shows the data introduced in burst-centered tensor, which forces the burst to be in the center of the window. Panels (b) shows how the continuous flow of data in the soft blue window is fragmented to be allocated in the tensor as shown in panel (c).

**Table 3.** Algorithms’ performance in terms of the MSE per sample.

Method	Weeks	MSE/Sample	
		Burst Length = 100	Burst Length = 200
Forward & Backward Predictors	-	1.11	2.23
SingleDecomp—CP	3	0.87	1.78
SingleDecomp—CP	7	0.80	1.58
SingleDecomp—TK	3	0.80	1.43
SingleDecomp—TK	7	0.71	1.28
DoubleDecomp—CP	3	0.55	1.05
DoubleDecomp—CP	7	0.52	1.02
DoubleDecomp—TK	3	0.55	1.04
DoubleDecomp—TK	7	<b>0.50</b>	<b>0.97</b>

Best results are indicated in bold text.

#### 4. Conclusions and Discussion

Machine learning methods were typically designed by assuming that training data is perfect and of infinite size. However, in real life, machine learning practitioners need to deal with imperfect training data or datasets of limited size. To alleviate the problem of incomplete/corrupted datasets or to increase the size of a training dataset, it is necessary to use powerful data models that can capture the essential features of the dataset. In this article, a unifying introduction to signal decomposition methods is presented, which are available in different flavors but sharing a common property: every signal in a dataset (sample) can be written as a linear combination of elementary, simpler components.

The main idea behind the reviewed decomposition models is that they are able to be learned from a limited or low-quality dataset. Once the right model for the dataset of interest is learned from the available samples, one can use it for different tasks: (1) to complete missing entries in data samples, (2) to compensate distortions or eliminate noise in data samples, and (3) to artificially create class-preserving new data samples.

We have demonstrated that low-rank, sparse coding and EMD decomposition methods are excellent candidates for models that can capture essential information of a dataset. All these methods can be applied to vector as well as to tensor datasets. However, some issues remain and there is space for improvement in future research. For example:

- The decomposition methods reviewed in this work for imputation of missing/corrupted values do not exploit the class label information in a supervised learning scenario. A possible further improvement of current methods is to incorporate label information into the decomposition models. We believe that missing data values could be better recovered if the class label of the corresponding data sample is known.
- EMD based data augmentation was developed in an ad-hoc fashion. We believe that more theoretical insights could be explored allowing future improvements, for example, by re-designing the way that IMFs are calculated in order to produce class-preserving artificial samples.

This review article illustrates the application of a variety decomposition methods to vector and tensor datasets in a wide range of technological areas including: classification of brain signals (EEG and iEEG), identification of face images and analysis of water network data. While this represents a non-exhaustive review of existing methods and applications of machine learning with low-quality datasets, we believe that it can be a useful reference for machine learning practitioners who are normally faced with incomplete, noisy or small datasets, and can inspire new methods to address these fundamental problems.

**Author Contributions:** Conceptualization, C.F.C., J.S.-C. and S.Z.; investigation C.F.C., J.S.-C., P.M.-P., S.Z. and T.T.; writing—Original draft preparation C.F.C., J.S.-C., P.M.-P. and T.T. All authors have read and agreed to the published version of the manuscript.

**Funding:** This research was funded in part by JST CREST Grant Number JPMJCR1784 and by the University of Vic—Central University of Catalonia (ref. R0947). J.S.-C. work is also based upon work from COST Action CA18106, supported by COST (European Cooperation in Science and Technology).

**Acknowledgments:** We are grateful to the anonymous reviewers for their valuable comments, which helped us to improve the first version of this manuscript.

**Conflicts of Interest:** The authors declare no conflict of interest.

#### References

1. Goodfellow, I.; Bengio, Y.; Courville, A. *Deep Learning*; MIT Press: Cambridge, MA, USA, 2016.
2. Harari, Y.N. Reboot for the AI revolution. *Nat. Publ. Group* **2017**, *550*, 324–327. [[CrossRef](#)]
3. Fatourech, M.; Bashashati, A.; Ward, R.K.; Birch, G.E. EMG and EOG artifacts in brain computer interface systems: A survey. *Clin. Neurophysiol.* **2007**, *118*, 480–494. [[CrossRef](#)]

4. Yu, J.; Lin, Z.; Yang, J.; Shen, X.; Lu, X.; Huang, T.S. Generative Image Inpainting With Contextual Attention. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition CVPR, Salt Lake City, UT, USA, 18–23 June 2018; pp. 5505–5514.
5. Zhang, M.; Chen, Y. Inductive Matrix Completion Based on Graph Neural Networks. In Proceedings of the International Conference on Learning Representations (ICLR), Addis Ababa, Ethiopia, 4–7 May 2020.
6. Mirkes, E.M.; Coats, T.J.; Levesley, J.; Gorban, A.N. Handling missing data in large healthcare dataset: A case study of unknown trauma outcomes. *Comput. Biol. Med.* **2016**, *75*, 203–216. [[CrossRef](#)] [[PubMed](#)]
7. Schölkopf, B.; Burges, C.; Vapnik, V. Incorporating Invariances in Support Vector Learning Machines. *ICANN 1996*, *1112*, 47–52.
8. Decoste, D.; Schölkopf, B. Training Invariant Support Vector Machines. *Mach. Learn.* **2002**, *46*, 161–190. [[CrossRef](#)]
9. Cireşan, D.C.; Meier, U.; Gambardella, L.M.; Schmidhuber, J. Deep, big, simple neural nets for handwritten digit recognition. *Neural Comput.* **2010**, *22*, 3207–3220. [[CrossRef](#)]
10. Dosovitskiy, A.; Fischer, P.; Springenberg, J.T.; Riedmiller, M.; Brox, T. Discriminative Unsupervised Feature Learning with Exemplar Convolutional Neural Networks. *IEEE Trans. Pattern Anal. Mach. Intell.* **2015**, *38*, 1734–1747. [[CrossRef](#)] [[PubMed](#)]
11. Ratner, A.J.; Ehrenberg, H.R.; Hussain, Z.; Dunmon, J.; Ré, C. Learning to Compose Domain-Specific Transformations for Data Augmentation. *Adv. Neural Inf. Process. Syst.* **2017**, *30*, 3239–3249.
12. Uhlich, S.; Porcu, M.; Giron, F.; Enekl, M.; Kemp, T.; Takahashi, N.; Mitsufuji, Y. Improving music source separation based on deep neural networks through data augmentation and network blending. In Proceedings of the 2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), New Orleans, LA, USA, 5–9 March 2017; pp. 261–265.
13. Lee, M.B.; Kim, Y.H.; Park, K.R. Conditional Generative Adversarial Network- Based Data Augmentation for Enhancement of Iris Recognition Accuracy. *IEEE Access* **2019**, *7*, 122134–122152. [[CrossRef](#)]
14. Hu, T.; Tang, T.; Chen, M. Data Simulation by Resampling—A Practical Data Augmentation Algorithm for Periodical Signal Analysis-Based Fault Diagnosis. *IEEE Access* **2016**, *7*, 125133–125145. [[CrossRef](#)]
15. Xie, F.; Wen, H.; Wu, J.; Hou, W.; Song, H.; Zhang, T.; Liao, R.; Jiang, Y. Data Augmentation for Radio Frequency Fingerprinting via Pseudo-Random Integration. *IEEE Trans. Emerg. Top. Comput. Intell.* **2019**, *4*, 1–11. [[CrossRef](#)]
16. Ding, J.; Chen, B.; Liu, H.; Huang, M. Convolutional Neural Network With Data Augmentation for SAR Target Recognition. *IEEE Geosci. Remote Sens. Lett.* **2016**, *13*, 1–5. [[CrossRef](#)]
17. Dao, T.; Gu, A.; Ratner, A.; Smith, V.; De Sa, C.; Ré, C. A Kernel Theory of Modern Data Augmentation. *Proc. Mach. Learn. Res.* **2019**, *97*, 1528–1537. [[PubMed](#)]
18. García-Laencina, P.J.; Sancho-Gómez, J.L.; Figueiras-Vidal, A.R. Pattern classification with missing data: A review. *Neural Comput. Appl.* **2009**, *19*, 263–282. [[CrossRef](#)]
19. Little, R.J.A.; Rubin, D.B. *Stat. Anal. Missing Data*; John Wiley & Sons: Hoboken, NJ, USA, 2014.
20. Batista, G.E.A.P.A.; Monard, M.C. A Study of K-Nearest Neighbour as an Imputation Method. *Hybrid Intell. Syst.* **2002**, *30*, 251–260.
21. Fessant, F.; Midenet, S. Self-Organising Map for Data Imputation and Correction in Surveys. *Neural Comput. Appl.* **2002**, *10*, 300–310. [[CrossRef](#)]
22. Yoon, S.Y.; Lee, S.Y. Training algorithm with incomplete data for feed-forward training neural networks. *Neural Process. Lett.* **1999**, *10*, 171–179. [[CrossRef](#)]
23. Bengio, Y.; Gingras, F. Recurrent Neural Networks for Missing or Asynchronous Data. *Adv. Neural Inf. Process. Syst.* **1995**, *8*, 395–401.
24. Ghahramani, Z.; Jordan, M.I. Supervised learning from incomplete data via an EM approach. *Adv. Neural Inf. Process. Syst.* **1994**, *6*, 120–127.
25. Goldberg, A.B.; Zhu, X.; Recht, B.; Xu, J.M.; Nowak, R.D. Transduction with Matrix Completion—Three Birds with One Stone. *Adv. Neural Inf. Process. Syst.* **2010**, *23*, 757–765.
26. Hazan, E.; Livni, R.; Mansour, Y. Classification with Low Rank and Missing Data. In Proceedings of the 32nd International Conference on Machine Learning, Lille, France, 6–11 July 2015.
27. Huang, S.J.; Xu, M.; Xie, M.K.; Sugiyama, M.; Niu, Z.; Chen, S. Active Feature Acquisition with Supervised Matrix Completion. *arXiv* **2018**, arXiv:1802.05380.

28. Smieja, M.; Struski, L.; Tabor, J.; Zielinski, B.; Spurek, P. Processing of missing data by neural networks. *Adv. Neural Inf. Process. Syst.* **2018**, *31*, 2719–2729.
29. S, K.P.F.R. LIII. On lines and planes of closest fit to systems of points in space. *Lond. Edinb. Dublin Philos. Mag. J. Sci.* **1901**, *2*, 559–572.
30. Karhunen, K. *Über lineare Methoden in der Wahrscheinlichkeitsrechnung*; Annales Academiae Scientiarum: Helsinki, Sana, 1947.
31. Loève, M. *Probability Theory*; Van Nostrand: Princeton, NJ, USA, 1963.
32. Bruckstein, A.M.; Donoho, D.L.; Elad, M. From Sparse Solutions of Systems of Equations to Sparse Modeling of Signals and Images. *SIAM Rev.* **2009**, *51*, 34–81. [[CrossRef](#)]
33. Elad, M.; Figueiredo, M.A.T.; Ma, Y. On the Role of Sparse and Redundant Representations in Image Processing. *Proc. IEEE* **2010**, *98*, 972–982. [[CrossRef](#)]
34. Davis, G.M.; Mallat, S.G.; Zhang, Z. Adaptive Time-frequency Decompositions. *Opt. Eng.* **1994**, *33*, 2183.
35. Tropp, J.A.; Gilbert, A.C. Signal Recovery From Random Measurements Via Orthogonal Matching Pursuit. *Inst. Electr. Electron. Eng. Trans. Inf. Theory* **2007**, *53*, 4655–4666. [[CrossRef](#)]
36. Needell, D.; Tropp, J. CoSaMP: Iterative Signal Recovery from Incomplete and Inaccurate Samples. *Appl. Comput. Harmon. Anal.* **2009**, *26*, 301–321. [[CrossRef](#)]
37. Chen, S.; Donoho, D.; Saunders, M. Atomic Decomposition by Basis Pursuit. *SIAM Rev.* **2001**, *43*, 129–159. [[CrossRef](#)]
38. Tropp, J.A.; Wright, S.J. Computational Methods for Sparse Solution of Linear Inverse Problems. *Proc. IEEE* **2010**, *98*, 948–958. [[CrossRef](#)]
39. Elad, M.; Aharon, M. Image Denoising Via Sparse and Redundant Representations Over Learned Dictionaries. *Image Process. IEEE Trans.* **2006**, *15*, 3736–3745. [[CrossRef](#)] [[PubMed](#)]
40. Mairal, J.; Bach, F.R.; Ponce, J.; Sapiro, G. Online Dictionary Learning for Sparse Coding. In Proceedings of the 26th Annual International Conference on Machine Learning (ICML), Montreal, QC, Canada, 14–18 June 2009; pp. 689–696.
41. Donoho, D.L. Compressed sensing. *Inst. Electr. Electron. Eng. Trans. Inf. Theory* **2006**, *52*, 1289–1306. [[CrossRef](#)]
42. Candès, E.; Wakin, M. An Introduction to Compressive Sampling. *Signal Process. Mag. IEEE* **2008**, *25*, 21–30. [[CrossRef](#)]
43. Bobin, J.; Starck, J.L.; Fadili, J.; Moudden, Y. Sparsity and Morphological Diversity in Blind Source Separation. *Image Process. IEEE Trans.* **2007**, *16*, 2662–2674. [[CrossRef](#)]
44. Elad, M.; Starck, J.L.; Querre, P.; Donoho, D.L. Simultaneous cartoon and texture image inpainting using morphological component analysis (MCA). *Appl. Comput. Harmon. Anal.* **2005**, *19*, 340–358. [[CrossRef](#)]
45. Huang, N.E.; Shen, Z.; Long, S.R.; Wu, M.C.; Shih, H.H.; Zheng, Q.; Yen, N.C.; Tung, C.C.; Liu, H.H. *The Empirical Mode Decomposition and the Hilbert Spectrum for Nonlinear and Non-Stationary Time Series Analysis*; Royal Society of London Proceedings Series A; The Royal Society: London, UK, 1998; pp. 903–998.
46. Tucker, L.R. Some mathematical notes on three-mode factor analysis. *Psychometrika* **1966**, *31*, 279–311. [[CrossRef](#)]
47. Kolda, T.; Bader, B. Tensor decompositions and applications. *SIAM Rev.* **2009**, *51*, 455–500. [[CrossRef](#)]
48. Cichocki, A.; Mandic, D.; De Lathauwer, L.; Zhou, G.; Zhao, Q.; Caiafa, C.; Phan, A.H. Tensor decompositions for signal processing applications: from two-way to multiway component analysis. *IEEE Signal Process. Mag.* **2015**, *32*, 145–163. [[CrossRef](#)]
49. Caiafa, C.F.; Cichocki, A. Computing sparse representations of multidimensional signals using Kronecker bases. *Neural Comput.* **2013**, *25*, 186–220. [[CrossRef](#)]
50. Caiafa, C.F.; Cichocki, A. Multidimensional compressed sensing and their applications. *Wiley Interdiscip. Rev. Data Min. Knowl. Discov.* **2013**, *3*, 355–380. [[CrossRef](#)]
51. Huang, J.; Zhou, G.; Yu, G. Orthogonal tensor dictionary learning for accelerated dynamic MRI. *Med. Biol. Eng. Comput.* **2019**, *57*, 1933–1946. [[CrossRef](#)] [[PubMed](#)]
52. Dinares-Ferran, J.; Ortner, R.; Guger, C.; Solé-Casals, J. A New Method to Generate Artificial Frames Using the Empirical Mode Decomposition for an EEG-Based Motor Imagery BCI. *Front. Neurosci.* **2018**, *12*, 1–9. [[CrossRef](#)] [[PubMed](#)]



53. Zhang, Z.; Duan, F.; Solé-Casals, J.; Dinares-Ferran, J.; Cichocki, A.; Yang, Z.; Sun, Z. A Novel Deep Learning Approach With Data Augmentation to Classify Motor Imagery Signals. *IEEE Access* **2019**, *7*, 15945–15954. [[CrossRef](#)]
54. Classification of Epileptic IEEG Signals by CNN and Data Augmentation. In Proceedings of the ICASSP 2020—2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Toronto, ON, Canada, 6–11 June 2020; pp. 926–930.
55. Al-Baddai, S.; Marti-Puig, P.; Gallego-Jutglà, E.; Al-Subari, K.; Tomé, A.M.; Ludwig, B.; Lang, E.W.; Solé-Casals, J. A recognition-verification system for noisy faces based on an empirical mode decomposition with Green's functions. *Soft Comput.* **2019**, *24*, 3809–3827. [[CrossRef](#)]
56. Akter, M.S.; Islam, M.R.; Iimura, Y.; Sugano, H.; Fukumori, K.; Wang, D.; Tanaka, T.; Cichocki, A. Multiband entropy-based feature-extraction method for automatic identification of epileptic focus based on high-frequency components in interictal iEEG. *Sci. Rep.* **2020**, *10*, 7044. [[CrossRef](#)] [[PubMed](#)]
57. Solé-Casals, J.; Caiafa, C.F.; Zhao, Q.; Cichocki, A. Brain-Computer Interface with Corrupted EEG Data: A Tensor Completion Approach. *Cogn. Comput.* **2018**, *10*, 1062–1074. [[CrossRef](#)]
58. Acar, E.; Dunlavy, D.M.; Kolda, T.G.; Mørup, M. Scalable tensor factorizations for incomplete data. *Chemom. Intell. Lab. Syst.* **2011**, *106*, 41–56. [[CrossRef](#)]
59. Liu, J.; Musialski, P.; Wonka, P.; Ye, J. Tensor Completion for Estimating Missing Values in Visual Data. *IEEE Trans. Pattern Anal. Mach. Intell.* **2012**, *35*, 208–220. [[CrossRef](#)]
60. Zhao, Q.; Zhang, L.; Cichocki, A. Bayesian CP Factorization of Incomplete Tensors with Automatic Rank Determination. *IEEE Trans. Pattern Anal. Mach. Intell.* **2015**, *37*, 1751–1763. [[CrossRef](#)]
61. Marti-Puig, P.; Martí-Sarri, A.; Serra-Serra, M. Different Approaches to SCADA Data Completion in Water Networks. *Water* **2019**, *11*, 1023. [[CrossRef](#)]
62. Marti-Puig, P.; Martí-Sarri, A.; Serra-Serra, M. Double Tensor-Decomposition for SCADA Data Completion in Water Networks. *Water* **2020**, *12*, 80. [[CrossRef](#)]
63. Ramoser, H.; Müller-Gerking, J.; Pfurtscheller, G. Optimal spatial filtering of single trial EEG during imagined hand movement. *IEEE Trans. Rehabil. Eng. Publ. IEEE Eng. Med. Biol. Soc.* **2000**, *8*, 441–446. [[CrossRef](#)] [[PubMed](#)]
64. Leys, C.; Ley, C.; Klein, O.; Bernard, P.; Licata, L. Detecting outliers: Do not use standard deviation around the mean, use absolute deviation around the median. *J. Exp. Soc. Psychol.* **2013**, *49*, 764–766. [[CrossRef](#)]
65. Andrzejak, R.G.; Schindler, K.; Rummel, C. Nonrandomness, nonlinear dependence, and nonstationarity of electroencephalographic recordings from epilepsy patients. *Phys. Rev. E* **2012**, *86*, 046206. [[CrossRef](#)]
66. Haibo, H.; Yang, B.; Garcia, E.A.; Shutao, L. ADASYN: Adaptive synthetic sampling approach for imbalanced learning. In Proceedings of the 2008 IEEE International Conference on Neural Networks (IEEE World Congress on Computational Intelligence), Hong Kong, China, 1–8 June 2008; pp. 1322–1328.
67. Liu, X.; Tanaka, M.; Okutomi, M. Single-Image Noise Level Estimation for Blind Denoising. *Image Process. IEEE Trans.* **2013**, *22*, 5226–5237. [[CrossRef](#)]
68. Tomasi, C.; Manduchi, R. Bilateral filtering for gray and color images. In Proceedings of the Sixth International Conference on Computer Vision, Washington, DC, USA, 4–7 January 1998; pp. 839–846.
69. Gonzalez, R.C.; Woods, R.E. *Digital Image Processing*; Prentice Hall: Upper Saddle River, NJ, USA, 2008.
70. Tukey, J.W. *Exploratory Data Analysis*; Addison-Wesley: Reading, MA, USA, 1977.
71. Lim, J.S. *Two-Dimensional Signal and Image Processing*; Prentice-Hall, Inc.: Upper Saddle River, NJ, USA, 1990.
72. Al-Baddai, S.; Al-Subari, K.; Tom, A.M.; Casals, J.S.; Lang, E.W. A Green's Function-Based Bi-Dimensional Empirical Mode Decomposition. *Inf. Sci.* **2016**, *348*, 1–17. [[CrossRef](#)]

**Publisher's Note:** MDPI stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.



© 2020 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<http://creativecommons.org/licenses/by/4.0/>).

Article

# Severity Classification of Parkinson's Disease Based on Permutation-Variable Importance and Persistent Entropy

Jigang Tong <sup>1,\*</sup>, Jiachen Zhang <sup>1</sup>, Enzeng Dong <sup>1</sup> and Shengzhi Du <sup>2</sup>

<sup>1</sup> Tianjin Key Laboratory for Control Theory & Applications in Complicated Systems, Tianjin University of Technology, Tianjin 300384, China; zhangjiachen1996@gmail.com (J.Z.); dongenzeng@tjut.edu.cn (E.D.)

<sup>2</sup> Department of Electrical Engineering, Tshwane University of Technology, Pretoria 0001, South Africa; dushengzhi@gmail.com

\* Correspondence: tjgtjut@gmail.com; Tel.: +86-13920279916

**Abstract:** Parkinson's disease (PD) is a neurodegenerative disease that causes chronic and progressive motor dysfunction. As PD progresses, patients show different symptoms at different stages of the disease. The severity assessment is inefficient and subjective when it comes to artificial diagnosis. However, abnormal gait was contingent and the subject selection was limited. Therefore, few-shot learning based on small sample sets is critical to solving the problem of insufficient sample data in PD patients. Using datasets from PhysioNet, this paper presents a method based on permutation-variable importance (PVI) and persistent entropy of topological imprints, and uses support vector machine (SVM) as a classifier to achieve the severity classification of PD patients. The method includes the following steps: (1) Take the data as gait cycles, and calculate the gait characteristics of each cycle. (2) Use the random forest (RF) method to obtain the leading factors differentiating the gait of patients at different severity levels. (3) Use time-delay embedding to map the data into a topological space, and use the topological data analysis based on permutation homology to obtain the persistent entropy. (4) Use the Borderline-SMOTE (BSM) method to balance the sample data. (5) Use the SVM to classify the samples for the severity levels of PD. An accuracy of 98.08% was achieved by 10-fold cross-validation, so our method can be used as an effective means of computer-aided diagnosis of PD, and has important practical value.

**Keywords:** Parkinson's disease; few-shot learning; permutation-variable importance; topological data analysis; persistent entropy; support-vector machine



**Citation:** Tong, J.; Zhang, J.; Dong, E.; Du, S. Severity Classification of Parkinson's Disease Based on Permutation-Variable Importance and Persistent Entropy. *Appl. Sci.* **2021**, *11*, 1834. <https://doi.org/10.3390/app11041834>

Academic Editor: Jordi Solé-Casals

Received: 16 January 2021  
Accepted: 17 February 2021  
Published: 19 February 2021

**Publisher's Note:** MDPI stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.



**Copyright:** © 2021 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

## 1. Introduction

Parkinson's disease (PD) is a common neurodegenerative disease characterized by the loss of dopamine in neurons in the brain, resulting in a series of complex network dysfunctions [1]. Such dysfunctions may cause significant effects on the gait of patients, such as an unstable walking posture, bradykinesia, tremor dominance, frequent falling, panic gait, and freezing of gait [2]. The onset of PD is a gradual process; in the progression of the disease, clinical patients show different severity. For PD patients with different severities of the disease, there are different means of treatment, so the severity evaluation can greatly strengthen the clinical management of patients by giving the targeted treatment. Currently, the most common PD rating criterion is the Hoehn and Yahr (HY) grading system [3], which divides the severity of PD into five levels (1 to 5, increasing in severity). However, the HY grading evaluation relies heavily on medical experts with specialized knowledge and clinical experiences, which is a time-consuming and low-efficiency process, and inevitably has a certain subjective judgment. Therefore, auxiliary means to assess PD severity is needed to improve the rating efficiency and reduce costs.

With the development of wearable sensing technology, gait-analysis technology based on human sensing data is being increasingly applied in the detection of PD. Among them, the ground reaction force (GRF) is widely used in PD symptom analysis as a common



quantitative measurement method for gait assessment [4,5]. As an important indicator of joint movement and muscle activity, the GRF during walking can be obtained by wearable insole sensors, which can well reflect the characteristics of an abnormal gait. These sensors have the advantages of small size, low cost, non-invasiveness, a wide range of application scenarios, and low energy consumption. Muniz et al. [6] used the GRF to evaluate the impacts on PD from the deep brain stimulation of the subthalamic nucleus (DBS-STU) and drug therapy. Petrucci et al. [7] studied the freezing of gait in patients with PD prediction and adjustment, in which the GRF was used as the main evaluation index and the auxiliary effect of ankle orthotics was observed. The ankle orthotics embodied in patients after auxiliary GRF captured significantly lower average vibration amplitude, which indicates that the seriousness of the PD patients is closely related to the GRF. In addition, using musculoskeletal modeling driven by depth sensors, Jeonghoon Oh et al. [8] compared the GRF among patients and healthy people, and found significant differences in the early peaks of the GRF. A large number of studies have shown that the abnormal gait patterns of PD patients are reflected in the GRF during walking, which can be an important feature of the PD research.

The GRF can well reflect the stability of gait, from which we can analyze and grade the severity of PD. Machine learning can effectively solve the problem of medical data analysis, and it has been widely used in related fields. From the perspective of gait data, many related scholars have applied machine-learning methods to conduct classification studies of PD, such as logistic regression, random forest, extreme gradient boosting, radial basis, and neural network [9–18]. In terms of the PD severity assessment using machine learning, Aite Zhao et al. [19] employed the GRF as gait data to identify the seriousness level using the two-channel method of long short-term memory (LSTM) and convolutional neural network (CNN). Similarly, Wei Zeng et al. [20] used neural networks and other methods for PD severity classification, in which the phase space reconstruction and empirical mode decomposition methods were used in data preprocessing. Balaji E et al. [21] used decision tree (DT), support vector machine (SVM), ensemble classifier (EC), and Bayesian classifier (BC) methods to classify the stages of PD, and achieved an effective evaluation of the severity. In a similar way, Enas Abdulhay et al. [22] and Tun Aurolo et al. [23] achieved effective PD grading by using medium Gaussian SVM and locally weighted random forest methods, respectively, with the shallow-learning method. However, in the above studies, the sample data was small and unbalanced. For instance, in Natasa Kleanthous et al.'s work [9], only 10 PD subjects were involved. In addition, compared with other shallow-learning methods, the deep-learning recognition methods based on neural network showed slightly worse results, with respect to time consumption and effectiveness in distinguishing the similar PD severity levels. The reason for this phenomenon is that deep learning is a supervised learning method based on big data, which relies heavily on a large number of high-quality labeled data. When the data is insufficient, problems such as overfitting occur in the model, thus reducing the recognition rate. Large sample sizes from a limited number of PD patients are associated with high clinical risks, uncontrollable repeatability, and high costs. These reasons make it difficult for many deep-learning models to be applied to PD research. In addition, some gait disorders such as panic gait, short gait, and frozen gait were contingencies in PD patients, which resulted in a small number of negative samples in the body-sensing data. Considering all kinds of factors, the identification of severity level by machine-learning methods is mainly based on small sample datasets, so few-shot learning with a lower sample-data requirement becomes the key to solve the problem.

In addition to few-shot learning, the undersampling and oversampling methods are the common means of balancing a dataset. For a PD disease dataset with too few samples, oversampling is used to expand the dataset, and noise data can be added to enhance the robustness of classifiers. When the model cannot effectively extract features from the existing data, the data can first be processed to find the most important parameters that dominate the differences among classes, or deeper feature analysis of the data can be carried out to make the classification and recognition effect more obvious. Fabienne

Reynard et al. [24] studied the dominant factors of stability during treadmill walking, using the random forest (RF) method [25] to measure the importance of the variables, while relatively insignificant variables were removed, which made the analysis more effective. Enas Abdulhay et al. [22] extracted only step time, stance time swing and footstrike profile from the GRF data to analyze and identify the diseases. In addition, in the study of Yan Yan et al. [26], the GRF data were reconstructed in a phase space, after mapping the data to the high-dimensional phase space for topological motion analysis, to study the gait fluctuation, and the extracted topological features were applied. The random forest (RF) method has shown a good performance in permutation-variable importance (PVI) and is widely used [27]. Therefore, RF ranking of the importance of gait characteristics can be used to obtain the main factors influencing the different severity levels in PD patients.

The walking process has strong nonlinear characteristics and can be regarded as a nonlinear dynamic system. Extracting deeper features of gait can enhance the differentiation of samples, which is beneficial to the classification of machine learning. The common method is to use topological data analysis (TDA) to obtain the topological imprint for further feature extraction [20,26].

In the classification of machine learning, traditional classifiers are commonly designed based on balanced datasets, and the losses of classifiers are biased toward the majority of classes [28]. Therefore, the imbalance of sample data may cause the insensitivity of the learning model to a minority of classes. However, in the study of abnormal gait patterns, usually only a small number of samples are available. In the method of balancing samples, the sampling method is often used to balance data, including oversampling [29], undersampling [30], and mixed sampling. In machine learning with a small sample size, the method of oversampling is usually adopted to balance the datasets. Among the oversampling methods, the synthetic minority oversampling technique (SMOTE) is considered to be the most effective [31]. The SMOTE method balances the number of minority classes by interpolation between the adjacent minority class samples, which increases the number of minority class samples and improves the classifier performance [32]. The Borderline-SMOTE method is used to synthesize new samples with only a few samples on the boundary, which can improve the distribution of the samples. However, during the composition of the minority classes, the SMOTE did not consider the class information of the nearest neighbor sample, which often overlays the sample, resulting in poor classification performance. The Borderline-SMOTE method was proposed to improve this problem [33]. Support vector machine (SVM) was first proposed by Vapnik et al. [34] as a solution to the dichotomy problem of linearly separable samples. In terms of recognizing abnormal gait, the SVM was successfully used in various pattern recognition problems [35]. Compared with other traditional learning models, the SVM has an excellent performance in solving few-shot learning problems [36]. Because it adopts the principle of structural risk minimization [37], the SVM model has strong generalization ability.

This paper addresses the PD severity-level classification with a small sample set. The PD gait dataset from Goldberger on PhysioNet [38] was used to demonstrate the proposed method. The dataset consisted of only 29 PD patients (15, 8, and 6 patients with HY ratings of 2, 2.5, and 3, respectively) and 18 healthy controls. The sample data was very small, so we considered three aspects to solve the small-sample learning problem: data, model, and algorithm [39]. When the training samples are insufficient, the neural network model with the objective of minimizing loss function tends to fit on a small number of samples, which results in low generalization capacity. However, many nonparametric methods do not need to train the optimization parameters, such as the embedded-learning (EL) method [40,41]. EL is a nonparametric method based on a measurement in which the prior knowledge of training set is used as a design source. In EL, the samples are embedded into a low-dimensional space, which makes the samples of different categories in the low-dimensional space easier to distinguish. The embedded data then can be enhanced in the aspects of the discrimination degree and the balance among the sample size of classes to optimize the performance of the learner. The GRF data is divided according to the gait

cycle, and then the categorized data is processed, and a series of gait characteristics is calculated. The variable importance is evaluated for the obtained characteristics by the RF method, and the variable of a bigger impact on the severity classification is reserved for further distinguishing features.

In order to reconstruct the phase space by embedding the obtained gait characteristics with time delay, the data is mapped to the topological space. The topological characteristics of the obtained point-cloud data are analyzed by using the persistent homology methods to obtain the topological signature of the gait data, such as persistent bar code, persistent scatter plot and persistent state plot. However, these topology imprints are challenging to be used as input to machine learning. For this reason, the persistent scattergram topology marks obtained by important gait parameters is calculated as the persistent entropy [42], which is more suitable for machine learning. The SVM is employed for few-shot learning in gait analysis.

In this study, a method based on permutation-variable importance and persistent entropy is proposed for the severity classification of PD. Based on the small dataset of gait, the dominant factors are extracted by permutation-variable importance, and the persistent entropy is proposed to transform the topological imprints into sample inputs more suitable for machine learning. The proposed method can fully improve the degree of differentiation between different disease categories and achieve a favorable effect, and has certain practical significance.

## 2. Materials and Methods

### 2.1. Subjects and Data Set

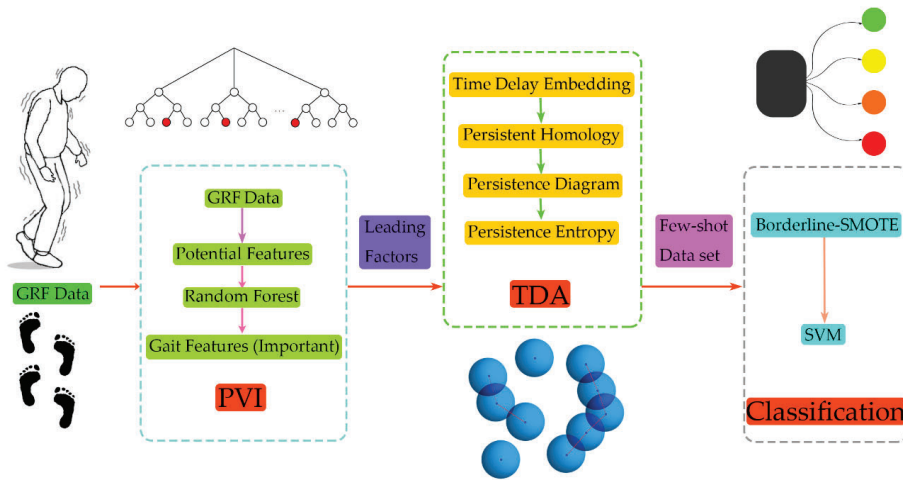
For this study, we use a gait database from PhysioNet provided by Goldberger [38]. The dataset consisted of GRF signals from PD patients and healthy controls. The gait data signals were collected from normal walking and dual-task walking. The normal walking data of patients and the control group was used in this paper for analysis. There were a total of 47 subjects in this dataset, including 29 PD patients and 18 normal controls. Among the 29 PD patients, there were 20 males and 9 females. The normal control group consisted of 10 males and 8 females. The mean ages of the patients and the control groups were 71 and 72, respectively. Among the PD patients, 15 subjects were HY grade 2, 8 were grade 2.5, and 6 were grade 3. Table 1 shows the basic information of the subjects involved in the experiment.

**Table 1.** Subject information.

Group	Number	Male	Female	Age	HY = 2	HY = 2.5	HY = 3
PD	29	20	9	71 ± 8	15	8	6
Co	18	10	8	72 ± 6	-	-	-

### 2.2. Analysis Method

The framework of the proposed method is shown in Figure 1. A total of 47 GRF data (29 PD patients with different disease grades and 18 normal subjects) were used. First, the GRF data were preprocessed, including categorizing subjects according to the gait cycle and calculating the gait characteristics of each gait cycle during walking, then a time series of the gait characteristics was obtained. During the gait-cycle division, the period should be as short as possible on the premise of guaranteeing the complete representation of the gait-cycle information for both the left and right feet. Therefore, we choose two gait cycles as the period and divided them into sections, so that the information in the original signal was completely retained.

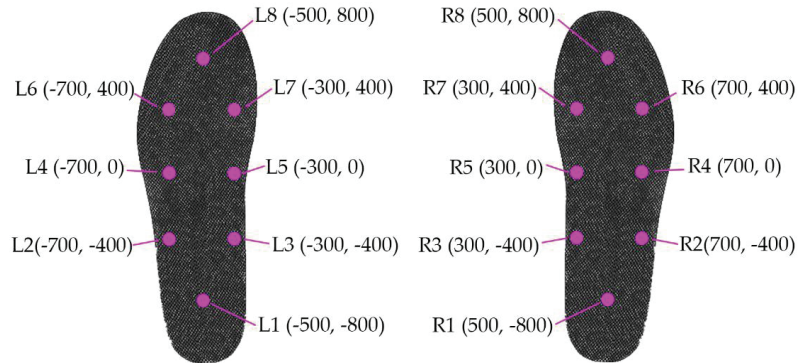


**Figure 1.** The processing framework of this study is divided into three parts: variable importance (PVI) analysis, topological data analysis (TDA), and severity classification. In the analysis of the importance of variables, the GRF data were first categorized according to the data in each gait cycle. The gait characteristics of each cycle were calculated, and the variable importance was ranked to select the most significant ones. In the TDA, phase-space reconstruction was carried out for each gait feature, and a persistent homology method was used to extract topology marks to obtain persistent scatter plots, then the persistent entropy of persistent scatter plots was calculated. In the stage of classification, the Borderline-SMOTE method was used to balance the samples, then the Support Vector Machine (SVM) was used to classify the data and obtain the obfuscation matrix for performance analysis.

Regarding the extraction of gait characteristics on the GRF, we referred to the method on the previous study [43]. In this way, potential characteristics were selected that affected the severity classification, including the coordinates of the center of pressure (CoP), stride time, gait phase, and sample entropy. The random forest method was used to evaluate the importance of these characteristics/variables, and the most significant ones were selected for further analysis. After obtaining the time-series data with a great influence on the difference, the time-delay embedding theorem was used to reconstruct the phase space, and the data were mapped to the phase space to obtain the data point cloud. The topology features of the obtained phase-space-data point cloud were extracted and the persistent entropy was calculated. The Borderline-SMOTE algorithm was used to enhance the data in the training dataset, and the balanced sample data was used as the input to train using SVM to realize the grade recognition of PD.

### 2.3. Data Description

The data recorded were the GRFs when subjects walked for about two minutes on flat ground at a pace of their preference. In the experiments, each subject had 16 force sensors under their feet, with eight sensors under each foot. Thus, we could study stride-to-stride dynamics and the variability of these time series. When a person is comfortable standing with both legs parallel to each other, sensor locations inside the insole can be described approximately in Figure 2, assuming the origin (0,0) is just between the feet, and the person is facing toward the positive Y-axis.



**Figure 2.** The pressure sensors L1–L8 and R1–R8 under the left and right feet, respectively.

The sampling frequency of the force sensors was 100 Hz, and the forces (N) were collected to obtain a time series of pressure data. In addition to the pressure data, two synthetic signals were generated, including the total sum of the pressure under the left and right feet. The resulting data contain 19 columns per row, with column 1 as time (s); columns 2–9 and 10–17 as the GRF (N) of the left and right feet, respectively; and column 18 as the sum of the GRF on the left foot and column 19 as that for the right. These data were used to fit the relationship between the pressure position and time, model the reaction pressure center as a function of time, and obtain the gait features such as stride time, swing time, etc.

#### 2.4. Preprocessing

##### 2.4.1. Data Partitioning

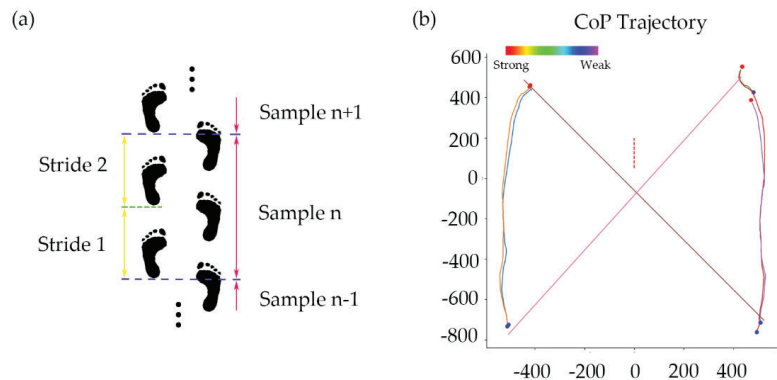
In this study, the dataset contained 16 independent force sensor signals and 2 synthetic pressure signals. The pressure magnitude and the position of a single sensor could not directly reflect the pressure-tracking distribution during the walk alone. To extract the pressure-tracking distribution, the pressure magnitude and position of individual sensors, the total pressure of all sensors are needed. The changing track of the plantar pressure center was calculated as follows.

$$x = \frac{\sum_{i=1}^8 x_i F_i}{F} \tag{1}$$

$$y = \frac{\sum_{i=1}^8 y_i F_i}{F} \tag{2}$$

where  $x_i$  and  $y_i$  are the X-axis and Y-axis coordinates of the  $i$ -th sensor of a foot,  $F_i$  is the force measured by the corresponding sensor, and  $F$  is the sum of the pressures under the foot.

According to the centers of pressure (CoP) obtained in Equations (1) and (2), the entire walking process was divided into two stride cycles. Each cycle began with the first touch of the left heel and ended with the third touch of the same heel (starting the next cycle). This ensured that there was at least one continuous step cycle for each foot. The gait characteristics of each cycle were extracted. The CoP track for each partition is shown in Figure 3. In order to exclude the influence of the unstable features when walking started, the first two stride cycles of each subject were excluded, but the middle 40 dividing cycles and a total of 80 stepping cycles were selected for analysis. The same criteria were applied to each subject to ensure the accuracy of the sampling.



**Figure 3.** (a) Schematic diagram of the period division. (b) Center of Pressure (CoP) path for each partition period. The color represents the pressure.

#### 2.4.2. Gait Features

From the track of CoP, we could further extract gait features for better reflection of the characteristics relevant to walking stability in the PD patients. The selected features were screened for visible differences among classes, which was conducive to the inaccurate identification of a small number of classes in the learning of the small sample dataset, so that it could have a better effect on the classifier training of disease grading. The trajectory of CoP was analyzed using linear and nonlinear analysis methods, and the corresponding gait characteristics are obtained; this could further find the most significant factors and realize more accurate grade identification.

In the calculation of the linear characteristics, we used the root mean square (RMS) of the two stride cycles as the results. The linear indicators we selected are as follows:

1. CoP distribution and its derivatives. The CoP distribution can well reflect the stability of gait and can be used in the analysis of PD. In the analysis of the coordinate distribution of CoP, we selected the RMS of mediolateral direction (X-axis); anterior-posterior (Y-axis) direction; total CoP coordinates; and the RMS of velocity, acceleration and jerk.
2. Gait phase ratio and stride time. In the study of abnormal gait, the proportions of gait time and stride time are usually very important gait characteristics that can clearly reflect the difference between patients and normal subjects. The proportion of the subject's walking gait can be calculated from the pressure and corresponding time. The length of time between the start of a heel touch on one side and the end of the next heel touch on that side is a step time. Among these, the period from the time when one foot heel touches the ground to the time when the toe is off the ground is the supporting phase of this gait cycle, and the difference between the stride time and the duration of the supporting phase is the swinging phase time of this stride cycle. The proportion of the gait time can be obtained by calculating the time duration of the gait and the time of the stride. The gait phase proportion and stride time of the two stride periods of each division period can be obtained by taking the RMS.
3. CoP efficiency and CoP track intersections (CSIP). These two characteristics are also considered, and they may reflect the stability of gait to some extent, and also help us analyze the walking pattern of PD patients. Both of these features can be obtained from the track of CoP. The CoP path efficiency is calculated through dividing direct CoP distance by the actual path that CoP traveled during the stance phase. In the stance phase, the CoP position moves forward under the support foot. When support moves to the other foot, the CoP position moves from one foot to the other [44].

Human walking, as a complex system, has strong nonlinear characteristics. The use of nonlinear analysis method to extract features can effectively analyze the gait characteristics of PD patients. In this study, we chose the sample entropy of CoP as a nonlinear index, which can reflect the degree of disorganization of and attention to walking, and can be used as an important sample input for disease classification and identification.

### 2.5. Permutation-Variable Importance

When using the small sample dataset to classify the severity of the disease, we chose to first calculate some gait characteristics, in order to find out the characteristics that dominated the difference of different categories and improve the discrimination degree of the samples. For the measurement of the importance of variables, this study used the random forest method to evaluate the importance of features. Using this method, the aim was to identify the dominant factors that influence the different manifestations of PD at different severity levels, and to exclude irrelevant characteristics. The measurement of the importance of variables can reduce the dimension of the input sample data. On one hand, it eliminates the influence of irrelevant factors, while on the other hand, it facilitates the subsequent processing of the data. The random forest method can be used to select the characteristics that have the greatest impact on the severity level, so as to reduce the number of features in the model building and make the classifier achieve good results in training. When we use the random forest method to obtain the importance of certain characteristics in disease classification, the specific steps are as follows:

1. For each decision tree, select the corresponding out-of-bag (OOB) data to calculate the out-of-bag data error, which is denoted as  $err1$ .
2. Random noise interference is added to such characteristics of all samples of out-of-bag data, and the out-of-bag data error is calculated again, denoted as  $err2$ .
3. The permutation-variable importance is obtained by Equation (3):

$$PVI = \frac{\sum_{i=1}^N err2_i - err1_i}{N} \quad (3)$$

where  $N$  is the number of decision trees in the random forests,  $err1_i$  is the OOB error of the  $i$ -th decision tree for the feature to be evaluated, and  $err2_i$  is the OOB error of the  $i$ -th decision tree for an assessment feature after noise interference is added to the feature.

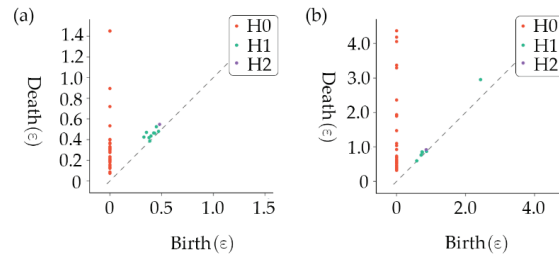
When the random noise is added, the accuracy of data outside the bag will decrease. When this feature is of high importance, the value of OOB error  $err2_i$  will increase significantly, and the calculated measurement value will increase, indicating that this feature has a great impact on the prediction results of disease grade identification, and thus indicates that this feature is of high importance. In this study, we measured the importance of variables in patients with PD and normal subjects. The results of our assessment of the importance of all the features are shown in Figure 4. We also ranked the evaluation results in order of importance in the two cases, calculated the average value of importance in the two cases, and selected the characteristics that rank in the top 15 for importance.







or holes will disappear, which means that these homology structures have a specific duration. We recorded homophones for time of birth and time of death, which we called the persistent homophones, resulting in the topological stamp. A persistence diagram was obtained for each homology structure by plotting a graph with the times of birth and death as the axes, as shown in Figure 5.



**Figure 5.** (a) The control subjects in the stance phase of the left foot (b) The Parkinson’s disease (PD) subjects in the stance phase. The abscissa is the appearance time of the structure, and the ordinate is the disappearance time. H0, H1, and H2 are the homology structures of 0-dimensions, 1-dimension and 2-dimension, respectively.

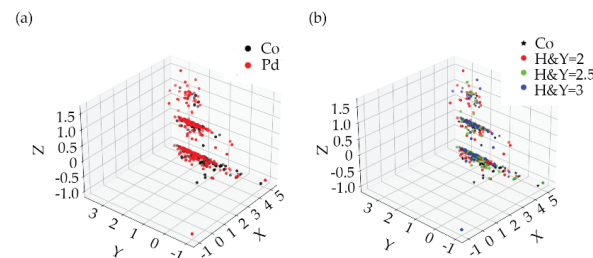
However, there was no machine-learning benefit available from persistent scatter diagrams, so we introduced persistent entropy as a treatment:

$$E(D) = -\sum_{i \in I} p_i \log(p_i) \tag{7}$$

$$p_i = \frac{d_i - b_i}{L_D} \tag{8}$$

$$L_D = \sum_{i \in I} (d_i - b_i) \tag{9}$$

where  $I$  is the set of points in a persistence diagram;  $b_i$  and  $d_i$  are the times of birth and death of the  $i$ -th point, respectively; and  $E(D)$  is the persistent entropy of the persistence diagram  $D$ . The persistent entropy distribution of control subjects and PD subjects is shown in the Figure 6.



**Figure 6.** (a) Persistent entropy between control and PD subjects. (b) Persistent entropy between control subjects and PD subjects with different disease grades.

In this way, we represented each persistence diagram as a persistent entropy with three numbers. Thus, the gait characteristics of each subject could be transformed into persistent entropy, which represents the information of each characteristic and greatly reduced the data dimension of the input sample.

### 2.8. Data Oversampling

According to the above methods, we obtained the persistent entropy of each subject's gait characteristics as the sample data for the classification training of PD. Obviously, there was still a significant imbalance in the sample data. In the data we used, there were twice and three times as many as grade 2 subjects as there were grade 2.5 and 3 subjects, respectively. Subjects with a severity level of 3 were regarded as the lowest group, accounting for only 20.7% of the total sample dataset. If the data is directly put into the classifier for learning, then the test results of the classifier will be biased to most classes, resulting in the problem of insensitivity to the identification of a few classes, which is very unfavorable to the training of the classifier. In order to avoid this situation, we use Borderline-SMOTE to balance the dataset. The Borderline-SMOTE [33] is an improved oversampling algorithm based on SMOTE that uses only a few class samples on the boundary to achieve the oversampling, thus improving the class distribution of the sample. The specific steps of Borderline-SMOTE are as follows:

1. Calculate the Euclidean distance between each sample point  $p_i$  and all the training samples, and get the  $m$  nearest neighbor of the sample point.
2. A small number of samples were divided. Assuming that  $m'$  of the  $m$  nearest neighbor samples belong to most of the samples ( $0 \leq m' \leq m$ ), there can be three situations, as follows: when  $m' = m$ ,  $p_i$  is considered as noise and data synthesis is not performed; when  $0 \leq m' \leq \frac{m}{2}$ ,  $p_i$  is considered as a safe sample and no data synthesis is performed; when  $\frac{m}{2} \leq m' \leq m$ ,  $p_i$  is divided into boundary samples and the data needs to be synthesized, the boundary sample is denoted as:

$$\{p'_1, p'_2, \dots, p'_{dnum}\}, (0 \leq dnum \leq pnum) \tag{10}$$

where  $dnum$  is the number of minority-class boundary samples and  $pnum$  is the total number of minority samples.

3. The  $K$  nearest neighbor between the boundary sample point  $p_i$  and the minority sample  $P$  is calculated. According to the sampling ratio  $U$ ,  $s$  (The number of  $s$  is  $K$  nearest neighbors multiplied by sampling ratio  $U$ ) and  $p'_i$  are selected for linear interpolation to synthesize a small number of samples.

$$Synthetic = p'_i + r_i \times d_j, (j = 1, 2, \dots, s) \tag{11}$$

where  $d_j$  identifies the distance between  $p'_i$  and its  $s$  neighbors, and  $r_j$  is a random number between 0 and 1.

4. A few kinds of synthetic samples and the original training samples are combined to form a new training sample.

By using the Borderline-SMOTE method, the sample set reached a balance of the class, and use the balanced dataset for classifier training in the following step. In this study, we only used the Borderline-SMOTE method during training to enhance the data, but the training set remained unchanged.

### 2.9. Machine-Learning Method

The classification of PD is essentially a multiclassification problem based on small sample data. To solve the few-shot learning problem, SVM is a novel few-shot learning method with a solid theoretical foundation that can achieve better results than other classifiers on the small sample training set. The reason why SVM has an excellent performance in few-shot learning is that it basically does not involve probability measurements or the law of large numbers. In essence, SVM avoids the traditional process from induction to deduction and achieves efficient classification and regression. At the same time, SVM can also solve the few-shot learning generalization ability, but is not strong. Since the optimization goal of SVM itself is to minimize the structured risk [37] rather than the empirical risk, the concept of the interval is used to obtain the structured description of data distribution, which

reduces the requirements for data size and data distribution. This gives SVM an excellent generalization ability. In addition, a small amount of support vectors determines the final result of SVM. Adding or deleting nonsupport vector samples has no effect on the model, which gives the SVM training model good robustness. For PD classification in this study, the dimension of the training sample was higher, and in aiming at this problem, the SVM provided a way to avoid the complexity of the high-dimensional space, the inner product function directly in this space, the kernel function, the solution of the recycling in the case of the linear separable method to directly solve the decision problem of the corresponding higher-dimensional space, and to simplify the solution of the higher-dimensional space problem. Compared with other algorithms such as the neural network, SVM, which is based on the principle of structural risk minimization, avoids overlearning problems, and has a strong generalization ability. SVM is a convex optimization problem, so the local optimal solution must be the global optimal solution.

SVM is a learning device to dichotomize linearly separable samples. In this study, we used the radial basis function (RBF) to convert the samples to the state of linear-separable or approximate linear-separable. The classification of PD is a multiclassification problem. The strategy of one vs. one (OvO) or one vs. rest (OvR) and a dichotomous classification algorithm can be adapted to classify PD using SVM. In this study, we need to classify 4 types of samples from 3 different classes of patients and normal subjects. OvR’s method is to take one sample as a class and treat the remaining samples of all types as another class to form four dichotomous problems and train a total of four models. OvO’s method combines two classes of samples each time to form six dichotomous problems and train a total of six models. When we classify, the samples to be tested are passed into all models, and the corresponding result of the model with the highest probability is the final result. Obviously, the OvO method has a higher accuracy, but it also takes a longer time. In this study, the sample size was small and there was no significant difference in the number of models generated by the two strategies, so we chose the OvO strategy with higher accuracy to solve the multiclassification problem of SVM.

2.10. Statistics

In this study, the classification of PD was a multiclassification problem. When evaluating the performance of the classifier, we paid more attention to the recognition accuracy and misjudgment between categories, in addition to the recognition accuracy of each category. In the evaluation of multiple categories, we transformed the problem of multiple categories into the problem of multiple dichotomies for performance evaluation. In this study, five indicators were used to evaluate the performance of the classifier, including global accuracy, single-class precision, single-class recall, inter-class precision, and inter-class recall. In the following equations, *T* indicates the classification is correct and *F* indicates the classification is incorrect, and *P* and *N* indicate whether the sample is positive or negative, respectively.

$$accuracy = \frac{ncorrect}{N} \tag{12}$$

where *accuracy* is global accuracy, *ncorrect* is the number of all predicted correct samples, and *N* is the total number of samples.

$$P_{class} = \frac{TP_{class}}{TP_{class} + FP_{class}} \tag{13}$$

$$R_{class} = \frac{TP_{class}}{TP_{class} + FN_{class}} \tag{14}$$

where *P<sub>class</sub>* and *R<sub>class</sub>* are single-class precision and single-class recall, and *class* is the category to be evaluated.

$$P_{p-n} = \frac{TP_{p-n}}{TP_{p-n} + FP_{p-n}} \tag{15}$$

$$R_{p-n} = \frac{TP_{p-n}}{TP_{p-n} + FN_{p-n}} \tag{16}$$

where  $P_{p-n}$  and  $R_{p-n}$  are inter-class precision and inter-class recall,  $p$  represents the positive class, and  $n$  represents the negative class.

### 3. Results

#### 3.1. SVM Classification

The data processing and classification of the classifier in this work were completed on a workstation including an Intel (R) Core (TM) i7-5930K@ 3.50 GHz, 6 CPU cores and 32.0 GB memory (Santa Clara, CA, USA). The models used were all written in a Python 3.7 environment using Giotto-TDA 0.3.1 and scikit-learn 0.23.1 under Ubuntu 16.04.7 LTS. In the classification training, we used 50%, 60%, 70%, 80%, and 90% of the datasets as the training set, and the rest of the samples as the test set for training, and conducted a 10-fold cross-validation on the model.

In the training of SVM, in order to get better parameters, we used the method of network search cross-validation to traverse various parameter combinations to determine the best parameters, which is very suitable for small sample sets. In SVM, the parameter  $C$  is the penalty coefficient. The higher  $C$  is, the more the classifier cannot tolerate errors, which will lead to overfitting, and the lower  $C$  is, the less likely there will be underfitting. In addition, we choose RBF as the kernel function of SVM, where the parameter  $\gamma$  affects the number of support vectors in the model. The relationship between the size of  $\gamma$  and the number of support vectors is: when  $\gamma$  is larger, the support vector is lower; when  $\gamma$  is smaller, the support vector is higher. Through the method of network search cross-validation, the two parameters are traversed on the interval, and all the values are combined. Each time, they are evaluated by a 10-fold cross-validation. Finally, the best value of the penalty coefficient was  $C = 1.0536$ , and the best value of  $\gamma$  in the RBF function was  $\gamma = 0.0188$ .

When the training set accounted for 50–90% of the training set, the model’s accuracy for the corresponding test results was 93.75%, 95.31%, 97.92%, 100%, and 100%. It can be seen that the trained model had a good effect on the recognition accuracy of different disease categories.

In the case of different proportions of training samples,  $P_{class}$  and  $R_{class}$  are shown in Tables 2 and 3 and the confusion matrix is shown in Figure 7.

Table 2. The precision of the model.

Training Samples	Class			
	$P_0$	$P_2$	$P_{2.5}$	$P_3$
50%	100.00%	100.00%	85.71%	91.67%
60%	85.71%	92.86%	90.00%	100.00%
70%	93.33%	100.00%	91.67%	100.00%
80%	100.00%	100.00%	100.00%	100.00%
90%	100.00%	100.00%	100.00%	100.00%

Table 3. The recall rate of the model.

Training Samples	Class			
	$R_0$	$R_2$	$R_{2.5}$	$R_3$
50%	86.67%	88.00%	100.00%	100.00%
60%	85.71%	100.00%	94.74%	100.00%
70%	93.33%	100.00%	100.00%	100.00%
80%	100.00%	100.00%	100.00%	100.00%
90%	100.00%	100.00%	100.00%	100.00%

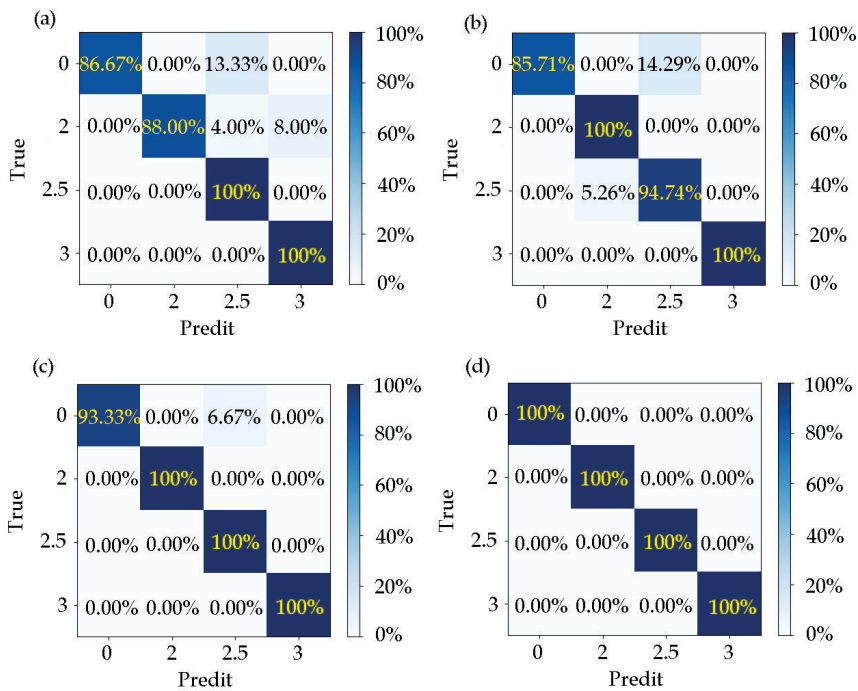


Figure 7. The confusion matrix of the test results with 50–80% (confusion matrix (a–d)) of the samples.

The results for the inter-class precision and recall ratio when the training set samples accounted for 50%, 60%, 70%, 80%, and 90% are shown in Tables 4–7.

Table 4. Inter-class precision and recall of Co as positive.

Training Samples	Positive = 0					
	$P_{0-2}$	$R_{0-2}$	$P_{0-2.5}$	$R_{0-2.5}$	$P_{0-3}$	$R_{0-3}$
50%	100.00%	100.00%	100.00%	86.67%	100.00%	100.00%
60%	100.00%	100.00%	100.00%	85.72%	100.00%	100.00%
70%	100.00%	100.00%	100.00%	93.33%	100.00%	100.00%
80%	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%
90%	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%

Table 5. Inter-class precision and recall of HY = 2 as positive.

Training Samples	Positive = 2					
	$P_{2-0}$	$R_{2-0}$	$P_{2-2.5}$	$R_{2-2.5}$	$P_{2-3}$	$R_{2-3}$
50%	100.00%	100.00%	100.00%	95.65%	100.00%	91.67%
60%	100.00%	100.00%	92.86%	100.00%	100.00%	100.00%
70%	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%
80%	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%
90%	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%

The data showed that the model did not misjudge patients as normal. When the proportion of training samples was 50%, there were cases in which the normal and disease grade 2 were misjudged as grade 2.5, and grade 2 was mistakenly judged as grade 3. When the proportion of training samples was 40%, there were cases in which the disease grade

of 2.5 was misjudged as level 2, and the normal level was wrongly judged as level 2.5. When the proportion of training samples was 30%, normal people were misjudged as the disease grade of 2.5. When the training samples accounted for 20% and 10%, there was no misjudgment. It can be seen that when the proportion of training samples increased, the learners acquired more information, which made the effect of the model gradually better.

**Table 6.** Inter-class precision and recall of HY = 2.5 as positive.

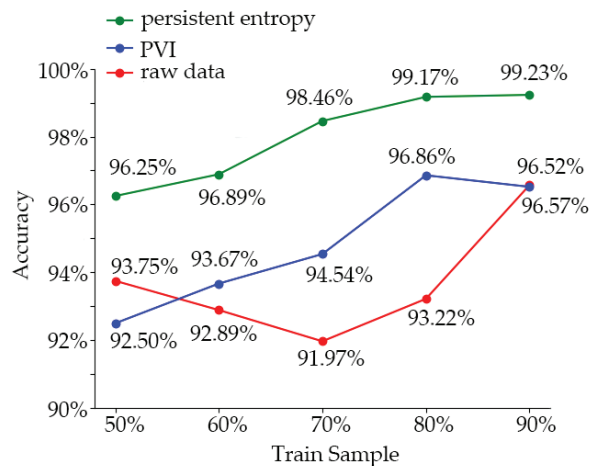
Training Samples	Positive = 2.5					
	$P_{2.5-0}$	$R_{2.5-0}$	$P_{2.5-2}$	$R_{2.5-2}$	$P_{2.5-3}$	$R_{2.5-3}$
50%	90.00%	100.00%	94.74%	100.00%	100.00%	100.00%
60%	90.00%	100.00%	100.00%	100.00%	100.00%	100.00%
70%	91.67%	100.00%	100.00%	100.00%	100.00%	100.00%
80%	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%
90%	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%

**Table 7.** Inter-class precision and recall of HY = 3 as positive.

Training Samples	Positive = 3					
	$P_{3-0}$	$R_{3-0}$	$P_{3-2}$	$R_{3-2}$	$P_{3-2.5}$	$R_{3-2.5}$
50%	100.00%	100.00%	91.67%	100.00%	100.00%	100.00%
60%	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%
70%	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%
80%	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%
90%	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%

### 3.2. Impact of Processing Strategies

In order to analyze the effect of the sample data-processing method used in this experiment, we used the dataset without processing and the dataset using only the variable-importance processing to train the learner. The training accuracy of the model was compared with the effect of the method used in this experiment. The training accuracy comparison results of the three groups of models are shown in Figure 8. And the comparison to other researches and summarized is shown in Table 8



**Figure 8.** The training accuracy of the raw data, permutation-variable importance processing and persistent entropy.

**Table 8.** Comparison with other methods.

	Yan Yan et al. [26]	Enas Abdulhay et al. [22]	Aite Zhao et al. [19]	Wei Zeng et al. [20]	Tunç Aşuroğlu et al. [23]	Present Method
Accuracy	87.10%	94.14%	98.70%	98.80%	99.00%	<b>99.23%</b>

From the comparison results, we can see that the training accuracy of the model trained by the combination of variable-importance processing and TDA persistent entropy was up to 99.23% (the training samples accounted for 90%). The training accuracy of the model trained with the dataset treated by the importance of variables did not increase with the increase of the proportion of training samples (96.86%, 80%; 96.52%, 90%), and maintained at this level. Without data processing, the training accuracy of the model trained by the learner appeared as a U-shaped curve from high to low and then to high; when the training set accounted for 50% to 90%, the training accuracy was 93.75%, 92.89%, 91.97%, 93.72%, and 96.57%, respectively. The reason for this is that SVM could well fit a small number of samples, while the features of the data without processing were not obvious, and the influence of irrelevant features was greater. As the number of samples increased, more complex information appeared, which reduced the training accuracy. When the number of samples increased further, the learner acquired more information, which made the training accuracy increase. In conclusion, the training effect of SVM in a small sample dataset was excellent, and the irrelevant features could be eliminated by variable-importance processing to avoid overfitting of the training model. Using topology analysis and persistent entropy training could further enhance the discrimination of samples and significantly improved the training accuracy.

#### 4. Discussion and Conclusions

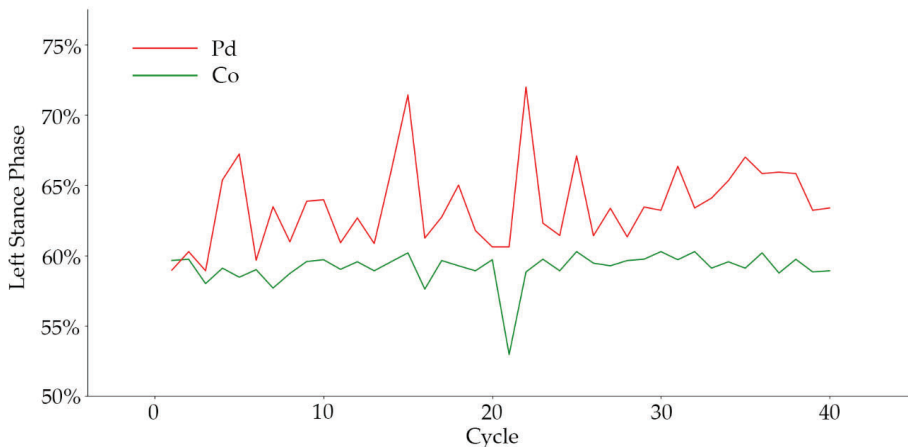
There is always a problem of insufficient samples in the recognition of PD. Similar to other studies of abnormal gait, the number of subjects with different disease grades of PD is usually very limited, and the samples are commonly unbalanced. These factors suggest that the grade recognition of PD is a few-shot learning problem.

In this paper, the common GRF dataset was used, which can show the walking pattern of PD patients well. However, from the point of view of the learning effect, the training accuracy curve of the model trained by GRF data showed a U-shape with an increase of the number of samples. The reason for this phenomenon is that the feature discrimination of untreated GRF data was not significant, and contained many irrelevant features/interferences. When the number of samples increased, the learner could not fit the new irrelevant information well, which led to the reduction of training accuracy. This indicates that the GRF data contained too much disturbing information. In another way, some characteristics did not change much among classes.

To solve this problem, we processed the original GRF sample data. The GRF sample data was first divided according to the gait cycle. Considering the existence of minority classes (such as the abnormal gait class), if the time span of GRF data used to calculate gait characteristics was too long, it could cause the loss of key information, and it would not be able to clearly characterize the abnormal gait problem. After the data partition, GRF was used to calculate potential gait features that may affect the classification of the disease grade. For the selection of gait features, we referred to the relevant research and the previous research [22,41], and selected a series of gait features that could be calculated from GRF data for further analysis. After the potential gait features were obtained, we measured the importance of gait features.

From the experiment results, we observed that the aggravation of PD was directly reflected in walking speed. When the severity of the disease worsened, serious gait disorders hindered the patient from normal speed walking, resulting in a slow movement. This conclusion was strongly supported by the experiments. In addition, walking speed was observed to also affect the stride time of patients, which is also reflected in the results.

In other gait features with significant influence, we found that the proportion of gait phase in the classification of disease grade had a high degree of differentiation. There were great differences in the proportion of support-phase and swing-phase time in patients with different disease grades, which also reflects the walking mode of patients with different severity levels. When abnormal gaits such as freezing gait and panic gait occurred, the proportion of gait phase changed significantly. In frozen gait, the proportion of support phase increased significantly. The frequency of the abnormal gait increased significantly when the disease grade was aggravated, and the change of the proportion of gait phase was more clear. For instance, the left-foot gait-phase ratio of PD patients to control subjects is shown in Figure 9.



**Figure 9.** The proportion of left foot support between control subjects and PD patients (one subject in each group).

At the same time, we demonstrated that the RMS of coordinates CoP velocity, CoP efficiency, and sample entropy have significant discrimination in the Y-axis direction. This result indicated that the component of the gait characteristic in the walking direction could significantly influence the classification of disease grade of PD. In addition, according to the importance of the left and right directions of each feature and the importance of CSIP coordinates, we found that there was no significant difference in gait symmetry among PD patients, so this symmetry could not be used as a basis for distinguishing disease grades; that is to say, the abnormal gait pattern of PD was not found in only one limb.

Considering the complexity of the human body, we analyzed the gait features. The results showed that the persistent entropy model was better than the model without topology data analysis. Although we could get good results by measuring the importance of variables, the training accuracy reached the peak when the proportion of training samples reached 80%. Increasing the number of training samples could not improve the training accuracy. This indicates that the effect of only using gait features to distinguish different PD grades encountered a bottleneck. When the persistent entropy was used as the training sample, the training accuracy of the learner broke through this bottleneck and reached 99.23%. The results showed that the TDA method could further extract the differences between gait features of different disease grades, and improved the discrimination among classes. This was due to the strong nonlinearity and complexity of human walking, and the SVM we used was essentially a linear classifier. The TDA method could map the gait feature data to the high-dimensional space and mine the sample features at a deeper level, which made the sample discrimination increase. Therefore, it was suitable for solving few-shot machine-learning problems related to human gait.



In addition, the training cost of samples processed by different methods is also different. The method of persistent entropy can be simplified to represent a class of gait features with only three numbers, which greatly reduces the dimension of the sample and significantly reduces the computation load during training.

In the problem of sample balancing, persistent entropy is used to strengthen the discrimination between different classes of samples, which makes the distance between different categories further. This avoids the blindness of the SMOTE algorithm in neighbor selection to a certain extent, and makes the synthesized samples achieve a better training effect. According to the misclassification of severity levels, there are some cases in which normal people are recognized as patients, or low-level cases are identified as high-level cases. This is because when there are too few training samples, the walking speed becomes the most important feature. When the walking speed of the older normal or mild patients is too slow, the learner will mistakenly classify them as a serious manifestation of the disease, resulting in misclassification. When the number of training samples increases, this kind of misclassification can be improved.

In summary, this paper proposed a few-shot learning method based on the measurement of permutation-variable importance and topological-imprint persistent entropy. The GRF was used as the basic data, Borderline-SMOTE was used as sample balancing method, and SVM was used as a classifier to identify the grade of PD. The proposed method achieved better results than when using original data. At the same time, the results of our study also indicated the leading factors of the differences among disease grades, which is valuable in further understanding the differential performance of different PD grades, revealing the walking characteristics of PD patients, and guiding the targeted health care.

**Author Contributions:** J.Z. and J.T. conceived the key idea; J.Z. analyzed the data and wrote the original draft; J.T. provided valuable suggestions for the experiments and reviewed the article; E.D. and J.Z. designed and carried out the experiments; and S.D. provided guidance for the analysis method and revised the paper. J.T. and J.Z. contributed equally to this work. All authors have read and agreed to the published version of the manuscript.

**Funding:** This work was supported by the Natural Science Foundation of Tianjin (No. 18JCY-BJC87700).

**Institutional Review Board Statement:** Not applicable.

**Informed Consent Statement:** Not applicable.

**Data Availability Statement:** Publicly available datasets were analyzed in this study. This data can be found here: [<https://physionet.org/content/gaitpdb/1.0.0/>], (accessed on 16 January 2021).

**Conflicts of Interest:** The authors declare no conflict of interest.

## References

- Lang, A.E.; Lozano, A.M. Parkinson's Disease. *Lancet* **1998**, *386*, 896–912.
- Jankovic, J.; McDermott, M.; Carter, J.; Gauthier, S.; Goetz, C.; Golbe, L.; Huber, S.; Koller, W.; Olanow, C.; Shoulson, I. Variable expression of parkinson's disease. *Neurology* **1990**, *40*, 1529. [[CrossRef](#)]
- Hoehn, M. Parkinsonism: Onset, progression and mortality. *Neurology* **1967**, *17*, 427–442. [[CrossRef](#)] [[PubMed](#)]
- Allen, J.L.; Kautz, S.A.; Neptune, R.R. Forward propulsion asymmetry is indicative of changes in plantarflexor coordination during walking in individuals with post-stroke hemiparesis. *Clin. Biomech.* **2014**, *29*, 780–786. [[CrossRef](#)]
- Roelker, S.A.; Bowden, M.G.; Kautz, S.A.; Neptune, R.R. Paretic propulsion as a measure of walking performance and functional motor recovery post-stroke: A review. *Gait Posture* **2019**, *68*, 6–14. [[CrossRef](#)] [[PubMed](#)]
- Muniz, A.M.S.; Liu, H.; Lyons, K.E.; Pahwa, R.; Liu, W.; Nobre, F.F.; Nadal, J. Comparison among probabilistic neural network, support vector machine and logistic regression for evaluating the effect of subthalamic stimulation in Parkinson disease on ground reaction force during gait. *J. Biomech.* **2010**, *43*, 720–726. [[CrossRef](#)] [[PubMed](#)]
- Petrucci, M.N.; Mackinnon, C.D.; Hsiao-Weckler, E.T. Modulation of Anticipatory Postural Adjustments Using a Powered Ankle Orthosis in People with Parkinson's Disease and Freezing of Gait. *Gait Posture* **2019**, *72*, 188–194. [[CrossRef](#)]
- Oh, J.; Eltoukhy, M.; Kuenze, C.; Andersen, M.S.; Signorile, J.F. Comparison of predicted kinetic variables between Parkinson's disease patients and healthy age-matched control using a depth sensor-driven full-body musculoskeletal model. *Gait Posture* **2020**, *76*, 151–156. [[CrossRef](#)]

9. Kleanthous, N.; Hussain, A.J.; Khan, W.; Liatsis, P. A new machine learning based approach to predict Freezing of Gait. *Pattern Recognit. Lett.* **2020**, *140*, 119–126. [[CrossRef](#)]
10. Vos, M.D.; Prince, J.; Buchanan, T.; Fitzgerald, J.J.; Antoniadis, C.A. Discriminating progressive supranuclear palsy from Parkinson's disease using wearable technology and machine learning. *Gait Posture* **2020**, *77*, 257–263. [[CrossRef](#)]
11. Wahid, F.; Begg, R.K.; Hass, C.J.; Halgamuge, S.; Ackland, D.C. Classification of Parkinson's Disease Gait Using Spatial-Temporal Gait Features. *IEEE J. Biomed. Health Inform.* **2015**, *19*, 1794. [[CrossRef](#)]
12. Hannink, J.; Thomas, K.; Cristian, F.P.; Jens, B.; Samuel, S.; Karl-Gunter, G.; Jochen, K.; Bjoern, M.E. Stride length estimation with deep learning. *arXiv* **2016**, arXiv:1609.03321.
13. Taha, K. Motion Cue Analysis for Parkinsonian Gait Recognition. *Open Biomed. Eng. J.* **2013**, *7*, 1–8.
14. Paredes, J.D.A.; Muñoz, B.; Agredo, W.; Ariza-Araújo, Y.; Orozco, J.L.; Navarro, A. A reliability assessment software using Kinect to complement the clinical evaluation of parkinson's disease. In Proceedings of the 2015 37th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC), Milano, Italy, 25–29 August 2015.
15. Rocha, A.P.; Choupina, H.; Fernandes, J.M.; Rosas, M.J.; Cunha, J.P.S. Kinect v2 based system for Parkinson's disease assessment. In Proceedings of the International Conference of the IEEE Engineering in Medicine & Biology Society, Milan, Italy, 25–29 August 2015.
16. Rocha, A.P.; Choupina, H.; Fernandes, J.M.; Rosas, M.J.; Cunha, J.P.S. Parkinson's disease assessment based on gait analysis using an innovative RGB-D camera system. In Proceedings of the Engineering in Medicine & Biology Society, Chicago, IL, USA, 26–30 August 2014.
17. Wang, S.; Chao, G.; Cai, Z.; Chen, H.; Liu, W. An efficient hybrid kernel extreme learning machine approach for early diagnosis of Parkinson's disease. *Neurocomputing* **2016**, *184*, 131–144.
18. Rodríguez-Martín, D.; Samà, A.; Pérez-López, C.; Cabestany, J.; Català, A. Posture transition identification on PD patients through a SVM-based technique and a single waist-worn accelerometer. *Neurocomputing* **2015**, *164*, 144–153. [[CrossRef](#)]
19. Zhao, A.; Qi, L.; Li, J.; Dong, J.; Yu, H. A Hybrid Spatio-temporal Model for Detection and Severity Rating of Parkinson's Disease from Gait Data. *Neurocomputing* **2018**, *315*, 1–8. [[CrossRef](#)]
20. Zeng, W.; Yuan, C.; Wang, Q.; Liu, F.; Wang, Y. Classification of gait patterns between patients with Parkinson's disease and healthy controls using phase space reconstruction (PSR), empirical mode decomposition (EMD) and neural networks. *Neural Netw.* **2019**, *111*, 64–76. [[CrossRef](#)]
21. Balaji, E.; Brindha, D.; Balakrishnan, R. Supervised machine learning based gait classification system for early detection and stage classification of Parkinson's disease. *Appl. Soft Comput.* **2020**, *94*, 106494.
22. Abdulhay, E.; Arunkumar, N.; Narasimhan, K.; Vellaiappan, E.; Venkatraman, V. Gait and tremor investigation using machine learning techniques for the diagnosis of Parkinson disease. *Future Gener. Comput. Syst.* **2018**, *83*, 366–373. [[CrossRef](#)]
23. Auroolu, T.; Ac, K.; Erda, C.B.; Toprak, M.K.; Oul, H. Parkinson's disease monitoring from gait analysis via foot-worn sensors. *Biocybern. Biomed. Eng.* **2018**, *38*, 760–772. [[CrossRef](#)]
24. Reynard, F.; Terrier, P. Determinants of gait stability while walking on a treadmill: A machine learning approach. *J. Biomech.* **2017**, *65*, 212. [[CrossRef](#)] [[PubMed](#)]
25. Breiman, L. Random forests. *Mach. Learn.* **2001**, *45*, 5–32. [[CrossRef](#)]
26. Yan, Y.; Omisore, O.M.; Xue, Y.C.; Li, H.H.; Wang, L. Classification of Neurodegenerative Diseases via Topological Motion Analysis—A Comparison Study for Multiple Gait Fluctuations. *IEEE Access* **2020**, *8*, 96363–96377. [[CrossRef](#)]
27. Anestis, A.A.; Lambert, L.; Jean, M.P.D. Random forests for global sensitivity analysis: A selective review. *Reliab. Eng. Syst. Saf.* **2020**, *206*, 107312.
28. Chen, B.; Xia, S.; Chen, Z.; Wang, B.; Wang, G. RSMOTE: A self-adaptive robust SMOTE for imbalanced problems with label noise. *Inf. Ences* **2020**, *206*, 107312.
29. Chawla, N.V.; Bowyer, K.W.; Hall, L.O.; Kegelmeyer, W.P. SMOTE: Synthetic Minority Over-sampling Technique. *J. Artif. Intell. Res.* **2002**, *16*, 321–357. [[CrossRef](#)]
30. Yen, S.J.; Lee, Y.S. Cluster-based under-sampling approaches for imbalanced data distributions. *Expert Syst. Appl.* **2009**, *36*, 5718–5727. [[CrossRef](#)]
31. Garcia, S.; Luengo, J.; Herrera, F. Tutorial on practical tips of the most influential data preprocessing algorithms in data mining. *Knowl. Based Syst.* **2016**, *98*, 1–29. [[CrossRef](#)]
32. Fernández, A.; García, S.; Herrera, F.; Chawla, N.V. SMOTE for learning from imbalanced data. *J. Artif. Intell. Res.* **2018**, *61*, 863–905. [[CrossRef](#)]
33. Han, H.; Wang, W.Y.; Mao, B.H. Borderline-SMOTE: A New Over-Sampling Method in Imbalanced Data Sets Learning. In Proceedings of the 2005 international conference on Advances in Intelligent Computing—Volume Part I, Hefei, China, 23–26 August 2005.
34. Cortes, C.; Vapnik, V.N. Support-Vector Networks. *Mach. Learn.* **1995**, *20*, 273–297. [[CrossRef](#)]
35. Begg, R.K.; Palaniswami, M.; Owen, B. Support Vector Machines for Automated Gait Classification. *IEEE Trans. Biomed. Eng.* **2005**, *52*, 828–838. [[CrossRef](#)]
36. Liu, Z.; Wang, L.; Zhang, Y.; Chen, C.L.P. A SVM controller for the stable walking of biped robots based on small sample sizes. *Appl. Soft Comput.* **2016**, *38*, 738–753. [[CrossRef](#)]
37. Abe, S. Introduction of Support Vector Machines for Pattern Classification-VI: Current Topics. *Syst. Control Inf.* **2009**, *53*, 205–210.

38. Goldberger, A.L.; Amaral, L.A.N.; Glass, L.; Hausdorff, J.M.; Ivanov, P.C.; Mark, R.G.; Mietus, J.E.; Moody, G.B.; Peng, C.K.; Stanley, H.E. PhysioBank, PhysioToolkit, and PhysioNet: Components of a new research resource for complex physiologic signals. *Circulation* **2000**, *101*, E215. [[CrossRef](#)]
39. Wang, Y.; Yao, Q.; Kwok, J. Generalizing from a Few Examples. *ACM Comput. Surv.* **2020**, *53*, 63.
40. Jia, Y.; Shelhamer, J.; Donahue, J.; Karayev, S.; Long, J.; Guadarrama, S.; Darrell, T. Caffe: Convolutional Architecture for Fast Feature Embedding. In Proceedings of the 22nd ACM international conference on Multimedia, New York, NY, USA, 3–7 November 2014; pp. 675–678.
41. Tong, L.; Hongbin, Z. Riemannian Manifold Learning. *IEEE Trans. Pattern Anal. Mach. Intell.* **2008**, *30*, 796. [[CrossRef](#)] [[PubMed](#)]
42. Rucco, M.; Castiglione, F.; Merelli, E.; Pettini, M. Characterisation of the Idiotypic Immune Network through Persistent Entropy. In *Proceedings of ECCS 2014*; Springer International Publishing: Cham, Switzerland, 2016.
43. Tong, J.; Zhang, J.; Dong, E.; Liu, C.; Du, S. The Influence of Treadmill on Postural Control. *IEEE Access* **2020**, *8*, 193632–193643. [[CrossRef](#)]
44. Shi, L.; Duan, F.; Yang, Y.; Sun, Z. The Effect of Treadmill Walking on Gait and Upper Trunk through Linear and Nonlinear Analysis Methods. *Sensors* **2019**, *19*, 2204. [[CrossRef](#)] [[PubMed](#)]

Article

# Convolution-GRU Based on Independent Component Analysis for fMRI Analysis with Small and Imbalanced Samples

Shan Wang, Feng Duan \*  and Mingxin Zhang

College of Artificial Intelligence, Nankai University, No.38 Tongyan Road, Jinnan District, Tianjin 300350, China; 2120190376@mail.nankai.edu.cn (S.W.); 1611260@mail.nankai.edu.cn (M.Z.)

\* Correspondence: duanf@nankai.edu.cn; Tel.: +86-022-85358718

Received: 17 September 2020; Accepted: 19 October 2020; Published: 23 October 2020



**Abstract:** Functional magnetic resonance imaging (fMRI) is a commonly used method of brain research. However, due to the complexity and particularity of the fMRI task, it is difficult to find enough subjects, resulting in a small and, often, imbalanced dataset. A dataset with small samples causes overfitting of the learning model, and the imbalance will make the model insensitive to the minority class, which has been a problem in classification. It is of great significance to classify fMRI data with small and imbalanced samples. In the present study, we propose a 3-step method on a small and imbalanced fMRI dataset from a word-scene memory task. The steps of the method are as follows: (1) An independent component analysis is performed to reduce the dimension of data; (2) The synthetic minority oversampling technique is used to generate new samples of the minority class to balance data; (3) A convolution-Gated Recurrent Unit (GRU) network is used to classify the independent component signals, indicating whether the subjects are performing episodic memory tasks. The accuracy of the proposed method is 72.2%, which improves the classification performance compared with traditional classifiers such as support vector machines (SVM), logistic regression (LGR), linear discriminant analysis (LDA) and k-nearest neighbor (KNN), and this study gives a biomarker for evaluating the reactivation of episodic memory.

**Keywords:** functional magnetic resonance imaging; independent component analysis; deep learning; recurrent neural network; functional connectivity; episodic memory; small sample learning

## 1. Introduction

Functional magnetic resonance imaging (fMRI) is a very effective non-invasive technique to study brain functions. The commonly mentioned fMRI mainly refers to BOLD-fMRI, which relies on the difference in the magnetization vector between oxyhemoglobin and deoxyhemoglobin to generate the fMRI signal, thereby obtaining changes in cerebral hemodynamics [1]. Due to its excellent non-invasive and high temporal and spatial resolution, fMRI method has become the most widely used method in the brain function researches. fMRI can accurately and reliably locate the cortical area of specific brain activity, which can be used to diagnose brain diseases. Another merit of fMRI is that it can track signal changes in real time and obtain time series of brain activities [2]. Therefore, a large number of brain science researchers began to study fMRI and introduced it to neuroscience field.

Through magnetic resonance imaging (MRI), researchers can get high-resolution anatomical images. At the same time, the function phase obtained by fMRI technology includes the signal changes over a period of time. After the brain function signals are obtained, the GLM model [3] can be used to calculate the brain activation levels of the subjects, including the individual level and the group level. The combination of anatomical images and function phases can reveal some patterns of human brain

activities, such as the activation of the brain area [4], the degree of lateralization [5], etc. However, most of the researchers focus on the obtained static images, and then get some qualitative patterns. An adult brain has tens of billions of cells. The number of the obtained voxels will be different according to the fMRI scanning interval. Taking the 3 mm scanning thickness as an example, the scanned four-dimensional data also contain about 270,000 voxels. If we only use activation maps, the temporal activity patterns of brain regions will be ignored, which cannot be accurately described by activation maps alone. However, if the activation intensity value of each voxel is manually calculated and observed, it is often difficult to achieve due to too much calculation and too many restricted conditions.

The application of machine learning algorithms alleviates the problems above. With the help of machine learning algorithms, the problem of insufficient computing power for manual calculations has been partly solved. Support vector machines (SVM) [6,7], random forest (RF) [8], logistic regression (LGR) [9] and other classifiers are widely used in the classification of fMRI data. In addition to classification, some machine learning methods have also been applied to other aspects of fMRI analysis. In the past studies, sparse coding [10,11] proved to be a good tool to focus on the activity of certain brain networks during tasks. Through the training of a large amount of data, the researchers obtained a lot of activity patterns of different brain regions under specific tasks. Even so, for hundreds of thousands of voxel signals in multiple brain regions, manual feature extraction is a very complicated work, although various feature extraction methods have been proposed [12,13]. Due to the complexity of the working mechanism of the human brain, it is difficult to say which extraction method is more appropriate for specific fMRI tasks. Therefore, as a supplement to various feature extraction methods, the independent component analysis (ICA) method was applied to the feature extraction of fMRI [14,15]. ICA assumes that the obtained fMRI signal is the result of superposition of multiple independent signal sources (spatially independent components). By blindly separating the fMRI signal, a spatially independent component map is obtained, which enhances the spatial connection of features.

The introduction of complex learning networks such as Convolutional Neural Networks (CNN) and Recurrent Neural Networks (RNN) [16,17] further increases the ability of fMRI data feature extraction and processing. These deep learning algorithms can automatically extract features from the input fMRI data to discover high-level information hidden in it. The CNN network convolves each three-dimensional image in the fMRI time series through the designed convolution kernel, which improves the comprehensive ability of the classifier to local information, so as to obtain relevant features between brain regions. In fact, CNN is also very common in the electroencephalography (EEG) process [18,19]. The RNN network is more focused on the correlation of data in time period, and the data at the past time point are also taken into consideration, which just meets the needs of learning fMRI data in time series. In addition to these two networks, some studies built a deep learning-based feature representation with a stacked auto-encoder to extract complex nonlinear relations [20]. In fact, these deep learning methods have indeed played a very important role in the data processing.

In the fMRI data acquisition process, the lack of subjects is a very common problem, which directly leads to a small sample size of fMRI data. Moreover, in the design of fMRI tasks, the number of different tasks is often inconsistent, which, at the same time, leads to the imbalance of fMRI data. As far as classification is concerned, how to classify data with imbalanced small samples is an important issue. Various data enhancement algorithms have been proposed to balance the data and expand the dataset, thereby improving the learning effect of the classification model (in fact, down sampling is also a usual strategy for balance data, but for data with a small sample size, the down sampling method will reduce the sample size and increase the difficulty for learning algorithms). Random oversampling is a simple and easy-to-use data balancing method, which can be achieved by random copying of the samples from the minority class, and has a good performance in some disease diagnosis applications [21]. However, random oversampling simply replicates samples from minority class, and does not add any new information, which makes it perform poorly when the sample size is small. In order to solve this problem, the synthetic minority oversampling technique (SMOTE) method was proposed to add new samples to minority classes by synthesizing data [22], thereby improving the

classification performance. Furthermore, the adaptive synthetic (ADASYN) method was proposed to enhance the synthesis of minority classes at the boundary [23]. Many studies have shown that these data balancing methods are effective [24–27].

In the task design of episodic memory, in order to ensure that the relevant brain regions are activated in the memory task, the difficulty of the tasks is inevitably increased. In general, in order to study the activation patterns of a certain brain area, multiple tasks with different contents and different times are often designed in the experiments, which results in an imbalance in the classification of data. If the learning algorithm is directly used to classify the brain states, there will be a very serious overfitting. This study used a public fMRI dataset from Openneuro [28] based on multi-object tracking memory tasks [29], which contains only 4 experimental stages, a total of 15 healthy subjects' experimental data. There is an imbalance in the number of samples between the two tasks. The ICA method is used to separate spatially independent components of the preprocessed data to reduce the dimension. The SMOTE algorithm is used to balance the training samples. The CNN network is used to extract relevant features, and the Gated Recurrent Unit (GRU) network [30] is used as a classifier to classify the experimental tasks to study whether the subjects carry out the episodic memory task will cause significant changes in the physiological state of the brain. It can help us explore the patterns of brain activities under different tasks, and it can also play an auxiliary role in the design of fMRI tasks.

## 2. Materials and Methods

### 2.1. Subjects and Dataset

The experiment involved 24 healthy subjects, all right-handed, 16 females, 8 males, aged 18–25 years old, with normal language, visual and auditory abilities, and no mental illness. Excluding one that was too active during the scanning process, a total of 23 subjects' data remained. However, due to data corruption in the public data set, only 15 subjects' data remain available.

The fMRI data scanning process was carried out at Princeton University, using a 3 Tesla whole-body Skyra MRI system (Siemens, Erlangen, Germany). The anatomical phase was collected by T1-weighted sequence, the voxel size was 1 mm × 1 mm × 1 mm, repetition time (TR) = 2530 ms; echo time (TE) = 3.37 ms, field of view (FOV) = 256 mm; 256 × 256 matrix. The functional phase is collected by T2\*-weighted echo-planar image (EPI), the voxel size is 3 mm × 3 mm × 3.9 mm, FOV = 192 mm; 64 × 64 matrix; TR = 2000 ms; TE = 33.0 ms.

### 2.2. Word Scene Experiment

The whole experiment is divided into six stages, and the public dataset contains fMRI data from the third stage to the sixth stage. This article mainly studies the data from the fifth stage.

Before the fifth stage of the experiment, all the subjects learned 30 word-scene pairs and performed a memory test to ensure that the subjects remembered these pairs. At the same time, in order to disturb the memory of the subjects, before the fifth stage, each subject was displayed with 16 lure words. These words were not matched with the scene pictures. When the subjects were familiar with these lure words, these words were used as a lure set in the fifth stage. In addition, these subjects also carried out multiple-object tracking (MOT) tasks in advance in order to familiarize themselves with the tasks in the fifth stage.

The main task of the fifth phase experiment is that the subjects recall the scenes associated with the words while performing the target tracking task. In the MOT task, the subjects will see 10 random points that do not overlap with each other on a black background, where the target point is red and the non-target point is green (in the case of multi-target tracking, there are five target point; there is only one target point for single target tracking). In addition, there is a white cross in the center of the screen. These points will be presented to the subjects for two seconds, after which all points will turn green and begin to move. Participants were asked to continuously track the originally red target point within 18 s. In addition, the fixed cross in the center is replaced by a word (which may or may

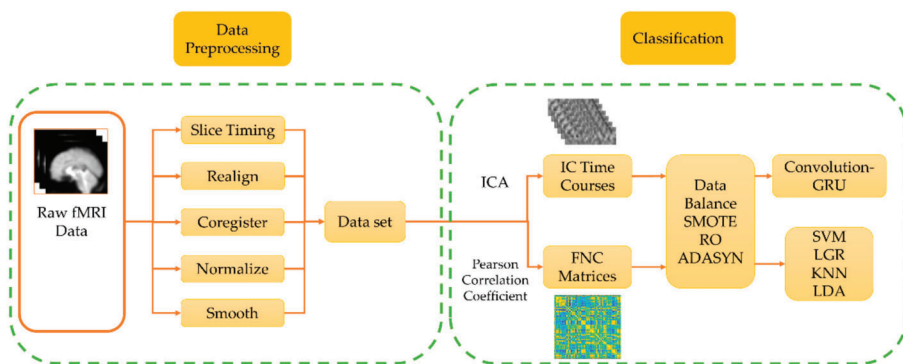


not be a word of the word scene pair) in white with a small white dot in the center. Subjects need to recall the scene paired with the word in as much detail as possible. Every five seconds, the center white dot will turn red. At that time, the subject needs to rate the specific degree of the scene he imagined. When the MOT task ends, all the points stop moving, and one of them is displayed in white. Subjects need to press the button to answer whether the white point was originally a target point or a non-target point. After three seconds, the subjects received one second of feedback, indicating whether the answer was correct or not. Finally, in order to disturb the imagination of the scene image after the experiment, the subjects needed to complete two calculation tasks. The sum of two numbers was displayed on the screen for 1.9 s. Subjects needed to press the button to answer whether the sum of the two numbers was odd or even. The text is displayed in white. When the subject answers correctly, the text is displayed in green; when the answer is incorrect, the text is displayed in red. The interval between the two calculation tests is 0.1s. After the two tests, the screen is fixed for 4 s before the next MOT task starts. The total time of each task is 32 s.

A total of 25 tasks were performed in each scan, of which there were ten single target tracking tasks with words from word-scene pairs (named as targ\_easy), ten multiple target tracking tasks with words from word-scene pairs (named as targ\_hard), and five multi-target tracking tasks with lure words (named as lure\_hard). The order of tasks is random. In the fifth stage, three fMRI scans were performed, resulting in a total of 75 tasks.

### 2.3. Data Analysis

Figure 1 shows the data processing framework of this study. A total of 15 subjects, 3 sessions, and 45 fMRI data are used. The fMRI data are preprocessed, and the spatially independent components of fMRI data and the functional connectivity (FNC) matrix between each brain area are calculated respectively. The SMOTE algorithm is used to enhance the training set data, and the enhanced data are used as input to train the classifier, and the performance of the classifier is compared.



**Figure 1.** The data processing framework of this study is divided into two parts: data preprocessing and classification. In the data preprocessing, the standard preprocessing process of functional magnetic resonance imaging (fMRI) is used. In data classification, independent component analysis (ICA) and Pearson correlation coefficients are performed on the processed data respectively to obtain independent component (IC) time courses and functional connectivity (FNC) matrix. After data balancing, the convolution-Gated Recurrent Unit (GRU) network and the traditional classifier are used for classification and the performance is compared.

#### 2.3.1. Pre-Processing

In this study, SPM12 [31] is used to preprocess the data (including slice timing, realign, coregister, normalize and smooth). First slice-timing is performed to correct the time point of each brain slice, and then the realign step is performed to correct the head movement of the subject. In order to locate

the points of the functional image on the anatomical image with higher resolution, the ‘coregister’ method is performed, and then, each image of the subject is normalized to the MNI standard brain space. The voxel size is resliced to  $3\text{mm} \times 3\text{mm} \times 3\text{mm}$ , and image data containing  $61 \times 73 \times 61$  voxels are obtained. Finally, in order to reduce noise and improve the signal-to-noise ratio, spatial smoothing filtering is performed on the data.

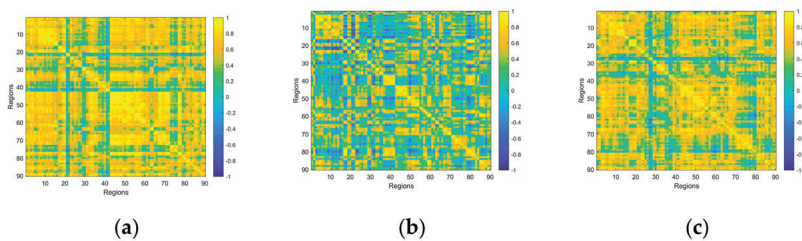
After pre-processing, in order to obtain the data of each multi-target tracking task, the fMRI data of each session need to be segmented. The entire scan lasts 810 s and contains 25 MOT tasks. Although the duration of each task is 32 s, considering that there are many processes that are not related to the MOT task (such as the fixed time of 2 s at the beginning, the calculation test, etc.), it may reduce the classification effect of the classifier, so we choose 4–18 s fMRI data of each task for training. Note that the start time of the first task is the 26 th second, and the scan time of the last task is insufficient, so these two pieces of data are discarded, and finally a total of 1080 data are left for use.

### 2.3.2. Functional Connectivity

Functional Connectivity is a common study point in brain researches, and some past studies reported FNC to be a good tool in disease analysis [32]. In this study, the AAL90 template [33] is used to divide the preprocessed data into 90 regions. We average the time series of all voxels in each area, and use the average time series as the signal of the area. In order to obtain the correlation between the signals of each brain region, the Pearson correlation coefficient is calculated on the time series of each two brain regions:

$$\rho_{X,Y} = \frac{\text{Cov}(X, Y)}{\sigma_X \sigma_Y} = \frac{E((X - \mu_X)(Y - \mu_Y))}{\sigma_X \sigma_Y} \quad (1)$$

$\mu_X$  is the mean value of the signal of region  $X$ ,  $\sigma_X$  is the standard deviation of the signal of region  $X$ . The range of  $\rho_{X,Y}$  is  $-1 \leq \rho_{X,Y} \leq 1$ . When  $0 < \rho_{X,Y} \leq 1$ , it indicates that the two sets of signals have the trend to be positively correlated, and the greater the  $\rho_{X,Y}$ , the stronger the relationship between the two signals. When  $\rho_{X,Y} = 1$ , the trend of the two signals is almost the same; when  $-1 \leq \rho_{X,Y} < 0$ , it means that the two signals have the trend to be negatively correlated, and the smaller the value, the stronger the negative correlation. When  $\rho_{X,Y} = -1$ , the trend of the signals is completely opposite. In particular, when  $\rho_{X,Y} = 0$ , it is generally considered that the signals are not related. After the calculation above, a  $90 \times 90$  FNC matrix is obtained, as shown in Figure 2. Then, these FNC matrices will be input into the traditional classifier as features for classification.



**Figure 2.** The FNC matrix of three tasks performed by a subject. (a) The FNC matrix of a single target tracking task with word from word-scene pairs. (b) The FNC matrix of a multiple target tracking task with lure word. (c) The FNC matrix of a multiple target tracking task with word from word-scene pairs. It should be noted that these three pictures only represent the functional connectivity of a certain task, but not the overall situation.

### 2.3.3. Independent Component Analysis

Independent component analysis is a method of extracting independent features (or independent signal sources) from data. In ICA, a basic assumption is that the observed data can be viewed as a linear superposition of multiple different independent components (or signal sources). Assuming that



$G = [g_1, g_2, \dots, g_m]^T$  is the observed signal, according to the previous assumption, there are independent signal sources  $S = [s_1, s_2, \dots, s_m]^T$  that satisfy:

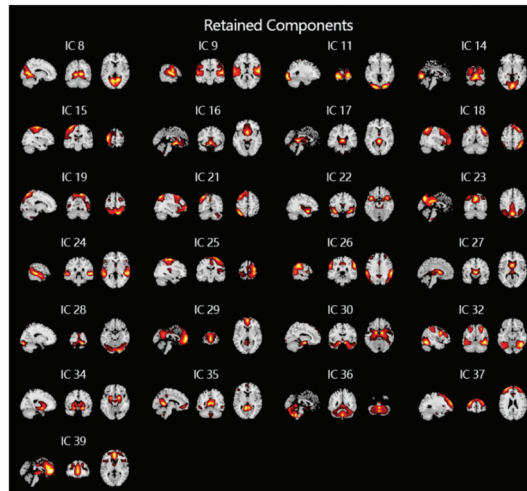
$$G = AS \tag{2}$$

where  $A$  is a matrix of  $n \times m$  (usually full rank), and the goal of ICA is to estimate a decomposition matrix  $B$  such that:

$$\hat{S} = BG \tag{3}$$

where  $\hat{S}$  is the best estimate of independent signal source  $S$ .

For fMRI signals, spatial ICA is often used to extract spatial features, which can extract time-series data of non-overlapping brain regions in space. In this study, the GIG-ICA method in the group ICA toolbox [34] (GIFT, <http://mialab.mrn.org/software/gift>) is used to analyze the pre-processed fMRI data of all subjects. Thirty-nine independent components are obtained at first. After manually removing artifacts unrelated to the experiment and signals from unrelated brain regions, 24 spatial independent components are finally obtained (shown Figure 3), as which greatly reduced the computational cost compared to using whole brain fMRI data directly.



**Figure 3.** The results of ICA of fMRI and removal of noise and artifacts. After performing ICA, a total of 39 independent components that do not overlap in space are obtained. After manually removing 14 noises and artifacts, the 25 independent components shown in the figure are finally obtained and used as input to the classifiers.

### 2.3.4. Data Oversampling

As introduced in the procedure, the MOT tasks are divided into three categories, namely: targ\_easy, targ\_hard, and lure\_hard. In each scan, there are 10 targ\_easy tasks, 10 targ\_hard tasks, and 5 lure\_hard tasks. Considering that the most significant difference between the three tasks is whether the word is from the word-scene pairs, so the data are divided into two categories, namely the target class and the lure class. Obviously, there is a serious imbalance in the number of samples of the two sets of data. As the majority class, the size of target class is three times that of the minority class. If the data are directly input into the classifier, it will make the classifier learn too much of the majority class data. If the data are directly input into the classifier, the test results will also be biased to the majority class, which is very disadvantageous for the training of the classifier. In order to balance the number of

training samples of the two categories, this study used Synthetic Minority Oversampling Technique (SMOTE) to balance the data set.

The basic idea of the SMOTE method is to use samples from the minority class to synthesize new samples and add them to the data set, so that the number of samples is balanced. The process of SMOTE algorithm is as follows:

1. First define a feature space and project all samples as points in this feature space. Then determine the sampling ratio according to the sample ratio of the majority class and the minority class.
2. For each minority sample  $(x, y)$ , find  $K$  samples closest to this sample according to the Euclidean distance, randomly select one  $(x_n, y_n)$  from the  $K$  samples, and construct a new one as follows:

$$(x_{new}, y_{new}) = (x, y) + rand(0, 1) \cdot ((x_n - x), (y_n - y)) \tag{4}$$

That is, randomly select a point on the line between the sample  $(x, y)$  and the nearest neighbor sample  $(x_n, y_n)$  as the new minority sample.

3. Repeat the above steps until the sample size is balanced.

After processed by the SMOTE method, the number of samples of the two classes is balanced and can be used to train the classifier in the subsequent steps. It should be noted that, in order to ensure the objectivity of the experimental method, this study only performs the SMOTE method on the training set, and the test set remains unchanged.

In addition to SMOTE, there are two other data oversampling methods that are also commonly used: the random oversampling method and the adaptive synthetic (ADASYN) method. The random oversampling method is easy to implement. Its principle is to randomly sample the minority samples to increase the number of minority samples. The ADASYN method is similar to SMOTE, and its basic process is as follows:

1. Calculate the degree of imbalance between the two classes:

$$p = \frac{n_s}{n_l} \tag{5}$$

where  $n_s$  is the sample number of the minority class and  $n_l$  is the sample number of the majority class. When  $p$  is less than the tolerance threshold, the oversampling process begins.

2. Calculate the number of samples to be synthesized:

$$M = (n_l - n_s) \cdot \beta \tag{6}$$

$\beta \in [0, 1]$  is the generation ratio, when  $\beta$  is 1, the number of the majority class and the balanced minority class are the same.

3. For each sample of the minority class, the  $K$  nearest neighbor is calculated according to the calculated Euclidean distance, and the ratio of the majority class  $r_i$  is calculated. Obviously,  $r_i \in [0, 1]$ .
4. Standardize  $r_i$ :

$$\hat{r}_i = \frac{r_i}{\sum_{k=1}^{n_s} r_k} \tag{7}$$

5. Calculate the number of new samples that need to be generated for each minority sample:

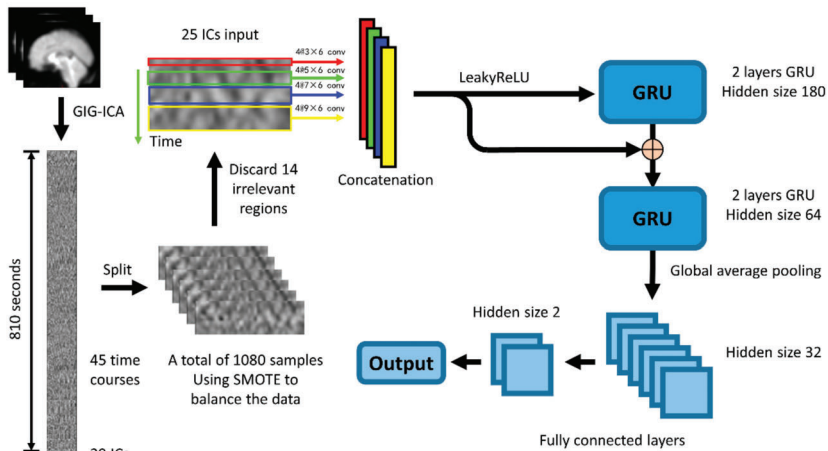
$$g_i = \hat{r}_i \cdot M \tag{8}$$

6. Generate new data in the same way as the second step of the SMOTE method until the quantity meets the requirements.

In order to test the effect of different data balancing methods on the classifier, this study also processed the training set using random oversampling and the ADASYN method, and compared the classification performances of the classifier.

### 2.3.5. Convolution-GRU Network

The structure of the Convolution-GRU network used in this study is shown in the Figure 4. The entire network consists of four 2-dimensional convolutional layers with convolution kernels of different sizes ( $3 \times 6$ ,  $5 \times 6$ ,  $7 \times 6$ ,  $9 \times 6$ ), a LeakyReLU layer, 2 gate recurrent units (GRU), a global average pooling layer and 2 fully connected layers.



**Figure 4.** The basic structure of the convolution-GRU model is to combine the spatial features obtained by convolution at four different time scales and send them to the GRU to learn the temporal features. Considering that negative values may appear in the process of convolution, the LeakyRelu layer is selected to ensure the learning of parameters. It should be noted that the number before the @ sign in this figure means the number of the channels of the convolution kernel.

First, in order to extract the spatial features in different time periods from the data (the extraction of temporal features is mainly carried out in GRU), 4 different scale convolution kernels are selected, in order to maintain the same length of the time series, padding is done in the direction of the time axis. After convolution, the data pass through the LeakyRelu layer and are concatenated together to form a  $7 \text{ (TRs)} \times 320 \text{ (features)}$  feature map, which will then be input into the GRU.

RNN has become one of the most commonly used time series artificial neural networks with its excellent time feature extraction ability. As a variant of LSTM network, GRU has similar performance to LSTM network, but it is easier to calculate than LSTM. In this study, two GRU models [14] with different hidden nodes connected in a feed-forward manner are used to extract high-dimensional time features from IC data. The feature map obtained in the previous step is first input into a GRU network with 180 hidden nodes, then the output of this GRU is concatenated with the previous feature map as the input of the GRU network with 64 hidden nodes.

Considering that the classifier used in this study is to classify the tasks performed by the subjects, the features of the brain activity of the subjects during the tasks are more valuable than the features at a single time point. Therefore, the output of the GRU network will go through a global average pooling layer to average the output on the entire time axis, instead of extracting only the output at the last time point to obtain a 64-dimensional feature vector. Finally, the feature vector passes through two fully connected layers to obtain the final binary classification result.

### 2.3.6. Statistic

In this study, we use 5 scores to evaluate the performance of the classifier; namely accuracy, majority (the large class) precision rate (LP), majority recall rate (LR), minority (the small class) precision rate (SP), and minority recall rate (SR). The calculation methods of precision rate (P) and recall rate (R) are as follows:

$$P = \frac{TP}{TP + FP}, R = \frac{TP}{TP + FN} \quad (9)$$

Among them TP, FP, TN, FN represent true positive, false positive, true negative, false negative. True and false indicate whether the classification is correct, positive and negative indicate whether the sample belongs to the positive or negative class, LP, LR, SP, SR are the results obtained when the majority class and the minority class are used as the positive class, respectively. It should be noted that the accuracy score is defined as follows:

$$accuracy = \frac{n_{correct}}{n_{all}} \quad (10)$$

$n_{correct}$  is the number of the samples which are classified correctly by the classifier and  $n_{all}$  is the number of the whole samples.

## 3. Results

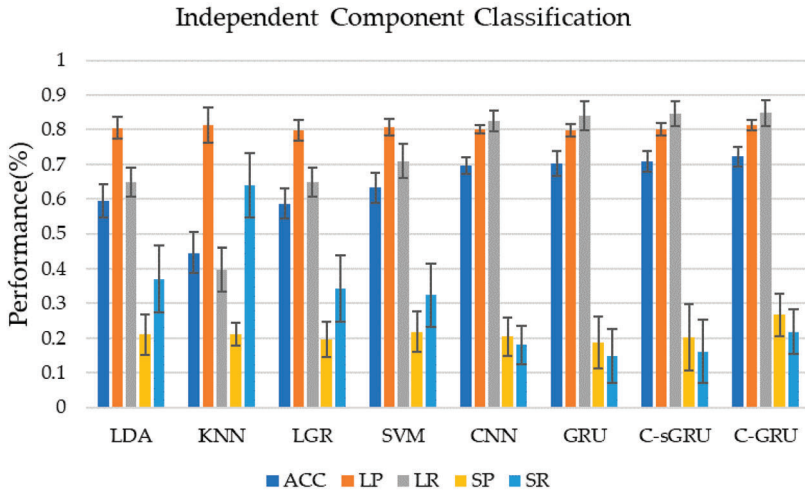
### 3.1. Convolution-GRU Classification

The training and classification of the classifier in this study are completed on a laptop computer, including an Intel (R) Core (TM) i7-9750H@ 2.60GHz, 6 CPU cores and an NVIDIA GeForce GTX 1660 Ti 6.0GB GPU, and 8.0 GB memory, the classification model used are all written in Python 3.8 environment using Pytorch 1.4.0 and Sklearn 0.22.2 under Windows 10. In order to compare with the classification performance of convolution-GRU classifier, we also trained four traditional classifiers (SVM (C = 0.01), k-nearest neighbor (KNN)(n\_neighbors = 2), linear discriminant analysis (LDA) (solver = 'lsqr'), LGR) and CNN and GRU networks. It should be noted that we have used independent component data to train these classifiers, but FNC data have no time series features, so only four traditional classifiers and CNN networks have been trained. Considering that the size of the data is small, 90% of the data set is used as the training set and 10% is used as the test set to allow the classifier to learn more information, and the model is tested using 10-fold cross validation.

During the training of the convolution-GRU network, the batch size is set to 64, and we used the Adam optimizer to reduce the cross-entropy loss of the model. The initial value of the learning rate is set to 0.01, and the learning rate is adjusted using the ReduceLRonPlateau method, that is, when the decrease in three consecutive training losses is not less than a certain threshold, the learning rate is reduced to half, and the threshold is set to 0.001.

Figure 5 and Table 1 show the classification performance of seven classification methods using independent component time series as training data in this study. Classification using the convolution-GRU network resulted in a classification accuracy of  $72.2 \pm 2.9\%$ , while using traditional classifiers (LDA, KNN, LGR, SVM) to classify time series of independent components yielded poor results. SVM gives the best performance in the traditional classifiers which achieved  $63.2 \pm 4.4\%$ . It indicates that the convolution-GRU network used in this study has achieved a large improvement in classification accuracy. In terms of the precision and recall of the two classes, the convolution-GRU model has achieved the best classification effect of the majority class, and the precision of the minority class has also reached  $26.7 \pm 6\%$ . While the precision of the majority class of the traditional classifier is comparable to that of the deep learning model, the recall of the majority class is lower, and the best performing SVM reaches only  $70.9 \pm 5.0\%$ . This indicates that the convolution-GRU model has fully learned the features of the majority class. For minority class, traditional classifiers often have a higher recall rate, but the precision rate is the same or lower than some simple deep learning models,

while the minority class precision rate of the convolution-GRU model reached  $26.7 \pm 6.1\%$ , which shows that this model can reduce the probability that samples from majority class are misclassified, thereby improving the credibility of the classification results. For deep learning models, when CNN, GRU, or convolution-single GRU models are used for classification, the performance of the majority class is similar to the convolution-GRU model, but the performance of the minority class is poor.



**Figure 5.** Classification results using independent component time series. Convolution-GRU achieved the highest classification accuracy rate of 72.2%, while traditional classifiers performed poorly, and none exceeded 65%. Deep learning models performed similarly in the majority class, however, convolution-GRU performed best in the minority class.

**Table 1.** Performance of the methods in IC classification.

Methods	ACC	LP	LR	SP	SR
LDA	0.5935 (0.0475)	0.8046 (0.0316)	0.6494 (0.0423)	0.2103 (0.0577)	0.3701 (0.0972)
KNN	0.4463 (0.0581)	0.8127 (0.0515)	0.3980 (0.0624)	0.2109 (0.0327)	<b>0.6400 (0.0924)</b>
LGR	0.5870 (0.0439)	0.7975 (0.0296)	0.6482 (0.0425)	0.1961 (0.0519)	0.3422 (0.0954)
SVM	0.6324 (0.0440)	0.8075 (0.0234)	0.7094 (0.0503)	0.2189 (0.0590)	0.3238 (0.0908)
CNN	0.6972 (0.0234)	0.8014 (0.0108)	0.8264 (0.0300)	0.2048 (0.0547)	0.1803 (0.0566)
GRU	0.7019 (0.0353)	0.7979 (0.0186)	0.8403 (0.0424)	0.1873 (0.0749)	0.1494 (0.0774)
C-sGRU	0.7093 (0.0302)	0.8018 (0.0175)	0.8460 (0.0350)	0.2024 (0.0945)	0.1617 (0.0924)
C-GRU	<b>0.7222 (0.0290)</b>	<b>0.8127 (0.0146)</b>	<b>0.8484 (0.0372)</b>	<b>0.2670 (0.0609)</b>	0.2182 (0.0640)

CNN: The convolutional layer of CNN used here is the same as the convolutional layer used in convolution-GRU. C-sGRU: convolution-single GRU, only one layer of GRU is used for classification after the convolutional layer. GRU: Use only GRU to classify time courses of independent components. It should be noted that the bold numbers in all tables in the paper indicate higher values in the column.

Considering that LP, LR, SP, and SR have different levels in each classifier, this article uses weighted F1 values to integrate these types of index to form a comprehensive index to evaluate the classifier used in this study. In the case of imbalanced sample sizes, the weighted F1 score can objectively evaluate the performance of the classifier compared to other scores. In fact, many studies have used this indicator [35,36]. The calculation method of the weighted F1 value is as follows:

$$weighted\_F1 = \frac{2}{\frac{1}{LP} + \frac{1}{LR}} \cdot \frac{n_l}{n_l + n_s} + \frac{2}{\frac{1}{SP} + \frac{1}{SR}} \cdot \frac{n_s}{n_l + n_s} \tag{11}$$

The calculated weighted F1 score of each classifier is shown in Table 2. It can be seen that under the weighted F1 score standard, the deep learning model still has great advantages over traditional classifiers. The weighted F1 value of the convolution-GRU network used in this paper reaches 0.6827, while the weighted F1 score of the KNN model is only 0.4801, which is also consistent with our previous analysis.

**Table 2.** Weighted F1 of the methods in IC classification.

Methods	Weighted F1
LDA	0.6061
KNN	0.4801
LGR	0.5987
SVM	0.6318
CNN	0.6582
GRU	0.6555
C-sGRU	0.6624
C-GRU	<b>0.6827</b>

Since we want to evaluate the comprehensive performance of the classifier, here we directly use the average of the precision and recall of the classifier to calculate the weighted F1 score. It can be seen from the table that the deep learning model still has great advantages over the traditional model.

### 3.2. Comparison of Different Oversampling Strategy

In order to compare the influence of different data balancing methods on the classification performance, we used three methods, random oversampling, SMOTE, and ADASYN, to balance the training set, and trained the convolution-GRU model. The classification results are shown in Table 3. It can be seen that the accuracy of the classifier is basically consistent under the three data balancing methods. The difference between the three methods is mainly reflected in the performance of the classification of minority data. As can be seen from the table, the data processed by the SMOTE method perform significantly better than the random oversampling method and the ADASYN method in minority class, and the precision and recall rates are the highest of the three methods. This shows that the data generated by SMOTE match the distribution of the original data more than the other two commonly used data balancing methods.

**Table 3.** Performance of different oversampling strategy.

Methods	ACC	LP	LR	SP	SR
RO	0.7046 (0.0489)	0.7977 (0.0166)	0.8450 (0.0668)	0.1939 (0.0831)	0.1435 (0.0841)
SMOTE	<b>0.7222 (0.0290)</b>	<b>0.8127 (0.0146)</b>	0.8484 (0.0372)	<b>0.2670 (0.0609)</b>	<b>0.2182 (0.0640)</b>
ADASYN	0.7157 (0.0259)	0.8031 (0.0124)	<b>0.8542 (0.0297)</b>	0.2157 (0.0657)	0.1617 (0.0613)

RO: Random Oversampling. Random oversampling, SMOTE, and ADASYN are common data balancing methods. In this study, these three methods are used to balance the data, and the classification performance of convolution-GRU is compared. It can be seen that the SMOTE method performs better than the other two methods in the accuracy of the minority class. It should be noted that all the three balancing methods are used in default parameters.

### 3.3. Classification of Using FNC Data

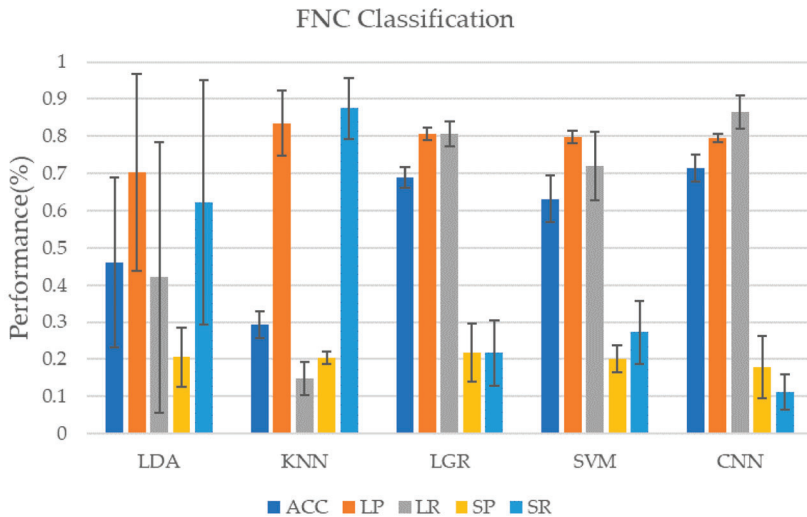
In addition to the ICA of the original data, we used the AAL90 template to obtain the FNC matrix of the subjects when performing the task. In order to compare with the classification results using independent component time series data for classification, we also used the FNC matrix as input to train the classifier. It should be noted that, because the FNC matrix we calculated is static data, not time series, no model containing GRU is selected for training. As shown in Table 4 and Figure 6, when using the CNN model for classification, the highest classification accuracy rate is achieved, reaching  $71.4 \pm 3.6\%$ , but the recall rate of the minority class is poor, only  $11.1 \pm 4.7\%$ , very unstable. LDA and KNN in traditional classifiers perform badly. These two classifiers are too biased to divide the data into the minority class, which makes the classification performance of the majority class worse.

The performance of LGR and SVM is more balanced. Although the accuracy and performance of most classes are not as good as the CNN model, they perform better in a few classes. Similarly, here also gives the weighted F1 value of each classifier when using FNC data for classification, as shown in Table 5. The deep learning model represented by CNN still maintains its advantage, reaching 0.6558, while the weighted F1 value of the KNN model is only 0.2716, which is the worst performance.

**Table 4.** Performance of the methods in FNC classification.

Methods	ACC	LP	LR	SP	SR
LDA	0.4602 (0.2284)	0.7033 (0.2649)	0.4208 (0.3646)	0.2059 (0.0803)	0.6229 (0.3296)
KNN	0.2935 (0.0361)	<b>0.8348 (0.0878)</b>	0.1482 (0.0448)	0.2043 (0.0175)	<b>0.8749 (0.0816)</b>
LGR	0.6889 (0.0278)	0.8052 (0.0166)	0.8068 (0.0331)	<b>0.2174 (0.0782)</b>	0.2171 (0.0878)
SVM	0.6315 (0.0630)	0.7980 (0.0174)	0.7209 (0.0920)	0.2003 (0.0361)	0.2729 (0.0856)
CNN	<b>0.7139 (0.0362)</b>	0.7954 (0.0111)	<b>0.8648 (0.0458)</b>	0.1794 (0.0832)	0.1110 (0.0472)

The CNN model used here contains two convolutional layers with a convolution kernel size of  $3 \times 3$ . In addition to LeakyRelu, a maximum pooling layer of size  $4 \times 4$  is also used.



**Figure 6.** Classification results using FNC matrix. Using the data of the subjects in the multi-object tracking (MOT) task to calculate the Pearson correlation coefficient to obtain the FNC matrix which is used as the input of the classifier. Since the FNC matrix itself is not a time course, so we chose CNN as the representative of the deep learning model for classification.

**Table 5.** Weighted F1 of the methods in FNC classification.

Methods	Weighted F1
LDA	0.4723
KNN	0.2716
LGR	<b>0.6588</b>
SVM	0.6259
CNN	<b>0.6558</b>

In order to compare the performance of the classifiers when using FNC data for classification, the weighted F1 score is also calculated. CNN and LGR model achieved the highest weighted F1 score.

In general, when using FNC as the input training classifier for classification, the three methods, LGR, SVM, and CNN, all have good performance in terms of accuracy, but the CNN method is too biased towards the learning of the majority class, resulting in the performance of the minority class is



poor. The SVM and LGR methods, although the accuracy is slightly lower, the performance is relatively balanced, which shows that in the classification of FNC data, the traditional classifier represented by SVM and LGR has more advantages. In addition, if we compare Table 1 with Table 4, we will find that for traditional models, LDA and KNN methods perform poorly on both data, while LGR performs better on FNC data, and the classification performance of SVM method on the two types of data are similar. For deep learning models, since independent component time series data has temporal features in addition to spatial features, the use of convolution-GRU network can comprehensively learn two features to achieve a better classification effect than using CNN alone. Comparing Table 2 with Table 5, it can be found that in the classification task of this study, deep learning models have better classification performance, while most traditional classifiers perform poorly. On the one hand, this shows that the deep learning model is more suitable for the classification task of this study than the traditional classification model. On the other hand, it also shows that the data used in this study have a high nonlinearity and are not suitable for classification with a linear classifier.

#### 4. Discussion

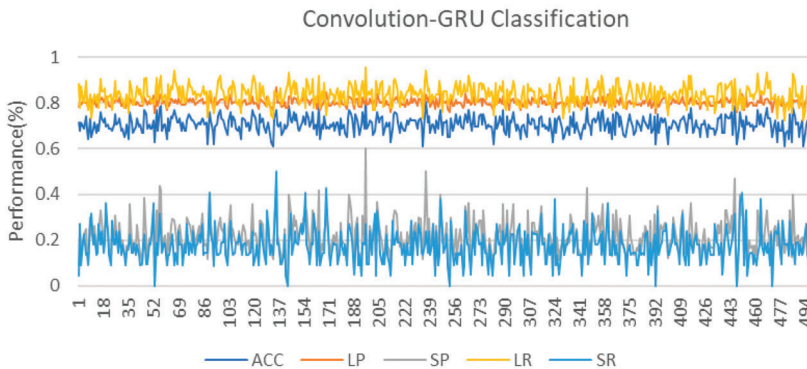
For a long time, the classification of data with a small and imbalanced sample size has been a troublesome problem, which is especially true for fMRI data. In this study, in order to solve the problem of small and imbalanced fMRI data, we proposed a three-step method. First, the ICA method is used to reduce the dimension of the data, and then the SMOTE method is used to generate the samples of the minority class to balance the number of samples. Finally, the convolution-GRU model is used and achieves an accuracy score of 72.2%. In this section, we mainly discuss the following three points: (1) The reason why the convolution-GRU method is more suitable for the fMRI data; (2) The influence of MOT task and data division on classification performance; (3) The relationship between ICA results and episodic memory studies.

In the classification of the public dataset used in this experiment, considering that too complex models may reduce the learning effect of small sample data, we used traditional classifiers such as LGR and SVM instead of CNN at first. However, as shown in the previous results, the traditional classifiers perform poorly on the fMRI data used in this study. It is because the feature dimension in this study is high and the nonlinearity is strong. SVM, LGR, LDA and other classifiers are essentially linear classifiers, which require a higher degree of linear separability of data. If they are used to classify high-dimensional nonlinear data, the effect will be reduced. Although the KNN method has no assumptions about the data and can be used for nonlinear classification, it is more sensitive to the size of the data and the degree of imbalance between the data, resulting a poor performance in classification. For these reasons, this study used ICA to reduce the dimension of the original data, and then used the SMOTE method to balance the number of two classes of samples. Finally, a convolution-GRU classifier is used to comprehensively learn the temporal and spatial features of the independent components of the data. While improving the classification performance of the majority class, the classification accuracy of the minority class is guaranteed. In addition, in this study, the FNC matrix is used as an input for classification, but since the FNC matrix does not contain temporal features, the classification performance is not as good as when independent components are used as training data.

As can be seen from the classification results, although the classification performance of the convolution-GRU classifier is better compared to traditional classifiers, the performance of the minority class is not as good as that of the majority class. On the one hand, it is due to the small sample size of the data used in this study and the high degree of non-linearity of the data. The dataset we used in this study comes from only 15 subjects, and due to the task design, the sample sizes of two classes are imbalanced. The random oversampling method only gives the random copies of the minority class, which means it cannot give more information about the class. The basic opinion of SMOTE and ADASYN is to generate new samples based on the Euclidean distance, when the data size is adequate, the two methods have enough chance to mix the samples from the minority class and generate new ones. However, for the fMRI data in this study, the lack of data makes the problem of the minority



class more serious. The generated data are easier to be influenced by the noise and individual bias. We think it is an important reason why there is a high difference between the classification of the majority and minority class. On the other hand, it is also related to the task. The data used in this study come from an episodic memory experiment. Using the words from word-scene pairs to do multi-target MOT and single-target MOT tasks ten times each, and only performing multi-target MOT tasks five times for the lure words, this caused an imbalance between the two types of data. In addition, too complex tasks are also one of the reasons for the small distinction between the two task states. The difficulty of the MOT task is increased in the task, which makes the subjects pay more attention to the target tracking situation of the MOT rather than the recall of the scene when completing the task. The design of the task adopts the block design method, and there is no long resting state between different tasks. This makes the brain activation state of the subject between tasks affect each other, resulting in a smaller degree of discrimination and an increasing of the difficulty in classification. However, this does not mean that the brain state of the subjects is the same when doing two types of task. In fact, in the process of cross-validation, we have also observed several ideal classification effects. The 10-fold cross-validation used to evaluate the performance of the classifiers in this study also reflects the influence of data partitioning on the performance of the classifiers. In order to study this effect, this study further conducted 500 cross-validations on the convolution-GRU network, and obtained the classifier performance that changed with different data division conditions, as shown in the Figure 7. It can be seen that as the data division continues to change, the performance of the classifier also fluctuates. That means, there are some cases where some samples from the minority class data that are more clearly distinguished from the majority class samples are divided into the training set, making the classifier learns more unique features of the minority class and improves the classification performance.



**Figure 7.** The result of 500 cross-validations on the convolution-GRU network. It should be noted that we are still performing 10-fold cross-validation, so 500 cross-validation is actually 50 rounds of 10-fold cross-validation. The changes in classification performance are almost entirely derived from data division and classifier parameter initialization.

The purpose of this study is to classify the fMRI data of subjects in different task states. The biggest difference between the two categories of tasks is whether the words shown to the subjects are the words in the word-scene pairs they memorized in the previous stages. Subjects associated the given words with the scene in the pre-stage of the task and memorized this association. Obviously, the memory of the word-scene pairs constitutes a simple episodic memory. In a successful retrieval process of episodic memory, when the test word is shown, the subjects will soon be familiar with the word and remember the scene associated with the word and even the other detail of the word-scene learning session. Therefore, the classification of the fMRI data of the two tasks in this study is actually to distinguish whether the tasks performed by the subjects caused the retrieval process of episodic

memory. Some past studies have shown that the recall process of episodic memory is related to the ventral parietal cortex [37,38]. There are also some studies show that the prefrontal lobe is also involved in the recall behavior of memory [39]. Compared to unsuccessful recall process, the neural activities of the “core recollection network” [40], which includes left angular gyrus, posterior cingulate cortex, medial prefrontal cortex, hippocampus, and parahippocampal cortex, are reported to be greater in successful recall process [39]. Similar results are observed in the ICA results of fMRI in this study, which shows that the task design of the word-scene pair successfully caused the recall process of the subject’s episodic memory. The ICA is proven to be a good tool to study the patterns of brain activities [15]. From past studies, we can learn the patterns indirectly by the weight of feature in classifiers, or some statistical methods, while the ICA gives us a chance to study the independent components in brain activities directly and can be combined with classification methods. However, in addition to the recall of the word-scene pairs, the MOT task is also introduced. This task is used as an interference task, which will compete with the activation of the episodic memory for visual resources and will interfere with the recall process. We think this is also the reason why the two tasks are difficult to distinguish.

In summary, this study proposes a convolutional GRU neural network based on the SMOTE method to classify small and imbalanced fMRI data from the episodic memory task with MOT, which obtained a better performance compared to traditional classifiers. At the same time, the classification results in this study also reflect whether there is a significant difference in the activation of relevant areas of the brain when the subjects perform the episodic memory task. This helps us to further understand the changes in the state of brain when people are intermittently performing the recall process of episodic memory, which has played a guiding role in the design of episodic memory-related experiments. It further reveals the operating mechanism of the memory function of the human brain under complex and diverse task environments and task conditions.

**Author Contributions:** Conceptualization, S.W.; methodology, S.W.; software, S.W.; validation, S.W., F.D. and M.Z.; formal analysis, S.W.; investigation, S.W. and M.Z.; resources, F.D. and M.Z.; data curation, S.W. and M.Z.; writing—original draft preparation, S.W.; writing—review and editing, S.W., F.D. and M.Z.; visualization, S.W. and M.Z.; supervision, F.D.; project administration, F.D.; funding acquisition, F.D. All authors have read and agreed to the published version of the manuscript.

**Funding:** This work was supported by the National Key R&D Program of China (No. 2017YFE0129700), the National Natural Science Foundation of China (No. 61673224), the Research Fellowship for International Young Scientists (No. 61850410526, 61850410524), and the Tianjin Natural Science Foundation for Distinguished Young Scholars (No. 18JCJC46100).

**Conflicts of Interest:** The authors declare no conflict of interest.

## References

1. Logothetis, N.K.; Pauls, J.; Augath, M.; Trinath, T.; Oeltermann, A. Neurophysiological investigation of the basis of the fMRI signal. *Nature* **2001**, *412*, 150–157. [[CrossRef](#)]
2. Friston, K.J.; Holmes, A.P.; Poline, J.B.; Grasby, P.J.; Williams, S.C.; Frackowiak, R.S.; Turner, R. Analysis of fMRI time-series revisited. *Neuroimage* **1995**, *2*, 45–53. [[CrossRef](#)]
3. Beckmann, C.F.; Jenkinson, M.; Smith, S.M. General multilevel linear modeling for group analysis in FMRI. *Neuroimage* **2003**, *20*, 1052–1063. [[CrossRef](#)]
4. Culham, J.C.; Brandt, S.A.; Cavanagh, P.; Kanwisher, N.G.; Dale, A.M.; Tootell, R.B.H. Cortical fMRI activation produced by attentive tracking of moving targets. *J. Neurophysiol.* **1998**, *80*, 2657–2670. [[CrossRef](#)]
5. Sommer, I.E.C.; Ramsey, N.F.; Kahn, R.S. Language lateralization in schizophrenia, an fMRI study. *Schizophr. Res.* **2001**, *52*, 57–67. [[CrossRef](#)]
6. LaConte, S.; Strother, S.; Cherkassky, V.; Anderson, J.; Hu, X.P. Support vector machines for temporal classification of block design fMRI data. *Neuroimage* **2005**, *26*, 317–329. [[CrossRef](#)] [[PubMed](#)]
7. Laconte, S.M.; Peltier, S.J.; Hu, X.P. Real-time fMRI using brain-state classification. *Hum. Brain Mapp.* **2007**, *28*, 1033–1044. [[CrossRef](#)] [[PubMed](#)]

8. Savio, A.; Grana, M. Local activity features for computer aided diagnosis of schizophrenia on resting-state fMRI. *Neurocomputing* **2015**, *164*, 154–161. [[CrossRef](#)]
9. Ryali, S.; Supekar, K.; Abrams, D.A.; Menon, V. Sparse logistic regression for whole-brain classification of fMRI data. *Neuroimage* **2010**, *51*, 752–764. [[CrossRef](#)] [[PubMed](#)]
10. Lv, J.L.; Jiang, X.; Li, X.; Zhu, D.J.; Chen, H.B.; Zhang, T.; Zhang, S.; Hu, X.T.; Han, J.W.; Huang, H.; et al. Identifying Functional Networks via Sparse Coding of Whole Brain fMRI Signals. In Proceedings of the International IEEE/EMBS Conference on Neural Engineering, CNE, San Diego, CA, USA, 6–8 November 2013; pp. 778–781.
11. Lv, J.L.; Ling, B.B.; Li, Q.Y.; Zhang, W.; Zhao, Y.; Jiang, X.; Guo, L.; Han, J.W.; Hu, X.T.; Guo, C.; et al. Task fMRI data analysis based on supervised stochastic coordinate coding. *Med. Image Anal.* **2017**, *38*, 1–16. [[CrossRef](#)]
12. Jang, H.; Plis, S.M.; Calhoun, V.D.; Lee, J.H. Task-specific feature extraction and classification of fMRI volumes using a deep neural network initialized with a deep belief network: Evaluation using sensorimotor tasks. *Neuroimage* **2017**, *145*, 314–328. [[CrossRef](#)] [[PubMed](#)]
13. Liu, Z.M.; de Zwart, J.A.; van Gelderen, P.; Kuo, L.W.; Duyn, J.H. Statistical feature extraction for artifact removal from concurrent fMRI-EEG recordings. *Neuroimage* **2012**, *59*, 2073–2087. [[CrossRef](#)] [[PubMed](#)]
14. Yan, W.Z.; Calhoun, V.; Song, M.; Cui, Y.; Yan, H.; Liu, S.F.; Fan, L.Z.; Zuo, N.M.; Yang, Z.Y.; Xu, K.B.; et al. Discriminating schizophrenia using recurrent neural network applied on time courses of multi-site FMRI data. *Ebiomedicine* **2019**, *47*, 543–552. [[CrossRef](#)] [[PubMed](#)]
15. Calhoun, V.D.; Liu, J.; Adali, T. A review of group ICA for fMRI data and ICA for joint inference of imaging, genetic, and ERP data. *Neuroimage* **2009**, *45*, S163–S172. [[CrossRef](#)]
16. Dvornek, N.C.; Ventola, P.; Pelphrey, K.A.; Duncan, J.S. Identifying Autism from Resting-State fMRI Using Long Short-Term Memory Networks. *Lect. Notes Comput. Sci.* **2017**, *10541*, 362–370. [[CrossRef](#)]
17. Huang, H.; Hu, X.T.; Zhao, Y.; Makkie, M.; Dong, Q.L.; Zhao, S.J.; Guo, L.; Liu, T.M. Modeling Task fMRI Data Via Deep Convolutional Autoencoder. *IEEE Trans. Med. Imaging* **2018**, *37*, 1551–1561. [[CrossRef](#)]
18. Zhang, Z.W.; Duan, F.; Sole-Casals, J.; Dinares-Ferran, J.; Cichocki, A.; Yang, Z.L.; Sun, Z. A Novel Deep Learning Approach with Data Augmentation to Classify Motor Imagery Signals. *IEEE Access* **2019**, *7*, 15945–15954. [[CrossRef](#)]
19. Sun, Z.; Huang, Z.H.; Duan, F.; Liu, Y. A Novel Multimodal Approach for Hybrid Brain-Computer Interface. *IEEE Access* **2020**, *8*, 89909–89918. [[CrossRef](#)]
20. Suk, H.I.; Shen, D.G. Deep Learning-Based Feature Representation for AD/MCI Classification. In Proceedings of the Medical Image Computing and Computer-Assisted Intervention—MICCAI 2013, Nagoya, Japan, 22–26 September 2013; Volume 8150, pp. 583–590.
21. Zhang, J.; Chen, L. Clustering-based undersampling with random over sampling examples and support vector machine for imbalanced classification of breast cancer diagnosis. *Comput. Assist. Surg.* **2019**, *24*, 62–72. [[CrossRef](#)]
22. Chawla, N.V.; Bowyer, K.W.; Hall, L.O.; Kegelmeyer, W.P. SMOTE: Synthetic Minority Over-sampling Technique. *J. Artif. Intell. Res.* **2011**, *16*, 321–357. [[CrossRef](#)]
23. He, H.B.; Bai, Y.; Garcia, E.A.; Li, S.T. ADASYN: Adaptive Synthetic Sampling Approach for Imbalanced Learning. In Proceedings of the IEEE International Joint Conference on Neural Networks, Hong Kong, China, 1–8 June 2008; pp. 1322–1328.
24. Eslami, T.; Saeed, F. Auto-ASD-Network: A Technique Based on Deep Learning and Support Vector Machines for Diagnosing Autism Spectrum Disorder using fMRI Data. In Proceedings of the 10th ACM International Conference on Bioinformatics, Computational Biology and Health Informatics, Niagara Falls, NY, USA, 7–10 September 2019; pp. 646–651.
25. Riaz, A.; Asad, M.; Alonso, E.; Slabaugh, G. Fusion of fMRI and non-imaging data for ADHD classification. *Comput. Med. Imag. Graph.* **2018**, *65*, 115–128. [[CrossRef](#)] [[PubMed](#)]
26. Koh, J.E.W.; Jahmunah, V.; Pham, T.H.; Oh, S.L.; Ciaccio, E.J.; Acharya, U.R.; Yeong, C.H.; Fabell, M.K.M.; Rahmat, K.; Vijayananthan, A.; et al. Automated detection of Alzheimer’s disease using bi-directional empirical model decomposition. *Pattern Recognit. Lett.* **2020**, *135*, 106–113. [[CrossRef](#)]
27. Faria, F.A.; Cappabianco, F.A.; Li, C.S.R.; Ide, J.S. Information Fusion for Cocaine Dependence Recognition using fMRI. In Proceedings of the International Conference on Pattern Recognition, Cancun, Mexico, 4–8 December 2016; pp. 1107–1112.

28. Poldrack, R.A.; Gorgolewski, K.J. OpenfMRI: Open sharing of task fMRI data. *Neuroimage* **2017**, *144*, 259–261. [[CrossRef](#)] [[PubMed](#)]
29. Poppenk, J.; Norman, K.A. Multiple-object Tracking as a Tool for Parametrically Modulating Memory Reactivation. *J. Cogn. Neurosci.* **2017**, *29*, 1339–1354. [[CrossRef](#)]
30. Cho, K.; van Merriënboer, B.; Gulcehre, C.; Bahdanau, D.; Bougares, F.; Schwenk, H.; Bengio, Y. Learning Phrase Representations using RNN Encoder-Decoder for Statistical Machine Translation. *arXiv* **2014**, arXiv:1406.1078.
31. Friston, K.J.; Holmes, A.P.; Worsley, K.J.; Poline, J.-P.; Frith, C.D.; Frackowiak, R.S.J. Statistical parametric maps in functional imaging: A general linear approach. *Hum. Brain Mapp.* **1994**, *2*, 189–210. [[CrossRef](#)]
32. Duan, F.; Huang, Z.; Sun, Z.; Zhang, Y.; Zhao, Q.; Cichocki, A.; Yang, Z.; Sole-Casals, J. Topological Network Analysis of Early Alzheimer’s Disease Based on Resting-State EEG. *IEEE Trans. Neural Syst. Rehabil. Eng.* **2020**, *28*, 2164–2172. [[CrossRef](#)]
33. Tzourio-Mazoyer, N.; Landeau, B.; Papathanassiou, D.; Crivello, F.; Etard, O.; Delcroix, N.; Mazoyer, B.; Joliot, M. Automated anatomical labeling of activations in SPM using a macroscopic anatomical parcellation of the MNI MRI single-subject brain. *Neuroimage* **2002**, *15*, 273–289. [[CrossRef](#)]
34. Egolf, E.A.; Calhoun, V.D.; Kiehl, K.A. Group ICA of fMRI Toolbox (GIFT). *Biol. Psychiatry* **2004**, *55*, 8S.
35. Zhang, X.; Kou, W.X.; Chang, E.I.C.; Gao, H.; Fan, Y.B.; Xu, Y. Sleep stage classification based on multi-level feature learning and recurrent neural networks via wearable device. *Comput. Biol. Med.* **2018**, *103*, 71–81. [[CrossRef](#)]
36. Martinc, M.; Pollak, S. Combining n-grams and deep convolutional features for language variety classification. *Nat. Lang. Eng.* **2019**, *25*, 607–632. [[CrossRef](#)]
37. Rugg, M.D.; King, D.R. Ventral lateral parietal cortex and episodic memory retrieval. *Cortex* **2018**, *107*, 238–250. [[CrossRef](#)] [[PubMed](#)]
38. Shimamura, A.P. Episodic retrieval and the cortical binding of relational activity. *Cogn. Affect. Behav. Neurosci.* **2011**, *11*, 277–291. [[CrossRef](#)] [[PubMed](#)]
39. King, D.R.; de Chastelaine, M.; Elward, R.L.; Wang, T.H.; Rugg, M.D. Recollection-Related Increases in Functional Connectivity Predict Individual Differences in Memory Accuracy. *J. Neurosci.* **2015**, *35*, 1763–1772. [[CrossRef](#)] [[PubMed](#)]
40. Elward, R.L.; Vilberg, K.L.; Rugg, M.D. Motivated Memories: Effects of Reward and Recollection in the Core Recollection Network and Beyond. *Cereb. Cortex* **2015**, *25*, 3159–3166. [[CrossRef](#)]

**Publisher’s Note:** MDPI stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.










© 2020 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<http://creativecommons.org/licenses/by/4.0/>).



Article

# Shadow Estimation for Ultrasound Images Using Auto-Encoding Structures and Synthetic Shadows

Suguru Yasutomi <sup>1,2,3,\*</sup>, Tatsuya Arakaki <sup>4</sup>, Ryu Matsuoka <sup>2,4</sup> , Akira Sakai <sup>1,2,5</sup>, Reina Komatsu <sup>2,4</sup>, Kanto Shozu <sup>6</sup> , Ai Dozen <sup>6</sup> , Hidenori Machino <sup>6,7</sup> , Ken Asada <sup>6,7</sup> , Syuzo Kaneko <sup>6,7</sup>, Akihiko Sekizawa <sup>4</sup> , Ryuji Hamamoto <sup>5,6,7</sup> and Masaaki Komatsu <sup>6,7,\*</sup> 

- <sup>1</sup> Artificial Intelligence Laboratory, Fujitsu Laboratories Ltd., 4-1-1 Kamikodanaka, Nakahara-Ku, Kawasaki, Kanagawa 211-8588, Japan; akira.sakai@fujitsu.com
- <sup>2</sup> RIKEN AIP-Fujitsu Collaboration Center, RIKEN Center for Advanced Intelligence Project, 1-4-1 Nihonbashi, Chuo-ku, Tokyo 103-0027, Japan; ryu@med.showa-u.ac.jp (R.M.); rkomatsu@med.showa-u.ac.jp (R.K.)
- <sup>3</sup> Department of Electronic and Information Engineering, Tokyo University of Agriculture and Technology, 2-24-16 Naka-cho, Koganei-shi, Tokyo 184-8588, Japan
- <sup>4</sup> Department of Obstetrics and Gynecology, Showa University School of Medicine, 1-5-8 Hatanodai, Shinagawa-Ku, Tokyo 142-8666, Japan; arakakit@med.showa-u.ac.jp (T.A.); sekizawa@med.showa-u.ac.jp (A.S.)
- <sup>5</sup> Biomedical Science and Engineering Track, Graduate School of Medical and Dental Sciences, Tokyo Medical and Dental University, 1-5-45 Yushima, Bunkyo-ku, Tokyo 113-8510, Japan; rhamamot@ncc.go.jp
- <sup>6</sup> Division of Molecular Modification and Cancer Biology, National Cancer Center Research Institute, 5-1-1 Tsukiji, Chuo-ku, Tokyo 104-0045, Japan; kshozu@ncc.go.jp (K.S.); adozen@ncc.go.jp (A.D.); hidenori.machino@riken.jp (H.M.); ken.asada@riken.jp (K.A.); sykaneko@ncc.go.jp (S.K.)
- <sup>7</sup> Cancer Translational Research Team, RIKEN Center for Advanced Intelligence Project, 1-4-1 Nihonbashi, Chuo-ku, Tokyo 103-0027, Japan
- \* Correspondence: yasutomi.suguru@fujitsu.com (S.Y.); masaaki.komatsu@riken.jp (M.K.); Tel.: +81-3-3547-5201 (M.K.)



**Citation:** Yasutomi, S.; Arakaki, T.; Matsuoka, R.; Sakai, A.; Komatsu, R.; Shozu, K.; Dozen, A.; Machino, H.; Asada, K.; Kaneko, S.; et al. Shadow Estimation for Ultrasound Images Using Auto-Encoding Structures and Synthetic Shadows. *Appl. Sci.* **2021**, *11*, 1127. <https://doi.org/10.3390/app11031127>

Academic Editors: Byung-Gyu Kim and Jordi Solé-Casals

Received: 19 December 2020

Accepted: 22 January 2021

Published: 26 January 2021

**Publisher's Note:** MDPI stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.



**Copyright:** © 2021 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

**Abstract:** Acoustic shadows are common artifacts in medical ultrasound imaging. The shadows are caused by objects that reflect ultrasound such as bones, and they are shown as dark areas in ultrasound images. Detecting such shadows is crucial for assessing the quality of images. This will be a pre-processing for further image processing or recognition aiming computer-aided diagnosis. In this paper, we propose an auto-encoding structure that estimates the shadowed areas and their intensities. The model once splits an input image into an estimated shadow image and an estimated shadow-free image through its encoder and decoder. Then, it combines them to reconstruct the input. By generating plausible synthetic shadows based on relatively coarse domain-specific knowledge on ultrasound images, we can train the model using unlabeled data. If pixel-level labels of the shadows are available, we also utilize them in a semi-supervised fashion. By experiments on ultrasound images for fetal heart diagnosis, we show that our method achieved 0.720 in the DICE score and outperformed conventional image processing methods and a segmentation method based on deep neural networks. The capability of the proposed method on estimating the intensities of shadows and the shadow-free images is also indicated through the experiments.

**Keywords:** ultrasound images; shadow detection; shadow estimation; deep learning; auto-encoders; semi-supervised learning

## 1. Introduction

Ultrasound (US) imaging is a popular modality of medical imaging. It is the first choice of diagnostic imaging because of these advantages: (i) it is noninvasive and has no side effects like X-rays and computed tomography (CT), (ii) equipment for US is smaller and cheaper than that of CT and magnetic resonance imaging (MRI), and (iii) it has higher temporal resolution (typically around 10–100 frames per second [1]) than CT and MRI. US imaging is used for a wide range of medical fields [1]; typically it is employed to

examine superficial organs, intra-abdominal organs, hearts, and fetuses. On the other hand, US imaging suffers from low spatial resolution (typically around  $10\ \mu\text{m}$ – $1\ \text{mm}$  [2]) and artifacts. Resulting images are often very noisy, and small findings and structures can be difficult to see.

To alleviate noise in US images and to support diagnosis, a number of technologies have been proposed. Recent equipment for US imaging comes with several techniques to improve the quality of US images [3]. For example, emitting sound waves from multiple different angles [4] and utilizing low-frequency bands that are hard to attenuate [5]. From the perspective of image processing, image enhancement methods have also been proposed [6,7].

Despite these techniques, acoustic shadows [8] (hereinafter simply referred to as shadows) that are common artifacts in US images are problematic. Shadows are shown as dark areas in US images. They are mainly caused by bones and air which reflect or absorb US emitted by probes. We cannot retrieve information of the areas that US does not reach, and thus, the areas are dark and dimmed. In some senses, shadows can be features for making a diagnosis; comet-tail artifact is known as a feature for finding gallstones [9], for example. However, we focus on situations that we are interested in structures of the organs such as fetal heart diagnosis. Because the regions of shadows have less information than shadow-free areas, clinicians can hardly make the right diagnosis if target organs are covered by shadows in such situations. Moreover, such shadows can degrade the performance of the image recognition methods for US images [10–15] although they are advancing lately with the rise of deep neural networks (DNNs) [16]. The only way to fundamentally avoid shadows is to move the probes so that the sound waves do not run into the obstacles. Shadows are basically unavoidable, but detecting such shadows is useful for assessing the quality of US images; whether the images can be used for diagnosis or image recognition techniques. Especially, for computer-aided diagnosis systems that detect structures such as [17], shadows can be critical for its performance. If shadows are detected while the US images are taking, we can notify the examiners whether the quality of the images is adequate or instruct them to retake the images if needed. Hence, shadows themselves have almost no information, but detecting them is crucial.

In this paper, we propose a shadow estimation method based on auto-encoding structures [18], a form of DNNs. The auto-encoding structures are constructed and trained to encode input images to feature vectors, decode the feature into images of estimated shadows and images of estimated shadow-free input, and then combine them to reconstruct the input. The structures enable us to obtain estimated shadows and estimated clean images at the same time. The primal target of the method is to estimate shadows, and the estimated clean images are supplementary outputs. The method is trained to localize shadows and estimate their intensity (or brightness) rather than just segmenting them as a pixel-wise binary classification (i.e., segmentation). Estimating the intensity is novel and motivated by the fact that shadows are often semi-transparent. By knowing the intensities of shadows, we can ignore detected shadows if they have low intensities. Considering the semi-transparency of shadows, labeling is quite difficult because the correct intensity of shadows are unknown; even annotating binary labels is difficult because of the ambiguity and variety of shadows. To address this problem, we introduce synthetic shadows as pseudo labels. If the target and method of the examination are fixed (i.e., the domain of US images is fixed), we can get to know the possible shapes of shadows. Based on the prior knowledge, we generate and inject simulated shadows with random shapes and intensities into the images and make the method learn them. In this way, shadows with any intensity can be learned without giving labeled data. Additionally, we also utilize pixel-wise binary labels if available in a semi-supervised fashion. An algorithm that estimates the intensity of the labeled shadows is proposed and the labels are turned into semi-transparent ones. We applied the proposed method to US images for fetal heart diagnosis and evaluated the performance. As a segmentation method, our method outperformed previous methods based on image processing in a situation without labels. Besides, in situations with labels,



it achieved comparable performance against a reference segmentation method based on DNNs. The effectiveness of the estimated shadow intensity was shown by the correlation between the estimation and the brightness of the input. The quality of the estimated input without shadows was also evaluated qualitatively.

## 2. Related Work

The necessity of detecting shadows in US images is known and methods that are based on rather traditional image processing have been proposed [19,20]. In [19], procedures of US image generation and the causes of shadows are modeled. The US images are analyzed along the scanlines and shadows are detected as ruptures of brightness. Segmentation based on random walks is employed in [20]. The idea of this method is that the upper parts of the images are more reliable because the probes are close. The method basically estimates confidence maps of US images but it can be considered as a shadow detection method. This random walk method has been improved by focusing on shadows caused by bones [21]. In recent years, DNN based methods have also been proposed and improved detection performance [22,23]. Generally, DNNs require many labeled data to achieve high performance but pixel-level labels of shadows are expensive. In [22,23], weakly supervised learning is applied to resolve this problem. Assuming that the image-level labels are low cost, many US images are annotated whether they have shadows or not. They illustrated that shadows can be detected effectively by training DNNs using these weakly labeled examples and a small amount of pixel-level annotated data. Since the image-level labels are actually also expensive and difficult to collect due to ambiguity and semi-transparency of shadows, in this study, we focus on utilizing unlabeled data supported by coarse domain-specific knowledge. A combination of the traditional shadow detecting method [19] and DNN based segmentation for US images is also proposed [24]. It shows that the segmentation results can be improved by knowing the presence of shadows and it is important to detect shadow precisely.

Auto-encoders [18] are popular unsupervised learning methods for DNNs. They consist of encoders and decoders, and the encoders compress an input into a latent vector and the decoders reconstruct it to the input. In this way, DNNs can learn features of training data like the principal component analysis [25]. Auto-encoding structures are simple, but there are many variants and applications [26,27] thanks to DNNs' high expression capability. Semi-supervised learning is one of the applications [27,28]. By efficiently extracting features from much unlabeled data in an unsupervised manner, classification problems are solved using a small amount of labeled data. Encoder-decoder structures, which are constructed just like auto-encoders but do not reconstruct the input, are often used for segmentation [29]. Especially, U-Net [30] is known as a standard method for medical images and it is applied to US images as well [12,31]. Encoder-decoder structures can be employed to generate images. For example, in [32], a two-way encoder-decoder structure generates relighted photos that come with lighting with desired direction and color temperature. It is trained in a supervised fashion to generate an intermediate shadow-free image and prior image of the desired lighting and to combine them into a final relighted output. We employ a similar structure for shadow estimation in US images, but we train it also as an auto-encoder to effectively utilize unlabeled data.

## 3. Materials and Methods

In this section, we introduce a DNN that has an auto-encoding structure for estimating US shadows and the datasets for evaluating the method. We describe the structure and propose a training method with unlabeled data based on our preliminary work [33]. Then the proposed method is extended to additionally use data with pixel-level labels in a semi-supervised fashion.

### 3.1. Datasets

We evaluate the performance of our method on US images of fetal heart diagnosis. Data for the experiments were acquired in Showa University Hospital, Showa University Toyosu Hospital, Showa University Fujigaoka Hospital, and Showa University Northern Yokohama Hospital. All the experiments were conducted in accordance with the ethical committee of each hospital. We collected 157 videos of 157 women who are 18–34 weeks pregnant. All the data are taken by convex probes with fetal cardiac preset on Voluson E8 or E10 (GE Healthcare, Chicago, IL, USA).

We converted 107 of the videos into 37,378 images and used them as an unlabeled dataset. From the remaining 50 videos, experts extracted 445 images with shadows and annotated them at a pixel-level. Annotated 445 images were split into a training dataset with 259 images, a validation dataset with 91 images, and a testing dataset with 95 images (corresponding to 30, 12, and 8 videos, respectively).

### 3.2. Restricted Auto-Encoding Structure for Shadow Estimation

Let  $x \in [0, 1]^{H \times W}$  be an input grayscale US image with a size of  $H \times W$ . Its brightness is assumed to be normalized to  $[0, 1]$ . We introduce an encoder DNN  $E: [0, 1]^{H \times W} \rightarrow \mathbb{R}^m$  and a decoder DNN  $D: \mathbb{R}^m \rightarrow [0, 1]^{H \times W \times 2}$ , where  $m$  is the number of dimensions of a latent vector. Note that the decoder  $D$  outputs an image with two channels. An auto-encoding procedure that reconstructs the input  $x$  into  $\hat{x}$  is defined as

$$\hat{x} = \hat{s} \odot \hat{c}, \quad (1)$$

$$\hat{x} := D(E(x)), \hat{s} := \hat{x}_1, \hat{c} := \hat{x}_2, \quad (2)$$

where  $\odot$ ,  $\hat{x}_i$ ,  $\hat{s}$  and  $\hat{c}$  are element-wise product, the  $i$ -th channel of  $\hat{x}$ , the estimated shadow image, and the estimated clean image without shadows, respectively. For each element in the estimated shadow  $\hat{s}$ , 1.0 means that no shadows expected in the pixel, and the lower the value, the intensity of the estimated shadow is higher. Figure 1a shows the proposed auto-encoding structure.

The reconstruction  $\hat{x}$  is given as Equation (1) because we assume that the input image with shadows is generated as an element-wise product of an ideal shadow-free input image and an image of semi-transparent shadows. This is different from the actual generation process of US images but we model US images with shadows in this way for simplicity.

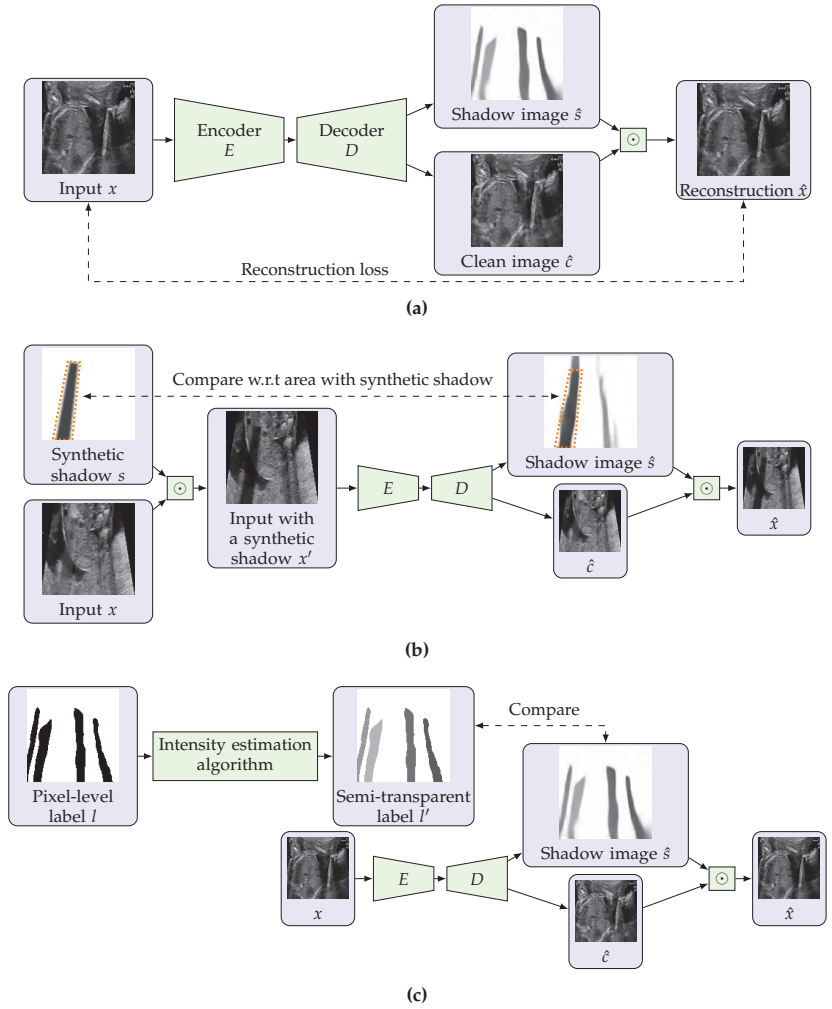
### 3.3. Training Using Unlabeled Data with Synthetic Shadows

Since the proposed model is based on auto-encoders, it is basically trained by minimizing a reconstruction loss given as

$$L_{\text{recon}}(x, \hat{x}) := \frac{1}{HW} \sum_{h=1}^H \sum_{w=1}^W (\hat{x}_{hw} - x_{hw})^2, \quad (3)$$

which is also known as the mean squared error (MSE). Of course, we cannot make the model split the input into the estimated shadow and the estimated clean image only by the reconstruction loss. To address this, we introduce synthetic shadows and a loss function that uses them as pseudo labels.

Annotating shadows is costly because the pixel-level label is expensive in the first place, and additionally, shadows are ambiguous. It is difficult to make a standard for labeling shadows that come in various intensities and are often blurred. However, the possible shapes of shadows are known when the domain is fixed; the target to be diagnosed and equipment such as probes are set. Once the shapes are determined, we can generate random plausible synthetic shadows in a rule-based manner. Then the synthetic shadows can be injected into the input image and can be used as pseudo labels. In this study, we focus on convex probes [1] that generate shadows shaped annular sectors. Details of the algorithm for generating shadows are described in Appendix A.



**Figure 1.** Overview of our shadow estimation method. (a) shows the proposed auto-encoding structure. (b,c) illustrate the learning process for unlabeled data and pixel-level labeled data, respectively. For unlabeled data, the estimated shadow  $\hat{s}$  is compared to the synthetic shadow with respect to the region that the synthetic shadow exists. For labeled data, the label is made semi-transparent based on the estimated intensity of labeled shadows, and  $\hat{s}$  is compared to it.

Assuming that a synthetic shadow image  $s \in [0, 1]^{H \times W}$  is given, we inject it to an input image  $x$  as follows;

$$x' = x \odot s, \tag{4}$$

and we use  $x'$  as a new input to the model. A loss function for training the model to predict shadows is defined as

$$L_{\text{synth}}(s, \hat{s}) := \frac{1}{HW} \sum_{h=1}^H \sum_{w=1}^W \mathbf{1}[s_{hw} \neq 1] (\hat{s}_{hw} - s_{hw})^2, \tag{5}$$

where  $\mathbf{1}[\cdot]$  is a function that returns 1 when the condition is met and returns 0 otherwise. Note that the loss  $L_{\text{synth}}$  evaluates the area that the synthetic shadow exists. This is because we do not know whether the original input already contains shadows. If the masking term  $\mathbf{1}[s_{hw} \neq 1]$  is omitted, the model learns the region without the synthetic shadow as a shadow-free region regardless of the presence of real shadows. Thanks to the mask, we can train the model to estimate the intensity of the synthetic shadow, but the whole estimated shadow tends to be dark. To prevent the estimated shadow  $\hat{s}$  from being too dark and to make the default output white, we also introduce an auxiliary regularization loss defined as

$$L_{\text{synthreg}}(\hat{s}) := \frac{1}{HW} \sum_{h=1}^H \sum_{w=1}^W |\hat{s}_{hw} - 1|. \tag{6}$$

A linear combination of the three losses introduced above is a loss for training with unlabeled US images. That is, the loss function is given as

$$L_{\text{unlabeled}}(x, s, \hat{x}, \hat{s}) := \lambda_{\text{recon}} L_{\text{recon}}(x, \hat{x}) + \lambda_{\text{synth}} L_{\text{synth}}(s, \hat{s}) + \lambda_{\text{synthreg}} L_{\text{synthreg}}(\hat{s}), \tag{7}$$

where  $\lambda_{\text{recon}}, \lambda_{\text{synth}}, \lambda_{\text{synthreg}} > 0$  are hyperparameters that decide the weight of each loss. This training procedure using unlabeled data and the synthetic shadows is illustrated in Figure 1b.

### 3.4. Use of Pixel-Level Labels and Extension to Semi-Supervised Learning

Pixel-level labels for US shadows are expensive as mentioned above, but if available, they can contribute to the improvement of the estimation performance. We assume that we have some pixel-level labeled data which labels are binary; whether each pixel is in shadows or not. Ideally, we expect labels that express semi-transparency of shadows but the correct intensities of shadows are unknown even for experts. Hence, we introduce a method to effectively utilize the binary labels for the proposed shadow estimation framework.

Let  $l \in \{0, 1\}^{H \times W}$  be a pixel-level binary label that represents where shadows exist. For each element in  $l$ , 0 and 1 correspond to a shadowed pixel and a shadow-free pixel, respectively. Recall that the proposed auto-encoding model is based on the idea: each input US image is considered to be generated by an element-wise product of the ideal shadow-free image and the shadow image. If the ideal shadow-free image  $x^*$  for an input  $x$  is available, we can calculate the intensity of the labeled shadow by

$$l_{hw}^* := \begin{cases} \frac{x_{hw}}{x_{hw}^*} & (l_{hw} = 0) \\ 1 & (l_{hw} = 1) \end{cases}. \tag{8}$$

However,  $x^*$  is actually unknown. Here, we estimate  $x^*$  as a mean brightness over the shadow-free area that is written as

$$\hat{x}^* := \frac{1}{\sum_{i,j} M_{ij}} \sum_{i=1}^H \sum_{j=1}^W M_{ij} x_{ij}, \tag{9}$$

where  $M \in \{0, 1\}^{H \times W}$  is a mask that represents the region without shadows. The mask  $M$  is given as

$$M_{hw} := \begin{cases} 1 & (l_{hw} = 1 \text{ and } x_{hw} > T) \\ 0 & (\text{otherwise}) \end{cases}, \tag{10}$$

where  $T \in [0, 1]$  is a given threshold to ignore almost completely black areas that have no shadows. In US images, liquids are shown in black. By thresholding, we can reject such areas and estimate  $x^*$  more precisely. Besides, for simplicity and stability of the training,

we assume that the intensity of each labeled shadow is constant. Hence, Equation (8) is rewritten and the resulting label with the estimated intensity is

$$l'_{hw} := \begin{cases} \frac{1}{\hat{x}^*} \cdot \frac{\sum_{z \in \mathcal{S}(x_{hw})} z}{|\mathcal{S}(x_{hw})|} & (l_{hw} = 0) \\ 1 & (l_{hw} = 1) \end{cases}, \tag{11}$$

where  $\mathcal{S}(x_{hw})$  is a subset of  $\{x_{ij}\} (i = 1, \dots, H, j = 1, \dots, W)$  that consists of  $x_{ij}$  inside the shadow that contains  $x_{ij}$ . Calculation of  $l'$  is summarized in Algorithm 1.

---

**Algorithm 1** Estimation of shadow intensities using a pixel-level binary label.

---

**Input:** A US image  $x \in [0, 1]^{H \times W}$ , a pixel-level label of shadows  $l \in \{0, 1\}^{H \times W}$ , and a threshold  $T$ .

**Output:** Semi-transparent label  $l' \in [0, 1]^{H \times W}$

- 1:  $M \leftarrow l \odot \mathbf{1}[x > T]$
  - 2:  $x^* \leftarrow \frac{1}{\sum_{i,j} M_{ij}} \sum_{i=1}^H \sum_{j=1}^W M_{ij} x_{ij}$
  - 3:  $l' \leftarrow l$
  - 4: **for** each labeled shadow  $l_c$  in  $l$  (i.e., each connected component  $l_c$  in  $l$  with a value 0) **do**
  - 5:   **for** each coordinate  $(i, j)$  that corresponds to  $l_c$  **do**
  - 6:      $l'_{ij} \leftarrow l'_{ij} + \frac{x_{ij}}{\hat{x}^* |l_c|}$
  - 7:   **end for**
  - 8: **end for**
- 

Using the estimated semi-transparent label  $l'$ , we can construct a loss function based on Equation (5) as follows;

$$L_{\text{label}}(l', \hat{s}) := \frac{1}{HW} \sum_{h=1}^H \sum_{w=1}^W (\hat{s}_{hw} - l'_{hw})^2. \tag{12}$$

In contrast to Equation (5), we do not need the masking term anymore because the label  $l'$  tells not only shadowed areas but also shadow-free areas. Besides, the regularization term given by Equation (6) is neither needed. The resulting loss function for labeled data is given as

$$L_{\text{labeled}}(x, l', \hat{x}, \hat{s}) := \lambda_{\text{recon}} L_{\text{recon}}(x, \hat{x}) + \lambda_{\text{label}} L_{\text{label}}(l', \hat{s}), \tag{13}$$

where  $\lambda_{\text{recon}}, \lambda_{\text{label}} > 0$  are hyperparameters. Figure 1c illustrates the training procedure with labeled data.

By switching Equations (7) and (13) according to the existence of labels, we can train the proposed model in a semi-supervised fashion that effectively utilizes both unlabeled data and labeled data. Assume that an unlabeled dataset  $\mathcal{D}_{\text{unlabeled}} = \{x^1, \dots, x^{N_{\text{unlabeled}}}\}$  and a labeled dataset  $\mathcal{D}_{\text{labeled}} = \{(x^1, l^1), \dots, (x^{N_{\text{labeled}}}, l^{N_{\text{labeled}}})\}$  are given, where  $N_{\text{unlabeled}}$  and  $N_{\text{labeled}}$  are the number of unlabeled data and that of labeled data, respectively. Then, a mini-batch  $\mathcal{B} \subset \mathcal{D}_{\text{unlabeled}} \cup \mathcal{D}_{\text{labeled}}$  can be drawn. The loss for the mini-batch is given as

$$L(\mathcal{B}) = \frac{1}{|\mathcal{B}|} \left[ \sum_{x \in \mathcal{B}_{\text{unlabeled}}} L_{\text{unlabeled}}(x, s, \hat{x}, \hat{s}) + \sum_{(x,l) \in \mathcal{B}_{\text{labeled}}} L_{\text{labeled}}(x, l', \hat{x}, \hat{s}) \right]. \tag{14}$$

Note that  $s$  and  $l'$  should be generated for each sample in the batch.

## 4. Results

### 4.1. Setting

The encoder  $E$  and the decoder  $D$  were constructed just like U-Net [30]. Details of the architecture are described in Appendix B. The model was optimized by Adam [34]. The size of the mini-batch was 32 and the number of training epochs was 10. The parameters of Adam were set to default except for the learning rate  $\alpha$  that was set  $10^{-4}$  initially and decayed to  $10^{-5}$  through 10 epochs of training. The weight for the reconstruction error  $\lambda_{\text{recon}}$  was fixed to 1.0. Other hyperparameters were set by grid search using the validation data. The search spaces of the hyperparameters are following:  $\lambda_{\text{synth}} = \lambda_{\text{label}} \in \{10^{-2}, 10^{-1}, 1, 10, 10^2, 10^3\}$ ,  $\lambda_{\text{synthreg}} \in \{10^{-3}, 10^{-2}, 10^{-1}, 1, 10\}$ , and  $v_{\text{min}} \in \{0.1, 0.5\}$  in Algorithm A1. Note that  $\lambda_{\text{synthreg}}$  was set to zero in the semi-supervised situations because we empirically found that the labels play the same role as the regularization term. The selected hyperparameters are shown in Appendix C. To stabilize the training, in semi-supervised situations, the labeled training dataset was oversampled so that the number of the labeled data was almost equal to the size of the unlabeled dataset. Specifically, the labeled data was simply repeated  $\lfloor N_{\text{unlabeled}}/N_{\text{labeled}} \rfloor$  times.

As a reference DNN based segmentation method, we used vanilla U-Net [30]. It was trained on pixel-wise cross-entropy and optimized by Adam with  $\alpha = 10^{-4}$ . The size of the mini-batch was 32. Because U-Net only uses the labeled dataset, the number of training epochs was set so that the number of iterations equals that of the proposed method to prevent underfitting. We empirically confirmed that the training of U-Net converged with this setting.

For both the proposed method and U-Net, the parameters of the models were saved every epoch. Then, the parameters that perform best for the validation dataset were selected. In addition, random cropping is applied to input data and they are resized to  $128 \times 128$  for both the methods.

The geometrical method [19] and the random walk method [20] are also used as references of methods that do not use labels. The parameters were in accordance with the original papers. Because the geometrical method outputs shadow detection results along scattered scanlines, we applied morphological closing filters to the results for adjusting to pixel-level detection.

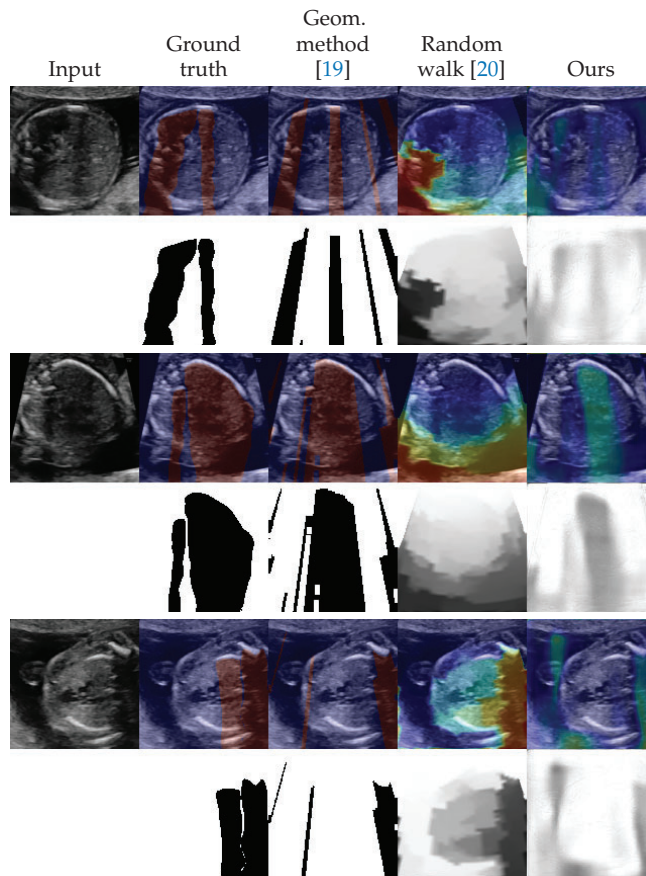
We show the experimental results on the testing dataset in the following sections. For all the experiments, the testing images are cropped so that they contain no meta-data and resized to  $128 \times 128$ . Quantitative results for the validation dataset is shown in Appendix D.

### 4.2. Shadow Detection

We evaluated the proposed method as a shadow detection method. In situations with labels, we investigated the performance on different numbers of labeled data. The numbers of labeled training data were set to 0, 42 (from 5 videos), 90 (from 10 videos), 177 (from 20 videos), and 259 (from all 30 videos). Since the proposed method estimates intensities of shadows as  $\hat{s}$ , a threshold to convert  $\hat{s}$  to binary was searched using the validation dataset. The threshold is selected from  $\{0.001, 0.002, \dots, 0.999\}$ . For the random walk method [20] which estimates confidence maps, we also searched and applied a threshold in the same way as the proposed method. The selected thresholds are shown in Appendix C. The detection performance was evaluated by the DICE score [35] which is also known as F1 measure. Table 1 shows the results in the DICE score. Figures 2 and 3 shows examples of shadow detection of the methods that do not use labels and the methods that use labels, respectively (for additional results, see Figures A2 and A3).

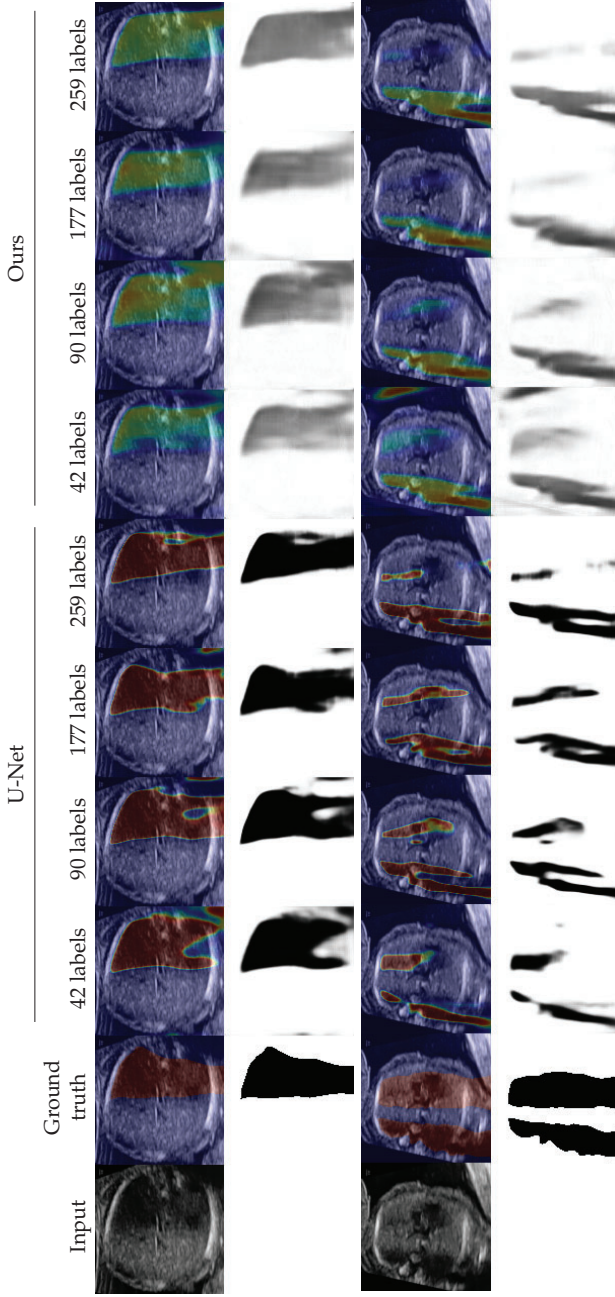
**Table 1.** Results of shadow detection evaluated in the DICE score. The scores are calculated for each testing image, and means over them are shown. The numbers in parentheses are the standard deviations.

Method	Number of Labeled Images				
	0	42 (5 Videos)	90 (10 Videos)	177 (20 Videos)	259 (30 Videos)
Geometric method [19]	0.193 (±0.210)	-	-	-	-
Random walk [20]	0.450 (±0.142)	-	-	-	-
U-Net [30]	-	0.610 (±0.184)	0.655 (±0.170)	0.681 (±0.136)	0.698 (±0.137)
Ours	0.578 (±0.164)	0.666 (±0.142)	0.686 (±0.148)	0.707 (±0.113)	0.720 (±0.151)



**Figure 2.** Examples of shadow detection results for the methods that do not use labels. The lower side of each example shows detection results, and the upper side shows them overlaid to the input image. For overlaid images, blue corresponds to low intensities and red corresponds to high intensities.





**Figure 3.** Examples of shadow detection results for the methods that use labels. The lower side of each example shows detection results, and the upper side shows them overlaid to the input image. For overlaid images, blue corresponds to low intensities and red corresponds to high intensities.

In the unlabeled situation, we can see that the proposed method outperformed the other two methods from Table 1. The standard deviation was large for the geometric method. The score of the geometric method is low but Figure 2 shows that performed well in some examples. The random walk method scored better than that of the geometric method and detected shadows that the geometric method could not. However, the outputs of the random walk method tended to simply reflect the distance from the probe. The proposed method estimated the shapes of the shadows well although the estimation results tend to be blurred. Figure 2 also indicates that the proposed method successfully estimated the intensities of the shadows.

In terms of the methods that use labeled data, that is, U-Net and the proposed method in semi-supervised situations, the proposed method performed slightly better than U-Net as Table 1 shows. We could see no clear trends for the standard deviations. The differences in the DICE score were larger when the numbers of labeled data were smaller. Figure 3 shows that the two methods detected shadows in almost the same performance. The proposed method expressed the intensities of the shadows while U-Net output the detection result in maps that are almost binary.

### 4.3. Shadow Intensity Estimation

We evaluated the performance of the proposed method in the estimation of intensities of shadows. Since we do not have the ground truth for shadow intensities, as a novel indicator, we calculated the correlations of the brightness of the input image and the estimation with respect to the area labeled as shadows. More specifically, given an input image  $x$ , a pixel-label  $l$ , and a shadow estimation  $\hat{s}$ , the indicator is calculated as follows;

$$\rho(x, l, \hat{s}) := \frac{\sum_{h=1}^H \sum_{w=1}^W (1 - l_{hw})(x_{hw} - \bar{x})(\hat{s}_{hw} - \bar{s})}{\sqrt{\sum_{h=1}^H \sum_{w=1}^W (1 - l_{hw})(x_{hw} - \bar{x})^2} \sqrt{\sum_{h=1}^H \sum_{w=1}^W (1 - l_{hw})(\hat{s}_{hw} - \bar{s})^2}}, \quad (15)$$

$$\bar{x} = \frac{\sum_{h=1}^H \sum_{w=1}^W (1 - l_{hw})x_{hw}}{\sum_{h=1}^H \sum_{w=1}^W (1 - l_{hw})}, \quad \bar{s} = \frac{\sum_{h=1}^H \sum_{w=1}^W (1 - l_{hw})\hat{s}_{hw}}{\sum_{h=1}^H \sum_{w=1}^W (1 - l_{hw})}, \quad (16)$$

which is Pearson’s correlation coefficient [36] that is masked using the label  $l$ . Table 2 shows the results. Since the conventional methods are designed for detecting shadows and not estimating their intensities, the coefficients for them were just for benchmarks. It illustrates that the proposed method achieved the largest coefficients for all the numbers of labeled data. This indicates that our method estimated the shadow intensities the most precisely among the methods we examined. The coefficients of U-Net are lower than that of the proposed method but it was stable. The methods based on image processing, the geometric method and the random walk, performed worse. Especially, the estimation of the random walk method is the worst and had almost no correlations to the input image, despite its better performance for shadow detection than the geometric method. The standard deviation of the random walk method was larger than other methods and the method seemed to be unstable for this indicator.

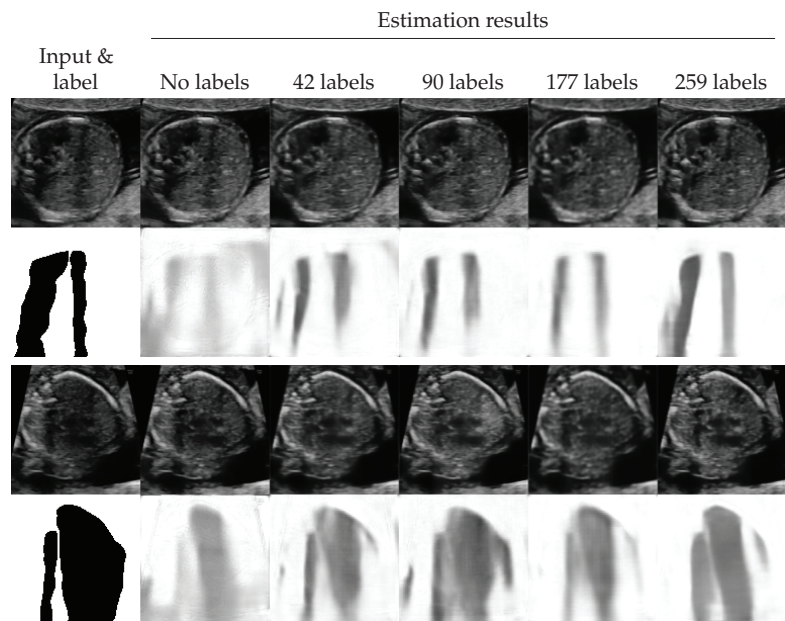
**Table 2.** Evaluation of the estimation of shadow intensities. Scores are the correlation coefficient calculated by Equation (15). The coefficients are calculated for each testing image, and means over them are shown. The numbers in parentheses are the standard deviations.

Method	Number of Labeled Images				
	0	42 (5 Videos)	90 (10 Videos)	177 (20 Videos)	259 (30 Videos)
Geometric method [19]	0.152 (±0.182)	-	-	-	-
Random walk [20]	-0.047 (±0.290)	-	-	-	-
U-Net [30]	-	0.308 (±0.150)	0.267 (±0.144)	0.262 (±0.158)	0.247 (±0.172)
Ours	0.351 (±0.155)	0.388 (±0.150)	0.414 (±0.159)	0.358 (±0.149)	0.349 (±0.162)

#### 4.4. Shadow Removal

The proposed method estimated shadows and shadow-free variants of the input images at the same time. We evaluated the quality of the estimated shadow-free images subjectively.

Figure 4 shows the examples of shadow removal performed by the proposed method (for additional results, see Figure A4). We observed that the areas with low-intensity shadows were efficiently enhanced. Additionally, the quality of the enhancement seemed better when the performance of shadow detection was better. In contrast, the shadows with high intensities tended to be just filled with blurred texture.



**Figure 4.** Examples of shadow removal results of the proposed method. The lower side of each example shows the labels and the detection results. The upper side shows the input images and the estimated shadow-free images.

## 5. Discussion

Among the shadow detecting methods which do not use labels, the proposed method detected shadows the most correctly. The other two methods could suffer from the difference of domains; the domains for which the methods were built and the domain of fetal heart diagnosis that we used in this paper. Besides, the standard deviation of the DICE score for the geometric method was large. This means that its detection performance varies among the testing images. This is probably because the method heavily depends on the domain for which it is designed. In that sense, our method has an advantage because it is data-driven. Although the proposed method also uses domain-specific knowledge, it only requires rough possible shapes of shadows that are determined mainly by the type of the probe. In the situations that the labels are available, the proposed method achieved comparable detection performance to U-Net which is a popular segmentation method for medical images. When only 42 labeled images were used for training, our method was better than U-Net. The auto-encoding structure possibly helped detection by extracting features by the unlabeled data. The maximum number of labeled data, 259 from 30 videos, was relatively small in the context of DNNs, but both the proposed method and U-Net performed well. In terms of detecting shadows, the dataset was clean and in a narrow domain, and it might be easy to detect shadows. From the perspective of data collection, we revealed that a couple of hundred of labeled data is enough for one domain. The amount is reasonable when it comes to accumulating data of different multiple domains.

One of the advantages of the proposed method is that it can estimate the intensities of shadows. Although we cannot obtain the ground truth of the intensities, our method estimated images of shadows which intensities are highly correlated to the brightness of the shadowed areas of the input images, at least. The correlation coefficients were higher than the other methods we used in the experiments. From the intensities of shadows, we can assess the quality of US images using them. If we detect shadows in a binary segmentation manner, the sizes of shadowed areas can be used for quality inspection. Our method can provide additional information, the intensities of shadows, and we can check US images based on it; an US image with large but light shadows should be allowed in some situations, for example.

Although removing shadows is not the main target of our work, notably, our method removed shadows of input US images without training on losses that directly lead the decoder to output clean image as  $\hat{c}$ . This result could come from the model structure that split the input into the estimated shadow and the estimated clean input and then compose them into the reconstruction. Besides, the estimated clean images were clear, thanks to the U-Net like structure that has skip-connections between the encoder and the decoder [37]. In a clinical sense, we cannot totally trust the estimated shadow-free images because the shadowed areas are completed statistically. It means that the images of popular and healthy cases are likely to appear even if the target of diagnosis has anomalies. However, generating shadow-free images can work as a pre-processing for image recognition techniques.

## 6. Conclusions

We proposed an auto-encoding structure-based DNN that estimates acoustic shadows in US images. The method estimates not only the location of shadows but also their intensities. We also introduced the loss functions for training the model on both unlabeled data and labeled data. For unlabeled data, synthetic shadows are generated using the knowledge that the probes decide the shapes of shadows, and used as pseudo labels. If binary pixel-level labels that tell us areas with shadows are given, they are effectively utilized by converting them to labels with estimated intensities. By experiments on US images for fetal heart diagnosis, we showed that our method detected shadows better than the conventional methods in the situation without labels and did better to the DNN-based segmentation method U-Net in the situations with labels available. In terms of estimating the intensities of shadows, the proposed method performed the best. Moreover,

we suggested the capability of our method in removing shadows as supplemental outputs, not just estimating them.

Although our method employs the auto-encoding structure that extracts features from input images, the difference in detection performance between fully-supervised U-Net and semi-supervised our method was small. One possible reason for this result is that the detection was easy; the images in the dataset were relatively clear and belonged to the narrow domain. Applying to datasets that are in different domains or have more variations of images has remained as one of the future work.

The proposed method can work with any US image recognition methods as a pre-processing. We can reject low-quality data based on the estimated shadow images. The use in such quality assessing ways is one of the possible future directions.

**Author Contributions:** Conceptualization, S.Y., T.A., R.M., and A.S. (Akira Sakai); methodology, S.Y., A.S. (Akira Sakai); software, S.Y.; validation, S.Y.; formal analysis, S.Y., T.A., and R.M.; investigation, S.Y., T.A., and R.M.; resources, T.A., R.M., R.K., and A.S. (Akihiko Sekizawa); data curation, S.Y., T.A., and R.M.; writing—original draft preparation, S.Y., T.A., and R.M.; writing—review and editing, A.S. (Akira Sakai), R.K., K.S., A.D., H.M., K.A., S.K., A.S. (Akihiko Sekizawa), R.H., and M.K.; visualization, S.Y.; supervision, R.H. and M.K.; project administration, S.Y., T.A., and R.M. All authors have read and agreed to the published version of the manuscript.

**Funding:** This work was supported by the subsidy for Advanced Integrated Intelligence Platform (MEXT), and the commissioned projects income for RIKEN AIP-FUJITSU Collaboration Center.

**Institutional Review Board Statement:** The study was conducted according to the guidelines of the Declaration of Helsinki, and approved by the Institutional Review Board of RIKEN, Fujitsu Ltd., Showa University, and the National Cancer Center (approval ID: Wako1 29-4).

**Informed Consent Statement:** This research protocol was approved by the medical ethics committees of the four collaborating research facilities, and data collection was conducted in an opt-out manner.

**Data Availability Statement:** Data sharing is not applicable owing to the patient privacy rights.

**Acknowledgments:** The authors wish to acknowledge Hisayuki Sano and Hiroyuki Yoshida for great supports, the medical doctors in the Showa University Hospitals for data collection.

**Conflicts of Interest:** R.H. has received the joint research grant from Fujitsu Ltd. The other authors declare no conflict of interest.

## Appendix A. Algorithm for Generating Synthetic Shadows

Algorithm A1 describes the process for generating synthetic shadows that correspond to convex probes. The parameters in the algorithm were set  $p = [-128, 64]$ ,  $(d_{\min}, d_{\max}) = (250, 290)[\text{deg}]$ ,  $(\theta_{\min}, \theta_{\max}) = (0, 10)[\text{deg}]$ ,  $(R_{\min}, R_{\max}) = (256, 256)$ ,  $r_{\min} = 128$ ,  $\delta_{\theta} = 1$ ,  $k = 10$ ,  $\sigma = 1.55$ , and  $v_{\max} = 1$  in all the experiments. The parameter  $v_{\min}$  is decided by the grid search as described in Section 4.1.

**Algorithm A1** Generation of annular sector shaped synthetic shadows. A function  $U(\cdot, \cdot)$  draws a sample from a uniform distribution.

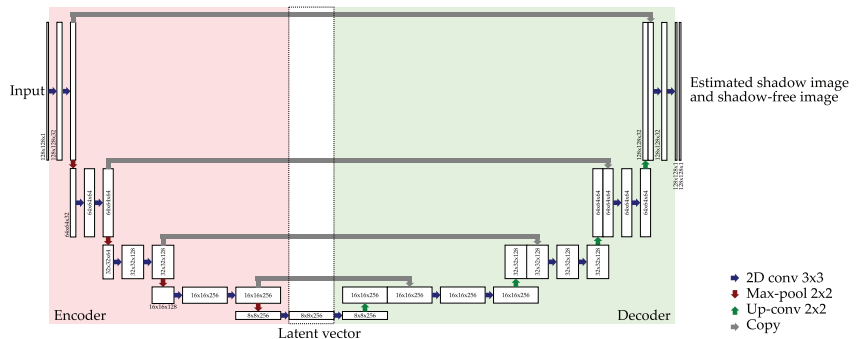
**Input:** Parameters for annular sectors (center coordinate  $p \in \mathbb{R}^2$ , range of direction  $d_{\min}, d_{\max}$ , range of angle  $\theta_{\min}, \theta_{\max}$ , range of outer radius  $R_{\min}, R_{\max}$ , and minimum inner radius  $r_{\min}$ ), blurring parameters  $\delta_\theta, k, \sigma$ , and range of shadow intensity  $v_{\min}, v_{\max}$ .

**Output:** Image of a synthetic shadow  $s \in [0, 1]^{H \times W}$ .

- 1:  $d \leftarrow U(d_{\min}, d_{\max})$ .
- 2:  $\theta \leftarrow U(\theta_{\min}, \theta_{\max})$ .
- 3:  $R \leftarrow U(R_{\min}, R_{\max})$ .
- 4:  $r \leftarrow U(r_{\min}, R)$ .
- 5:  $v \leftarrow U(v_{\min}, v_{\max})$ .
- 6:  $s \leftarrow 0_{H,W}$  (a zero matrix shaped  $H \times W$ ).
- 7: **for**  $i = -(k-1)/2, \dots, (k-1)/2$  **do**
- 8:   Let  $s_k \in [0, 1]^{H \times W}$  be a image that filled with 1 inside an annular sector which center is  $p$ , outer radius is  $r$ , angle is  $d + (i\delta_\theta)$ , and direction is  $\theta$ , and 0 otherwise.
- 9:    $s \leftarrow s + s_k$ .
- 10: **end for**
- 11:  $s \leftarrow v(s / \max(s))$ .
- 12:  $s \leftarrow 1 - s$ .
- 13: Apply Gaussian blur with variance  $\sigma^2$  to  $s$ .

### Appendix B. Details of DNNs

The encoder and the decoder of the proposed method had almost the same structure as U-Net [30]. Its network architecture is shown in Figure A1. To stabilize the training, we used leaky ReLU [38] as an activation function for convolution layers.



**Figure A1.** Detailed architecture of the encoder and the decoder for the proposed method.

### Appendix C. Selected Hyperparameters

The hyperparameters for the experiments are shown in Table A1. These are selected by the grid search as described in Section 4.1.



**Table A1.** Hyperparameters selected in the experiments by the grid search.

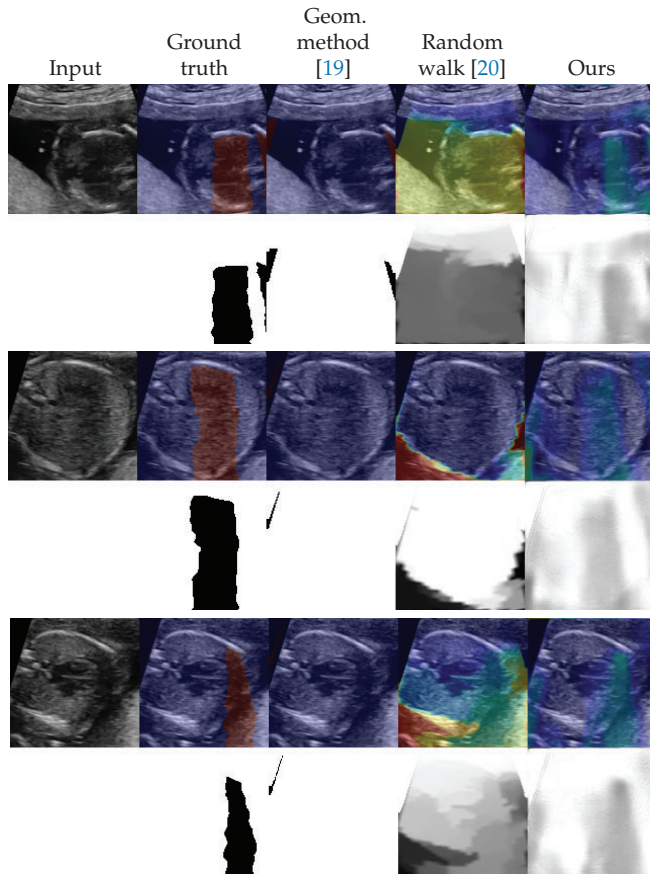
Hyperparameter	Number of Labeled Images				
	0	42 (5 Videos)	90 (10 Videos)	177 (20 Videos)	259 (30 Videos)
Threshold for random walk [20]	0.996	-	-	-	-
Threshold for the proposed method	0.865	0.870	0.890	0.894	0.885
$\lambda_{\text{synth}} = \lambda_{\text{label}}$	0.996	1	1	10	10
$\lambda_{\text{synthseg}}$	0.996	$10^{-3}$	0	0	0
$v_{\text{min}}$	0.996	0.1	0.5	0.1	0.5

**Appendix D. Additional Results**

Figures A2 and A3 shows the additional examples of the shadow detection results. Figures A2 and A3 correspond to Figures 2 and 3, respectively. Table A2 shows the shadow detection results in the the DICE scores for the validation dataset. Its trend is similar to that for the testing dataset that is shown in Table 1.

Table A3 shows the result of the shadow intensity estimation for the validation dataset. The results are similar to these for the testing dataset that are shown in Table 2.

Figure A4 shows the additional examples of the shadow removal results. It corresponds to Figure 4.



**Figure A2.** Additional examples of shadow detection results for the methods that do not use labels. The lower side of each example shows detection results, and the upper side shows them overlaid to the input image. For overlaid images, blue corresponds to low intensities and red corresponds to high intensities.

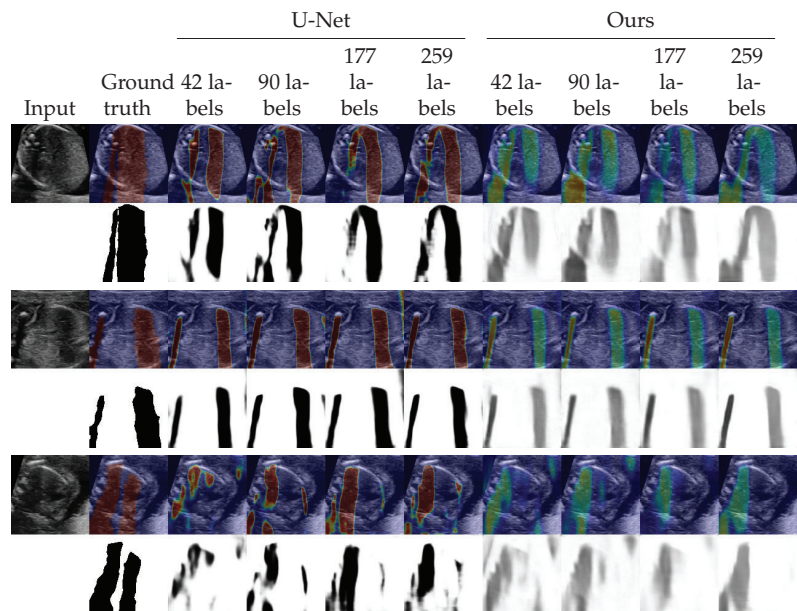


**Table A2.** Results of shadow detection for the validation dataset evaluated in the DICE score. The scores are calculated for each validation image, and means over them are shown. The numbers in parentheses are the standard deviations.

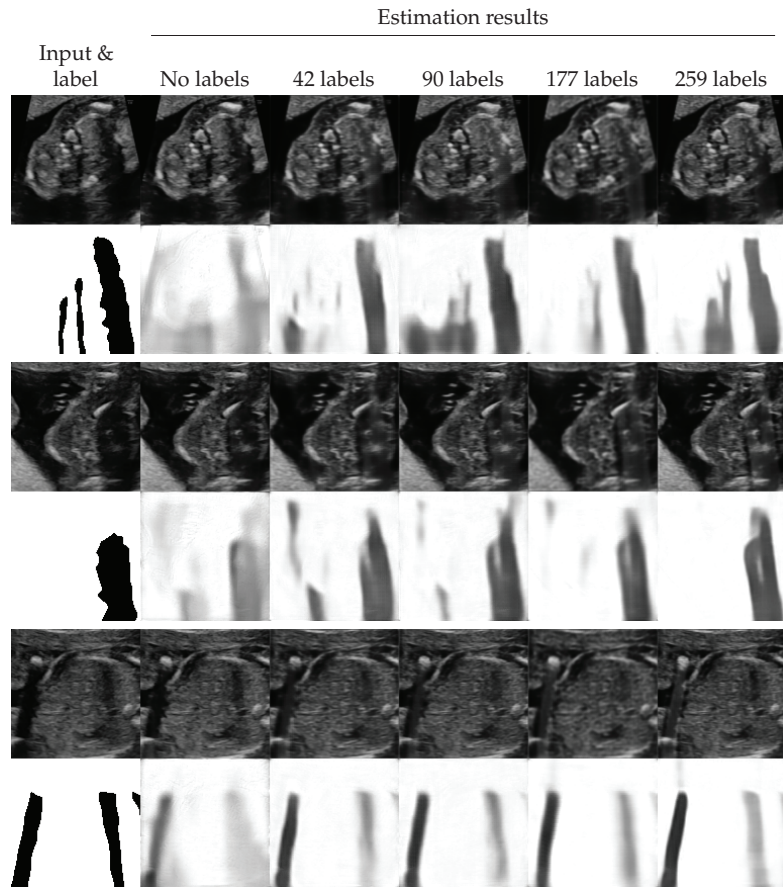
Method	Number of Labeled Images				
	0	42 (5 Videos)	90 (10 Videos)	177 (20 Videos)	259 (30 Videos)
Geometric method [19]	0.201 (±0.213)	-	-	-	-
Random walk [20]	0.349 (±0.151)	-	-	-	-
U-Net [30]	-	0.539 (±0.220)	0.575 (±0.215)	0.636 (±0.176)	0.657 (±0.181)
Ours	0.491 (±0.180)	0.615 (±0.176)	0.640 (±0.201)	0.676 (±0.157)	0.692 (±0.172)

**Table A3.** Evaluation of the estimation of shadow intensities for the validation dataset. Scores are the correlation coefficient calculated by Equation (15). The coefficients are calculated for each validation image, and means over them are shown. The numbers in parentheses are the standard deviations.

Method	Number of Labeled Images				
	0	42 (5 Videos)	90 (10 Videos)	177 (20 Videos)	259 (30 Videos)
Geometric method [19]	0.194 (±0.131)	-	-	-	-
Random walk [20]	-0.054 (±0.295)	-	-	-	-
U-Net [30]	-	0.282 (±0.170)	0.267 (±0.158)	0.262 (±0.168)	0.210 (±0.187)
Ours	0.353 (±0.190)	0.426 (±0.131)	0.420 (±0.140)	0.338 (±0.153)	0.310 (±0.168)



**Figure A3.** Additional examples of shadow detection results for the methods that use labels. The lower side of each example shows detection results, and the upper side shows them overlaid to the input image. For overlaid images, blue corresponds to low intensities and red corresponds to high intensities.



**Figure A4.** Additional examples of shadow removal results of the proposed method. The lower side of each example shows the labels and the detection results. The upper side shows the input images and the estimated shadow-free images.

## References

1. Szabo, T.L. *Diagnostic Ultrasound Imaging: Inside Out*; Academic Press: Cambridge, MA, USA, 2004.
2. Moran, C.M.; Pye, S.D.; Ellis, W.; Janeczko, A.; Morris, K.D.; McNeilly, A.S.; Fraser, H.M. A Comparison of the Imaging Performance of High Resolution Ultrasound Scanners for Preclinical Imaging. *Ultrasound Med. Biol.* **2011**, *37*, 493–501. [[CrossRef](#)]
3. Sassaroli, E.; Crane, C.; Scorza, A.; Kim, D.S.; Park, M.A. Image Quality Evaluation of Ultrasound Imaging Systems: Advanced B-Modes. *J. Appl. Clin. Med. Phys.* **2019**, *20*, 115–124. [[CrossRef](#)]
4. Entekin, R.R.; Porter, B.A.; Sillesen, H.H.; Wong, A.D.; Cooperberg, P.L.; Fix, C.H. Real-Time Spatial Compound Imaging: Application to Breast, Vascular, and Musculoskeletal Ultrasound. *Semin. Ultrasound CT MRI* **2001**, *22*, 50–64. [[CrossRef](#)]
5. Desser, T.S.; Jeffrey, R.B., Jr.; Lane, M.J.; Ralls, P.W. Tissue Harmonic Imaging: Utility in Abdominal and Pelvic Sonography. *J. Clin. Ultrasound* **1999**, *27*, 135–142. [[CrossRef](#)]
6. Ortiz, S.H.C.; Chiu, T.; Fox, M.D. Ultrasound Image Enhancement: A Review. *Biomed. Signal Process. Control* **2012**, *7*, 419–428. [[CrossRef](#)]
7. Perdios, D.; Vonlanthen, M.; Besson, A.; Martinez, F.; Arditi, M.; Thiran, J. Deep Convolutional Neural Network for Ultrasound Image Enhancement. In Proceedings of the 2018 IEEE International Ultrasonics Symposium, Kobe, Japan, 22–25 October 2018.
8. Feldman, M.K.; Katyal, S.; Blackwood, M.S. US Artifacts. *RadioGraphics* **2009**, *29*, 1179–1189. [[CrossRef](#)] [[PubMed](#)]
9. Ziskin, M.C.; Thickman, D.I.; Goldenberg, N.J.; Lapayowker, M.S.; Becker, J.M. The Comet Tail Artifact. *J. Ultrasound Med.* **1982**, *1*, 1–7. [[CrossRef](#)]
10. Noble, J.A.; Boukerroui, D. Ultrasound Image Segmentation: A Survey. *IEEE Trans. Med. Imaging* **2006**, *25*, 987–1010. [[CrossRef](#)]

11. Brattain, L.J.; Telfer, B.A.; Dhyani, M.; Grajo, J.R.; Samir, A.E. Machine Learning for Medical Ultrasound: Status, Methods, and Future Opportunities. *Abdom. Radiol.* **2018**, *43*, 786–799. [[CrossRef](#)]
12. Liu, S.; Wang, Y.; Yang, X.; Lei, B.; Liu, L.; Li, S.X.; Ni, D.; Wang, T. Deep Learning in Medical Ultrasound Analysis: A Review. *Engineering* **2019**, *5*, 261–275. [[CrossRef](#)]
13. Drukker, L.; Noble, J.A.; Papageorgiou, A.T. Introduction to Artificial Intelligence in Ultrasound Imaging in Obstetrics and Gynecology. *Ultrasound Obstet. Gynecol.* **2020**, *56*, 498–505. [[CrossRef](#)] [[PubMed](#)]
14. Dozen, A.; Komatsu, M.; Sakai, A.; Komatsu, R.; Shozu, K.; Machino, H.; Yasutomi, S.; Arakaki, T.; Asada, K.; Kaneko, S.; et al. Image Segmentation of the Ventricular Septum in Fetal Cardiac Ultrasound Videos Based on Deep Learning Using Time-Series Information. *Biomolecules* **2020**, *10*, 1526. [[CrossRef](#)] [[PubMed](#)]
15. Shozu, K.; Komatsu, M.; Sakai, A.; Komatsu, R.; Dozen, A.; Machino, H.; Yasutomi, S.; Arakaki, T.; Asada, K.; Kaneko, S.; et al. Model-Agnostic Method for Thoracic Wall Segmentation in Fetal Ultrasound Videos. *Biomolecules* **2020**, *10*, 1691. [[CrossRef](#)] [[PubMed](#)]
16. LeCun, Y.; Bengio, Y.; Hinton, G. Deep Learning. *Nature* **2015**, *521*, 436–444. [[CrossRef](#)] [[PubMed](#)]
17. Komatsu, M.; Sakai, A.; Komatsu, R.; Matsuoka, R.; Yasutomi, S.; Shozu, K.; Dozen, A.; Machino, H.; Hidaka, H.; Arakaki, T.; et al. Detection of Cardiac Structural Abnormalities in Fetal Ultrasound Videos Using Deep Learning. *Appl. Sci.* **2021**, *11*, 371. [[CrossRef](#)]
18. Vincent, P.; Larochelle, H.; Lajoie, I.; Bengio, Y.; Manzagol, P.A. Stacked Denoising Autoencoders: Learning Useful Representations in a Deep Network with a Local Denoising Criterion. *J. Mach. Learn. Res.* **2010**, *11*, 3371–3408.
19. Hellier, P.; Coupé, P.; Morandi, X.; Collins, D.L. An Automatic Geometrical and Statistical Method to Detect Acoustic Shadows in Intraoperative Ultrasound Brain Images. *Med. Image Anal.* **2010**, *14*, 195–204. [[CrossRef](#)]
20. Karamalis, A.; Wein, W.; Klein, T.; Navab, N. Ultrasound Confidence Maps Using Random Walks. *Med. Image Anal.* **2012**, *16*, 1101–1112. [[CrossRef](#)]
21. Hacihaliloglu, I. Enhancement of bone shadow region using local phase-based ultrasound transmission maps. *Int. J. Comput. Assist. Radiol. Surg.* **2017**, *12*, 951–960. [[CrossRef](#)]
22. Meng, Q.; Baumgartner, C.; Sinclair, M.; Housden, J.; Rajchl, M.; Gomez, A.; Hou, B.; Toussaint, N.; Zimmer, V.; Tan, J.; et al. Automatic Shadow Detection in 2D Ultrasound Images. In *Data Driven Treatment Response Assessment and Preterm, Perinatal, and Paediatric Image Analysis, Granada, Spain, 16 September 2018*; Springer: Berlin/Heidelberg, Germany, 2018; pp. 66–75.
23. Meng, Q.; Sinclair, M.; Zimmer, V.; Hou, B.; Rajchl, M.; Toussaint, N.; Oktay, O.; Schlemper, J.; Gomez, A.; Housden, J.; et al. Weakly Supervised Estimation of Shadow Confidence Maps in Fetal Ultrasound Imaging. *IEEE Trans. Med. Imaging* **2019**, *38*, 2755–2767. [[CrossRef](#)]
24. Hu, R.; Singla, R.; Yan, R.; Mayer, C.; Rohling, R.N. Automated Placenta Segmentation with a Convolutional Neural Network Weighted by Acoustic Shadow Detection. In Proceedings of the 2019 41st Annual International Conference of the IEEE Engineering in Medicine and Biology Society, Berlin, Germany, 23–27 July 2019; pp. 6718–6723.
25. Goodfellow, I.; Bengio, Y.; Courville, A.; Bengio, Y. *Deep Learning*; MIT Press: Cambridge, MA, USA, 2016.
26. Kingma, D.P.; Welling, M. Auto-Encoding Variational Bayes. *arXiv* **2014**, arXiv:1312.6114.
27. Makhzani, A.; Shlens, J.; Jaitly, N.; Goodfellow, I.; Frey, B. Adversarial Autoencoders. *arXiv* **2016**, arXiv:1511.05644.
28. Rasmus, A.; Berglund, M.; Honkala, M.; Valpola, H.; Raiko, T. Semi-Supervised Learning with Ladder Networks. In *Advances in Neural Information Processing Systems, Montréal, Canada, 7–10 December 2015*; Curran Associates, Inc.: Red Hook, NY, USA, 2015; Volume 28, pp. 3546–3554.
29. Garcia-Garcia, A.; Orts-Escolano, S.; Oprea, S.; Villena-Martinez, V.; Martinez-Gonzalez, P.; Garcia-Rodriguez, J. A Survey on Deep Learning Techniques for Image and Video Semantic Segmentation. *Appl. Soft Comput.* **2018**, *70*, 41–65. [[CrossRef](#)]
30. Ronneberger, O.; Fischer, P.; Brox, T. U-Net: Convolutional Networks for Biomedical Image Segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention, Proceedings of the MICCAI 2015: Medical Image Computing and Computer-Assisted Intervention, Munich, Germany, 5–9 October 2015*; Springer: Berlin/Heidelberg, Germany, 2015; pp. 234–241.
31. Chen, C.; Qin, C.; Qiu, H.; Tarroni, G.; Duan, J.; Bai, W.; Rueckert, D. Deep Learning for Cardiac Image Segmentation: A Review. *Front. Cardiovasc. Med.* **2020**, *7*, 25. [[CrossRef](#)] [[PubMed](#)]
32. Wang, L.W.; Siu, W.C.; Liu, Z.S.; Li, C.T.; Lun, D.P.K. Deep Relighting Networks for Image Light Source Manipulation. In Proceedings of the 2020 European Conference on Computer Vision, Glasgow, UK, 23–28 August 2020.
33. Yasutomi, S.; Arakaki, T.; Hamamoto, R. Shadow Detection for Ultrasound Images Using Unlabeled Data and Synthetic Shadows. *arXiv* **2019**, arXiv:1908.01439.
34. Kingma, D.P.; Ba, J. Adam: A Method for Stochastic Optimization. In Proceedings of the 3rd International Conference on Learning Representations, San Diego, CA, USA, 7–9 May 2015.
35. Carass, A.; Roy, S.; Gherman, A.; Reinhold, J.C.; Jesson, A.; Arbel, T.; Maier, O.; Handels, H.; Ghafoorian, M.; Platel, B.; et al. Evaluating White Matter Lesion Segmentations with Refined Sørensen-Dice Analysis. *Sci. Rep.* **2020**, *10*, 8242. [[CrossRef](#)] [[PubMed](#)]
36. Lane, D.; Scott, D.; Hebl, M.; Guerra, R.; Osherson, D.; Zimmer, H. *Introduction to Statistics*; Rice University: Houston, TX, USA, 2003. Available online: <https://open.umn.edu/opentextbooks/textbooks/459> (accessed on 10 December 2020).

37. Isola, P.; Zhu, J.Y.; Zhou, T.; Efros, A.A. Image-To-Image Translation With Conditional Adversarial Networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017.
38. Maas, A.L.; Hannun, A.Y.; Ng, A.Y. Rectifier Nonlinearities Improve Neural Network Acoustic Models. In *ICML Workshop on Deep Learning for Audio, Speech and Language Processing, Atlanta, USA, 16–21 June 2013*; PMLR: Cambridge, MA, USA, 2013.

Article

# Investigating Health-Related Features and Their Impact on the Prediction of Diabetes Using Machine Learning

Hafiz Farooq Ahmad <sup>1</sup>, Hamid Mukhtar <sup>2,\*</sup>, Hesham Alaqail <sup>1</sup>, Mohamed Seliaman <sup>3</sup> and Abdulaziz Alhumam <sup>1</sup>

<sup>1</sup> Computer Science Department, College of Computer Sciences and Information Technology (CCSIT), King Faisal University, P. O. Box 400, Al-Ahsa 31982, Saudi Arabia; hfahmad@kfu.edu.sa (H.F.A.); hisham@mail.net.sa (H.A.); aahumam@kfu.edu.sa (A.A.)

<sup>2</sup> Computer Science Department, College of Computers and Information Technology (CCIT), Taif University, P. O. Box 11099, Taif 21944, Saudi Arabia

<sup>3</sup> Information Systems Department, College of Computer Sciences and Information Technology (CCSIT), King Faisal University, P.O. Box 400, Al-Ahsa 31982, Saudi Arabia; seliamanme@kfu.edu.sa

\* Correspondence: h.mukhtar@tu.edu.sa

**Abstract:** Diabetes Mellitus (DM) is one of the most common chronic diseases leading to severe health complications that may cause death. The disease influences individuals, community, and the government due to the continuous monitoring, lifelong commitment, and the cost of treatment. The World Health Organization (WHO) considers Saudi Arabia as one of the top 10 countries in diabetes prevalence across the world. Since most of its medical services are provided by the government, the cost of the treatment in terms of hospitals and clinical visits and lab tests represents a real burden due to the large scale of the disease. The ability to predict the diabetic status of a patient with only a handful of features can allow cost-effective, rapid, and widely-available screening of diabetes, thereby lessening the health and economic burden caused by diabetes alone. The goal of this paper is to investigate the prediction of diabetic patients and compare the role of HbA1c and FPG as input features. By using five different machine learning classifiers, and using feature elimination through feature permutation and hierarchical clustering, we established good performance for accuracy, precision, recall, and F1-score of the models on the dataset implying that our data or features are not bound to specific models. In addition, the consistent performance across all the evaluation metrics indicate that there was no trade-off or penalty among the evaluation metrics. Further analysis was performed on the data to identify the risk factors and their indirect impact on diabetes classification. Our analysis presented great agreement with the risk factors of diabetes and prediabetes stated by the American Diabetes Association (ADA) and other health institutions worldwide. We conclude that by performing analysis of the disease using selected features, important factors specific to the Saudi population can be identified, whose management can result in controlling the disease. We also provide some recommendations learned from this research.

**Keywords:** machine learning; prediction; feature importance; feature elimination; hierarchical clustering



**Citation:** Ahmad, H.F.; Mukhtar, H.; Alaqail, H.; Seliaman, M.; Alhumam, A. Investigating Health-Related Features and Their Impact on the Prediction of Diabetes Using Machine Learning. *Appl. Sci.* **2021**, *11*, 1173. <https://doi.org/10.3390/app11031173>

Academic Editor: Jordi Solé-Casals  
Received: 18 December 2020  
Accepted: 23 January 2021  
Published: 27 January 2021

**Publisher's Note:** MDPI stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.



**Copyright:** © 2021 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

## 1. Introduction

Diabetes mellitus (DM) is one of the most common chronic diseases worldwide. In 2019, the International Diabetes Federation (IDF) announced that the number of adults that are diagnosed with diabetes is approximately 463 million of the world's population [1]. In addition, IDF considers the Middle East as one of the highest regions in diabetes prevalence, and the World Health Organization (WHO) places Saudi Arabia as the highest among Middle Eastern countries [2] and fifth in the top 10 countries known for a high diabetes incidence rate in the world. It is expected that Saudi Arabia is heading to a higher position by 2035 [3]. The cost of medical treatment is also affected by the rapid growth of the number of individuals with diabetes, representing a large burden on government

health expenses. According to recent estimates, the cost of diabetes incurred by the Saudi government is at 17 billion Riyals and if those with glucose intolerance (pre-diabetes) progressed at the current observed rate, the total cost would be 43 billion Riyals [4] in the coming years. Besides, Saudi Arabia is known for its rapid growth in population and has encountered soaring economic development in the recent four decades, leading to lifestyle changes due to urbanization.

These changes have led to an increasing rate of chronic diseases. Many studies conducted to address the rapid growth of Diabetes Mellitus have either the objective to quantify the status of diabetics in the country [3,5], identifying the most frequently performed self-care behaviors [6], identifying factors related to diabetes control [7], or apply mathematical [8] or machine learning models for diabetes prediction [9]. All these efforts are related to the increasing demand to enhance healthcare quality and control the elevated growth rate of diabetes in the kingdom.

It is essential for federal or local governments to perform national or local screening and educate people through awareness programs. There is a need to invest in novel ways to prevent and help in the early detection of such an expensive disease [10]. Early prevention can limit the complications and their impact on the person's quality of life, resulting in a reduced cost with a positive impact on the community and the health system [11]. An upfront cost in the form of early investments by the governments can result in long-term benefits to the overall society.

We believe that the existing efforts may have their own benefits and usefulness in tackling the diabetes issues in Saudi Arabia however, there is a need to devise mechanisms for efficient, cost-effective, and easily-available solutions for diabetes identification in the general population. Given the constant rise in the diabetic population in the country, it is imperative not only to identify diabetics from non-diabetic persons but at the same time, the factors associated with diabetes should also be delineated. By knowing and overcoming these factors, people can act to control them in time. Clinics and hospitals can also identify patients-at-risk by evaluating these factors.

Considering the above-mentioned context and the need for time, we are motivated to develop a solution for predicting diabetics from non-diabetic patients from the electronic health records obtained from local Saudi hospitals. Therefore, the goal of this research paper is to develop predictive classifiers and models to investigate real diabetic patients' data gathered from different Saudi hospitals and regions, utilizing different metrics. Although the obtained records have very few health-related attributes including lab test results such as cholesterol tests (HDL and LDL), and the diabetes-specific tests (FPG and HbA1c), our objective is to identify those factors that can be controlled by the patients-at-risk or non-diabetics as precautions for avoiding diabetes. Previous work in this direction has used a much larger number of variables in different contexts. Thus, the current work presents a novel perspective on diabetes prediction. The insights obtained from this work in the prediction of Diabetes Mellitus (DM) and its associated risk factors can be useful at different levels: To support and strengthen the existing findings of DM medical research, particularly, in the context of Saudi Arabia, to assist the community in understanding the causes and prevention of diabetes, and to help the government to allocate efforts in the right direction to minimize the effect of the growing number of diabetes patients.

With these objectives in mind, we developed a model that used five machine learning classifiers to evaluate two datasets; each one containing some basic attributes about the patients (age, gender, weight, height, presence of hypertension, and the level of physical activity), the results of cholesterol laboratory tests (HDL and LDL) and one of these two tests: HbA1c in one dataset and FPG in another dataset. With these limited sets of features, we evaluated each dataset and identified the importance of various features and their role in diabetes prediction. To improve the performance of our classifiers, we also applied feature elimination using feature permutation and hierarchical clustering techniques. Finally, using the rank-based correlation method, we also identified and analyzed correlation among various features and their impact on diabetes risk and prediction.



The remainder of this paper is structured as follows. In Section 2, we briefly explain DM followed by the related work done in diabetes classification. In Section 3, we explain our research methodology from data collection and preprocessing to the process of feature engineering and dataset creation. Section 4 explains the development of machine learning models for diabetes prediction and classification with a focus on improving the model performance through feature elimination. The results are then discussed in Section 5. Section 6 discusses the outcome and benefit of our research and Section 7 concludes this article.

## 2. Background and Related Work

### 2.1. Diabetes Mellitus (DM) and Risk Factors

DM is a set of endocrine disorders resulting in high levels of blood glucose in the human body due to a deficiency in insulin excretion or insulin action and sometimes both. It causes direct and indirect complications responsible for significant illness and death [5]. There are different types of diabetes, but the most common ones are Type 1 Diabetes (T1D) and Type 2 Diabetes (T2D). Type 2 (T2D) is the most common form of diabetes as around 90% of diabetic patients are T2D. The remaining 10% is classified as T1D or gestational diabetes, which may occur during pregnancies.

The blood test for the measurement of Hemoglobin A1c (HbA1c) level is clinically significant in prediabetes and diabetes diagnosis [12]. Similarly, the glucose in plasma of fasting subjects is accepted as a diagnostic criterion for diabetes [13]. Moreover, according to the American Diabetes Association (ADA) there is more harmony between blood tests such as FPG and HbA1c when compared to two types of blood tests in the separation of HbA1c [14]. Most of the existing work achieve good results for diabetes prediction only when they include these tests in their input to the machine learning model along with a myriad of other features [15–20]. However, as the number of features is reduced, such predictions cannot be made with greater accuracy, and in absence of these tests, it becomes impossible to identify the diabetic status of a patient with high certainty.

From a medical point of view, it is possible to avoid DM at an early stage or at least control its complications [21]. For example, individuals with a certain range of FPG and HbA1c, are considered as prediabetic patients [12]. Their early diagnosis can help in preventing their transition to becoming diabetic or in their recovery into the non-diabetic stage. However, identification of factors that can lead a person to transition to the status of diabetes in a population is a challenge, albeit some studies have identified factors such as hypertension or body size among some of the associated risk factors with diabetes [22]. Other studies have identified certain conditions that can only be determined through various blood or imaging tests [23–26]. The unavailability of such tests at most health facilities and the associated costs may prevent people from diagnosing with diabetes and, thus, a large part of the population remains undiagnosed until it is very late in the treatment process [27].

Despite these difficulties associated with the diagnosis of diabetes, prediction of diabetes using machine learning techniques has gained significant attention from the medical and informatics research community. Below, we identify some of the recent efforts in this direction.

### 2.2. Related Work

There are different Machine Learning (ML)-based methods for diabetes prediction as well as methods that use feature selection. We will review them next.

### 2.3. ML-Based Methods

Othmane et al. [28] applied and evaluated four ML algorithms (decision tree, K-nearest neighbors, artificial neural network, and deep neural network) to predict patients with or without type 2 diabetes mellitus. These techniques were trained and tested on two diabetes databases: One obtained from Frankfurt hospital (Germany), and the other one,



the openly available, well-known Pima Indian dataset (<https://www.kaggle.com/uciml/Pima-indians-diabetes-database>). These datasets contained the same features composed of risk factors and some clinical data such as the number of pregnancies, glucose levels, blood pressure, skin thickness, insulin, BMI (Body Mass Index), age, and diabetes pedigree function. The results compared using different similarity metrics give a classification accuracy of more than 90% and up to 100% in some cases. Similarly, many other approaches trained their models on similar features. For example, in [15–20]) the authors used the Pima Indian diabetes dataset by modifying the preprocessing steps, applying different algorithms and adjusting their hyperparameters to generate improved results. The limiting factor of these approaches is the inclusion of some features like skin thickness, insulin, and diabetes pedigree function, which are generally not available or recorded. Moreover, factors like skin thickness may result in the classification based on ethnic function, thus, preventing a wide-range applicability of the approach.

Lai et al. [29] built a predictive model to better identify Canadian patients at risk of having Diabetes Mellitus based on patient demographic data and the laboratory results. Their data included the patient features age, sex, fasting blood glucose, body mass index, high-density lipoprotein, triglycerides, blood pressure, and low-density lipoprotein. They built predictive models using Logistic Regression and Gradient Boosting Machine (GBM) techniques achieving good sensitivity results. But the authors did not mention their performance in accuracy or specificity, which usually has better sensitivity as a trade-off. Thus, their performance cannot be generalized. Like this, many research works have compared the performance of several machine learning using the selected metrics, while a different metric may give a poor performance on the same model. Many other approaches for diabetes classification concluded that a certain type of algorithms can give better results for prediction without considering the issue of the generality of their models [30,31].

A number of other studies have used the National Health and Nutrition Examination Survey (NHANES) (<https://wwwn.cdc.gov/nchs/nhanes/>) from the US Center for Disease Control (CDC) for the prediction of diabetes or other diseases. The NHANES data was initiated in 1999 and is growing every year in the number of records as well as the variables it considers in its surveys. These studies, while utilizing the main NHANES dataset, use some subset of variables for disease prediction or classification tasks. For example, Yu et al. [32] identified 14 important variables—age, weight, height, BMI, gender, race and ethnicity, family history, waist circumference, hypertension, physical activity, smoking, alcohol use, education, and household income—for training their machine learning models. Using two different classification schemes, they achieved 83.5% and 73.2% results for the area under the Receiver Operating Characteristic (ROC) curve. Semerdjian and Frank [33] added two more variables—cholesterol and leg length—in their analysis. By applying an ensemble model using the output of five classification algorithms they were able to predict the onset of diabetes with an AUC (Area Under Curve) of 83.4%. In both these studies, the number of variables (14 and 16) was significantly higher than would normally be available in most EHRs. Hospitals supporting the record of these variables may also not have the values for all these variables for maximum patients. This limits the generality or wide applicability of the approaches.

The study by Dinh et al. [34] used the NHANES dataset and various machine learning algorithms to predict variables that are a major cause for the development of diabetes and cardiovascular diseases. They also considered the prediction of prediabetes and undiagnosed diabetes. Logistic regression, support vector machines, random forest, and gradient boosting algorithms were used to classify the data and predict the outcome for the diseases. The authors used ensemble models by combining the performance of the weaker models to improve accuracy. In diabetes classification, they used 123 variables and achieved good prediction performance. A distinguishing aspect of their work was that the dataset was further categorized into laboratory dataset (containing laboratory results) versus non-laboratory (survey data only) dataset. Laboratory results were any feature variables within the dataset that were obtained via blood or urine tests. The purpose of the non-laboratory dataset was

to enable a performance analysis of machine learning models in cases where laboratory results were unavailable for patients, supporting the detection of at-risk patients based only on a survey questionnaire. According to their results, waist size, age, self-reported weight, leg length, and sodium intake were five major predictors for diabetes patients. The study found that machine learning models based on survey questionnaires can give automated identification mechanisms for patients at risk of diabetes. In non-laboratory data, the most important features included waist size, age, weight (self-reported and actual), leg-length, blood pressure, BMI, household income, etc. [34]. The exact number of variables used in non-laboratory data is not reported by the authors, and, thus, it cannot be concluded if their approach can be useful in general situations.

#### 2.4. Feature-Based Methods

Feature selection has been used previously for improving the classification performance in different medical situations. Particularly, Matín-González et al. have proved that by performing feature selection, the results of the classifier can be improved for the prediction of success or failure in Noninvasive Mechanical Ventilation (NIMV) in Intensive Care Units (ICU) [35]. Similarly, Akay [36] used F-score feature selection-based SVM model, and Chen et al. [37] used SVM with rough-set based feature selection for the improved diagnosis of breast cancer. Liaqat et al. [38] performed a premier study on developing deep learning-based classifiers for atrial fibrillation. They built six models based on feature-based approaches and DL approaches. However, their features are manually extracted while the DL methods are trained on raw data without any feature engineering, as they perform implicit feature selection. It is unclear how they performed the manual feature extraction, but manual feature extraction is not a preferred approach if this can be done automatically, as explained later in our case.

Amer et al. applied a feature engineering approach to gain clinical insight and, thus, improve the ICU mortality prediction in field conditions [39]. The authors used only linear hard margin SVM as it maximizes the separation between different classes. Feature selection was performed using statistical and dynamic feature extraction with an evaluation performed after each step. Any misclassifications after these two stages were investigated manually. A final phase of feature fine-tuning consists of seven steps and utilizes the vital signs as opposed to the selection of dimensions in the previous stages. Results were then obtained by evaluating the various combinations of feature selection performed in different stages. The interesting aspect of their approach is the combination of different features at different stages and improving the results step-by-step. A conclusion of the work was that different profiles of patients required a different set of features for efficiently predicting the mortality of patients.

Tomar and Agarwal used the hybrid feature selection technique [40] on three different datasets of diabetes, hepatitis, and breast cancer. Their model adopted Weighted Least Squares Twin SVM (WLSTSVM) as a classification approach, sequential forward selection as a search strategy, and correlation feature selection to evaluate the importance of each feature. In contrast, we applied permutation importance for feature selection, which is known to be a faster technique without the need for a selection strategy. Once the features were found, we could use any of the ML models to classify and predict the data.

Specific for the case of diabetes, Balakrishnan et al. used SVM ranking with backward search for feature selection in T2D databases [41], where they proposed a specific feature selection approach for finding an optimum feature subset that enhanced the classification accuracy of Naive Bayes classifier. Ephzibah [42] constructed a combined model using genetic algorithms and fuzzy logic for feature subset selection. However, genetic algorithms have their own associated costs, and the proposed approach did not justify the cost compared to the achieved accuracy. On the same lines, Aslam et al. [43] used genetic programming with Pima Indian diabetes dataset by generating subsets of original features by adding features one by one in each subset using the sequential forward selection method. The approach is not only costly but their results using 10-fold cross validation with KNN

(K-Nearest Neighbor) and a specific configuration of genetic programming yielded an accuracy of about 80.5%, which is not up to par with other contemporary approaches.

Rodríguez-Rodríguez et al. [44] applied feature selection on T1D Mellitus (T1D) patients using variables like sleep, schedule, meal, exercise, insulin, and heart-rate. Using time-series data of these features and the Sequential Input Selection Algorithm (SISAL), they ranked features according to their importance with respect to their relevance in the blood glucose level prediction.

One approach for feature selection consists of clustering that has been mostly used for dimensionality reduction in text classification. But hierarchical agglomerative clustering that organizes features into progressively larger groups [45] have been used in structural classification. Ienco and Meo [46] used hierarchical clustering for improving the accuracy of classification on 40 datasets from the UCI Irvine[47]. Their experimental results show that the hierarchical clustering method of feature selection outperforms the ranking methods in terms of accuracy. On the diabetes data, they achieved an accuracy of 77.47% using the Naïve Bayes' classifier and 75.26% using the J48 based classifier. There are two limitations of their work. First, the accuracy is not as good as reported by other approaches on the same dataset. Second, the performance reported is only the accuracy of classification, but it is worse in other evaluations. Compared to their approach, we report a much higher performance in accuracy, precision, and recall scores.

Considering the above analysis, we conclude that most of the existing approaches (1) use features which are not generalizable, (2) use a large number of features that cannot be obtained in many real-world scenarios, (3) develop specific models that may not be generalizable, and (4) report only a specific metric for evaluation while ignoring other metrics that may have worse performance as an issue of trade-off between the various evaluation metrics. We approach the problem of diabetes prediction while considering these limitations.

In our approach, we use a minimum number of features, reducing them further by feature elimination. We apply five different classification models to avoid model-specific performance. We report that all the models performed equally well on the metrics of accuracy, precision, recall, and F1-score, implying that our data or features are not bound to specific models. Finally, our analysis also includes the identification of those factors that can have an indirect impact on the complications of diabetes.

### 3. Materials and Methods

We begin with data collection, preprocessing, feature engineering, and label assignment to explain how we obtain two different datasets from the same subset of features.

#### 3.1. Data Collection

The anonymized Electronic Health Records have been acquired from five different Saudi hospitals across three regions: The Central region, the Western region, and the Eastern region. It contains data of around 3000 patients collected over two years from 2016 until 2018 through different departments such as outpatient, inpatient, and emergency. The obtained dataset consists of 16 features of numerical, binominal, polynomial, and date type. The initial features along with a brief description of each are listed in Table 1.

**Table 1.** The set of features selected in our dataset for classification of diabetic and prediabetic patients.

No.	Feature Name	Feature Type	Feature Description
1	Date of birth	Date	Values in date format
2	Gender	Binominal	F: Female, M: Male
3	Height	Numerical	Values in Centimeter (cm)
4	Weight	Numerical	Values in Kilograms (kg)
5	Hypertension (HTN)	Binominal	Values as Yes, No
6	Fasting Plasma Glucose (FPG)	Numerical	Lab test results measured in mmol/L
7	Hemoglobin A1c (HbA1c)	Numerical	Lab test results measured in percentage (%)
8	High-density lipoprotein (HDL)	Numerical	Lab test results in mmol/L
9	High-density lipoprotein (LDL)	Numerical	Lab test results in mmol/L
10	Physical Activity Level	Categorical	Values in L: Low, M: Medium, H: High
11	Diagnosis start date	Date	Values in date format
12	Primary diagnosis	Categorical	Values available in ICD10 code format.
13	Secondary diagnosis	Categorical	Values available in ICD10 code format.
14	Primary diagnosis full name	Categorical	Values indicate diagnosis full name
15	Secondary diagnosis full name	Categorical	Values indicate diagnosis full name
16	Region	Categorical	Values indicate the region of the patient whether in central, western or eastern region.

### 3.2. Data Preprocessing

In the data preprocessing phase, data is prepared to be suitable for cleansing and classification. The data is cleansed using normalization and transformation of some columns (features) for example, the birth date was used to generate the age of the patient. In addition, many patients were missing important feature values like Fasting Plasma Glucose (FPG) and Hemoglobin A1c (HbA1c). Since both features were used to initially classify a person as diabetic or non-diabetic, to establish the ground truth, all the instances that did not have these feature values were removed. As the number of missing features was very high, replacing the missing values for both features was not desirable. After filtering the patients, our dataset decreased to 225 eligible patients for classification. However, 43 out of 225 patients were missing HDL and LDL values. HDL is considered as “Good Cholesterol”—higher HDL means better state—while LDL is considered as “Bad Cholesterol” therefore, lower LDL values are desirable. In the experiments, when HDL and LDL values were used, the records with missing values were dropped. FPG and HbA1c values were also transformed using the American Diabetes Association (ADA) as reference for the different value ranges [48].

### 3.3. Subject Exclusion

In this study, we excluded subjects whose age was less than 19 years to focus on the prediction of T2D by reducing the chances of T1D, which usually develops in children and adolescents. Previous work [32–34] also excluded similar data as well as data indicated as gestational diabetes, which is relevant to pregnant women however, since we lack information on pregnancy, we did not perform this step. By excluding these subjects, we were finally left with 162 instances.

### 3.4. Feature Engineering

Of the 16 features mentioned in Table 1, we had to apply techniques to modify some features to make them suitable for ML algorithms for improved classification. We proceeded as follows. The date of birth was replaced by the age feature. All the features containing diagnosis information (primary, secondary, and their full names) were removed as the diagnosis of patients included multiple diagnoses, most importantly T1D and T2D, and was removed to avoid leaking the classification information into the machine learning model. Finally, the region and diagnosis start date features were also removed.

After the initial feature selection, the dataset obtained consisted of 10 features: Age, height, weight, gender, Hypertension (HTN), Physical Activity Level (PAL), lab tests of Lipoprotein levels (HDL and LDL), Fasting Plasma Glucose (FPG) and Hemoglobin A1c (HbA1c). We would like to mention that age, height, weight, HDL, LDL, FPG, and HbA1c were all numerical features, while gender (M or F), HTN (Yes or No), and PAL (L, M, or H) were categorical features containing text or literals. As our implementation is done in the scikit-learn library (<http://scikit-learn.org/>), whose methods require numerical data for efficient processing, we converted the categories to numerical values. Instead of replacing text with numbers (e.g., L:0, M:1, H:2), we used one-hot encoding to prevent the implicit ordering caused by the numeric values.

At this stage, our data processing steps were finished. Before starting the analysis, it was imperative to identify each record as representing data for a diabetic or non-diabetic patient. In other words, each record was to be labeled with an appropriate class.

### 3.5. Label Assignment

The appropriate references to use for evaluating diabetes were the “Standards of Medical Care in Diabetes—2018” from the American Diabetes Association (ADA) [49] and considering the algorithm proposed by the American Association of Clinical Endocrinologists (AACE) [48]. Two medical experts were also consulted who guided in the diagnosis of diabetes including the factors related to predicting the development of diabetes among people. Their suggestions agreed with the ADA and AACE specifications. Based upon these criteria, any of the FPG or HbA1c laboratory tests could be used to classify patients into either a Diabetic (Y) or Non-Diabetic (N) class. Thus, we proceeded to create two different datasets based on the class labeling scheme. Using an algorithm, the data was automatically labeled in the datasets with either of these classes using the criteria.

#### 3.5.1. Dataset-1: HbA1c Labeling

In this case, a dataset was created by labeling each instance as diabetic if the value of HbA1c  $\geq 6.5\%$  (48 mmol/mol) otherwise it was classified as non-diabetic. This labeling resulted in  $n = 79$  ( $\approx 49\%$ ) instances as diabetic and  $n = 83$  ( $\approx 51\%$ ) as non-diabetic. We can see that the dataset is quite balanced.

#### 3.5.2. Dataset-2: FPG Labeling

In this case, a dataset was created by labeling each instance as diabetic if the value of FPG  $\geq 126$  mg/dL (7.0 mmol/L) otherwise it was classified as non-diabetic. This labeling resulted in  $n = 62$  ( $\approx 38\%$ ) instances as diabetic and  $n = 100$  ( $\approx 62\%$ ) as non-diabetic. Although the dataset with FPG labeling is not quite balanced as in the case of HbA1c labeling, it cannot be characterized as imbalanced either.

Thus, we get two labeled datasets with 8 common features (age, weight, height, gender, PAL, HTN, LDL, and HDL) and using one of the FPG and HbA1c features as input and the other as the label in each dataset. For convenience, we refer to these datasets as HbA1c-labeled and FPG-labeled datasets, where the HbA1c-labeled dataset contains FPG as an input feature and vice-versa.

## 4. Model Development

To analyze the effect of the choice of the HbA1c or FPG labeling attributes on the datasets with the remaining attributes common between the two datasets, we performed the task of diabetes prediction using five machine learning classifiers. Each classifier was evaluated against both datasets. The details of the classifiers and results of the predictions will be discussed in Section 5.

After getting the prediction results on the initial datasets, we planned on improving the results further by performing further analysis and evaluation through feature importance and feature elimination.

#### 4.1. Feature Importance and Feature Elimination

Feature selection aims at filtering out features that may carry redundant information. It is a widely-recognized important task in machine learning with the aim of reducing the chances of overfitting of a model on a dataset [46]. There are several ways to select features for a model. One way is to use those features which are important in the predictive power to affect the classification accuracy. Based on the score assigned to each feature, its usefulness in predicting a target variable can be estimated. Many models provide an intrinsic mechanism to rank the features according to the value of their coefficients (e.g., in Support Vector Machines or Logistic Regression) or using the split-criteria (e.g. in Decision Tree and Random Forest). The correlation between various features can also be used to discover more relevant and important features [50].

While classifiers like linear SVM and linear logistic regression are suitable for interpretation in the form of linear relationships among the variables, they fail to discover complex, non-linear dependencies in the data. Decision trees are suitable for finding interpretable non-linear prediction rules, but there have been some concerns about their instability and lack of smoothness [51]. Similarly, RF models are found to be biased by giving importance to categorical variables with a large number of categories [52,53]. More explicit and advanced mechanisms include the method of Recursive Feature Elimination (RFE) which provides the flexibility of choosing the number of features to select or the algorithm used in choosing the features [54]. The impact of an exploratory variable on a response variable is usually interpreted in isolation, this is usually inappropriately interpreted as an impact for business or medical insight purposes [55].

Permutation importance is one technique recently proposed for identifying measures of feature importance [53,55]. It is a reliable technique that directly measures variable importance by observing the effect on model accuracy by randomly shuffling each predictor variable. In addition, it does not rely on internal model parameters, such as linear regression coefficients, and can be used with other models such as those developed using RF.

Feature elimination aims to reduce the number of input variables when developing a predictive model. The objective is to remove the features that may be non-informative or redundant predictors in the model [56]. By reducing the input variables, not only is the computational cost of modeling reduced, but it may also result in improving the performance of the model. By eliminating weak predictors, we can also improve the generality of the model [55]. Although our datasets comprise a small number of features as well as a relatively small number of instances, we were more concerned with improving the performance of the models through feature selection and elimination. Thus, we applied permutation importance followed by hierarchical clustering to identify the features that could be eliminated.

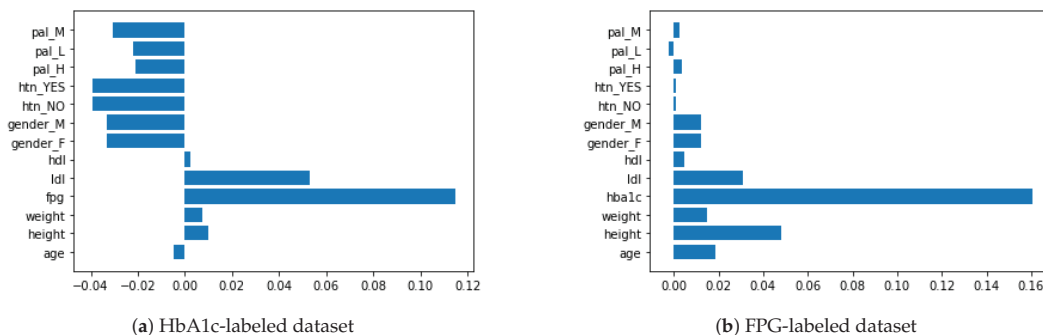


Figure 1. Permutation importance applied to the 9 features in HbA1c- and FPG-labeled datasets.

Figure 1 depicts the permutation importance applied to the two datasets. Although we can observe the importance of the FPG and HbA1c features in the HbA1c- and FPG-labeled

datasets, respectively, the importance of the remaining variables is not consistent, given that they have all the eight features in common. This inconsistency is because the categorical features have been broken down using the one-hot encoding (as explained in the feature engineering subsection), resulting in collinearity among the features. For example, HTN (Yes/No) and Gender (Male/Female) are inversely correlated features. This is also evident in the case of Figure 1, where the collinear features have almost identical importance values. To avoid multicollinearity, as in our case, a strategy is to cluster features that are correlated and only keep one feature from each cluster. We applied Spearman’s correlation by ranking the values of the variables and then running a standard Pearson’s correlation on those ranked variables as proposed by Parr et al. [53]. This resulted in a linkage matrix that is used to infer three main clusters, divided into further subclusters, as shown in the dendrograms for each dataset in Figure 2.

Compared to the permutation importance shown in Figure 1, the agglomerative hierarchical clustering in Figure 2 is consistent for both the datasets. Notably, the HTN=Yes feature is in the same cluster as the label (FPG or HbA1c) of the dataset, which implies they are close in their importance. Similarly, the height and Gender=M features are in the same cluster and are among the least predictive features. The final step is to flatten the cluster to its cluster components. By using the distance among the clusters as a criterion for cluster flattening, obtained from the linkage matrix computed earlier, we can identify the features that can be eliminated from the dataset. This process was applied to both datasets and Gender=M was the only feature that could be eliminated.

With one feature less than the total number of initial features, we performed the classification task once again. This is explained in the next section.

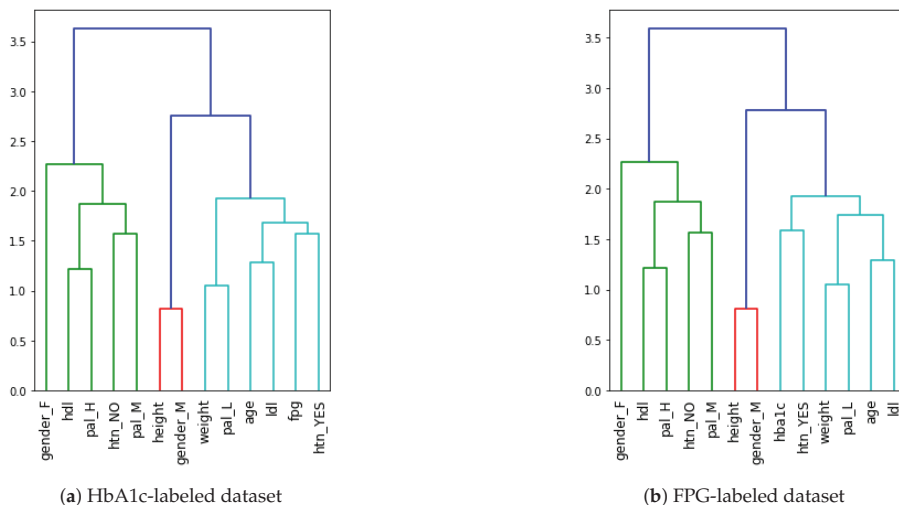


Figure 2. Dendrograms showing hierarchical clustering of the features in our datasets.

#### 4.2. Selection of Machine Learning Classifiers

We chose five machine learning classifiers to evaluate the two datasets. These include three simple learners: Logistic Regression (LR), Support Vector Machines (SVM), and Decision Tree (DT) and two ensemble learners: Random Forest (RF) and Ensemble Majority Voting (EMV). RF uses a set of homogenous decision trees as its base classifiers while the EMV classifier was composed of the three simple learners LR, SVM, and DT, and used hard voting that considered the majority for predicting the class label for each instance in the test set. The rationale for choosing these is based on their previous performance reports in similar situations [9,31,32]. As our objective was to understand the factors contributing to



the classification, we chose not to use any neural networks-based classifier in our analysis due to their “black-box” nature of interpretation of the model [34,57].

### 5. Results

For each dataset, two types of experiments were performed with all the classifiers. In the first experiment, all nine input features were used. In the second experiment, we performed feature selection and elimination before training and evaluating the classifiers, which resulted in eliminating one feature (Gender=M) from the dataset. With the eight final features, we performed the prediction task once again.

#### 5.1. Performance of Machine Learning Classifiers

To measure the performance of each classifier, we used the widely-accepted performance statistics: Accuracy, precision, recall, and F1-score [58]. For model evaluation, we used 10-fold cross-validation in all experiments. The RF classifier used  $n = 100$  estimators with max depth set to 40. Other parameters were left as default by the scikit-learn library. Both datasets were evaluated with the same model configurations. To allow reproducing the same splits across different experiments, we used the same seed for generating the random state for both datasets.

Table 2 and 3 describe the comparative performance of the five classifiers against each performance metric in the two experiments. The metrics represent the weighted average of the cross-validation. We learn the following from these tables:

- The performance of all classifiers was better in the FPG-labeled dataset as compared to the HbA1c-labeled dataset;
- SVM performed best on the HbA1c-labeled dataset while RF performed best on the FPG-labeled dataset;
- There was no change in the performance of SVM after feature elimination in both cases, while all the other classifiers saw an improvement or a decrease in the performance after feature elimination;
- The performance of DT and EVM classifiers improved, but that of RF decreased, after feature elimination in both cases.

The classification results are comparable to existing approaches for diabetes classification [9,31,32,34].

Table 2. Performance evaluation of the HbA1c-labeled dataset.

	Evaluation with 9 Features				Evaluation with 8 Features			
	Accuracy	Precision	Recall	F1-Score	Accuracy	Precision	Recall	F1-Score
Logistic Regression	80.86	80.95	80.86	80.83	80.86 ↔	80.95 ↔	80.86 ↔	80.83 ↔
SVM	<b>82.10</b>	<b>82.30</b>	<b>82.10</b>	<b>82.05</b>	<b>82.10</b> ↔	<b>82.30</b> ↔	<b>82.10</b> ↔	<b>82.05</b> ↔
Decision Tree	74.07	74.07	74.07	74.06	75.31 ↑	75.34 ↑	75.31 ↑	75.28 ↑
Random Forest	81.48	81.91	81.48	81.38	80.86 ↓	81.61 ↓	80.86 ↓	80.70 ↓
Ensemble	77.78	78.14	77.78	77.66	78.40 ↑	78.86 ↑	78.40 ↑	78.26 ↑

Table 3. Performance evaluation of the FPG-labeled dataset.

	Evaluation with 9 Features				Evaluation with 8 Features			
	Accuracy	Precision	Recall	F1-Score	Accuracy	Precision	Recall	F1-Score
Logistic Regression	83.33	83.31	83.33	83.04	82.72 ↓	82.62 ↓	82.72 ↓	82.45 ↓
SVM	84.57	84.74	84.57	84.22	84.57 ↔	84.74 ↔	84.57 ↔	84.22 ↔
Decision Tree	80.86	81.50	80.86	81.03	82.72 ↑	83.01 ↑	82.72 ↑	82.81 ↑
Random Forest	<b>88.27</b>	<b>88.31</b>	<b>88.27</b>	<b>88.29</b>	<b>87.65</b> ↓	<b>87.90</b> ↓	<b>87.65</b> ↓	<b>87.72</b> ↓
Ensemble	83.95	83.84	83.95	83.84	84.57 ↑	84.47 ↑	84.57 ↑	84.43 ↑

## 6. Discussion

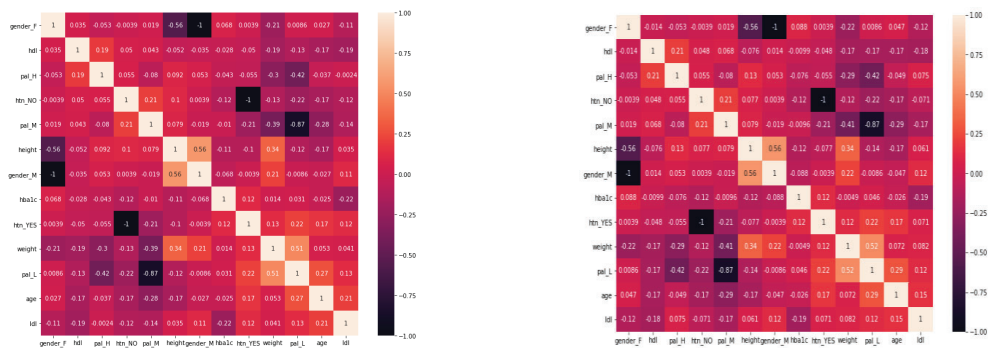
Together with other features in the form of lab tests (LDL and HDL) as well as patient's information (age, gender, height, weight, hypertension, and physical activity level), we used HbA1c and FPG as features in two separate datasets for classifying an instance as diabetic or non-diabetic. In our experiments, we found that all five different classifiers predicted with better performance on the FPG-labeled dataset that included HbA1c as one of the input features. This implies that HbA1c can be used as a superior variable than FPG for diabetes prediction. This is consistent with the previous work as well. In a previous study on Vietnamese patients [59], researchers collected overnight fasting blood samples from 3523 individuals (of which 2356 were women). Like our case, diabetes was diagnosed with an HbA1c value  $\geq 6.5\%$  or an FPG level  $\geq 7$  mmol/l. It was concluded that HbA1c testing had a higher sensitivity for identifying patients at risk for diabetes vs FPG, and therefore may have a greater impact on the diagnoses, cost, burden, and treatment of patients with diabetes [59].

When compared with the existing approaches, we can identify some distinguishing features of our approach. We used only a limited number of basic features (age, gender, height, weight, presence of hypertension, and physical activity level) and three laboratory tests (HDL, LDL, HbA1c, or FPG) to predict if a person has diabetes or not. In contrast, most of the existing approaches use many attributes. For example, Dinh et al. [34] initially used 123 features in diabetes prediction and even after removing the various laboratory tests, they were left with a much higher number of features (the exact number is not known). Finding these many features in real-world data is rarely possible. So, we proposed a mechanism whereby only with a few features could we infer the role played by them in the classification of a person into diabetic or non-diabetic.

The strategy for the identification of the contribution of each feature through the feature importance is also significant in the current analysis. Mostly, a correlation analysis is performed directly to identify such hidden patterns from data (e.g., as in [44]). However, as can be seen in Figure 1, visualizing the feature importance for different features does not reveal the same information as we have inferred from our results. Thus, we had to carry out a certain transformation in the form of clustering and distance evaluation to perform feature elimination. While we did not use correlation in the prediction task, the ranked correlations obtained during an intermediate step of our model development can be used to add to our findings.

### 6.1. Analysis of Diabetes Risk Factors

Figure 3a shows the correlation matrix of the initial dataset obtained after feature engineering, without applying any transformations, and Figure 3b shows the correlation matrix obtained after the ranked correlation based on the Spearman's ranking. While there have been small changes in some of the values, after the transformation, the correlation between the variables largely remains the same during the transformation process. Thus, the transformation has not affected the original relationship between the variables and the data remains integrated. The values of the correlation between some of the features of interest are shown in Table 4. We have organized the table into three sections: Lab results of HDL and LDL, hypertension, and personal attributes of age, weight, and height.



(a) FPG-labeled dataset before any transformations

(b) FPG-labeled dataset after going through the transformation process

**Figure 3.** Comparison of correlation before and after hierarchical clustering based on Spearman's ranking

From the correlation analysis, we can establish the following information:

1. When we compare LDL to HbA1c and FPG, LDL is more correlated with HbA1c than FPG. On the other hand, HDL has almost no impact on HbA1c (close to 0);
2. The presence of hypertension is correlated with an increase in age as well as with a lower level of physical activity. Lower PAL is associated with hypertension while medium PAL is associated with the absence of hypertension;
3. Hypertension is also correlated with increasing levels of HbA1c and FPG, with an almost similar impact on both;
4. A higher level of physical activity has a good impact on HDL (the “good” cholesterol), while a low level of physical activity may cause higher levels of LDL (the “bad” cholesterol);
5. As the age of a person increases, so does LDL, meaning that younger people have comparatively small levels of dangerous cholesterol as compared to the older ones;
6. The level of physical activity decreases with the age. Thus, older people lack physical activities;
7. The level of physical activity of a person has also a strong correlation with the weight of a person, i.e., lower PAL indicates more weight while higher PAL is correlated with less weight of a person. Also, males have more weight when compared to females;
8. The height of a person is negatively correlated with both HbA1c and FPG. Accordingly, shorter people may be at higher risk of diabetes. This is also in accordance with existing findings [60,61]. In comparison, when we evaluate the weight of a person against HbA1c and FPG, there is almost no correlation between them (0.01);
9. There is no significant, direct relation between PAL and either HbA1c or FPG ( $< 0.05$  | in all cases of HbA1c and FPG with all PALs). Thus, we conclude that PAL has effects on weight, HTN, LDL, and HDL, which then have an impact on HbA1c and FPG levels, leading to diabetes.

These insights give us some hints into the diabetic disease, its development, and associated complications.

**Table 4.** Correlation between various features after ranking.

LDL and HDL			Hypertension			Age, Weight, and Height		
Feature 1	Feature 2	Corr	Feature 1	Feature 2	Corr	Feature 1	Feature 2	Corr
LDL	HbA1c	−0.19	HTN=Yes	Age	0.17	Age	LDL	0.15
LDL	FPG	−0.11	HTN=Yes	PAL=L	0.22	Age	PAL=L	0.29
LDL	PAL=L	0.12	HTN=No	PAL=M	0.21	Weight	PAL=L	0.52
HDL	PAL=H	0.21	HTN=Yes	HbA1c	0.12	Weight	PAL=H	−0.29
HDL	FPG	−0.14	HTN=Yes	FPG	0.13	Weight	Gender=M	0.22
HDL	HbA1c	−0.01				Height	HbA1c	−0.12
						Height	FPG	−0.12

*6.2. Recommendations*

With insights from the current work, we can present some recommendations. First, we can see that even with limited data, patients can be pre-screened for diabetes, and in case of their classification as diabetic, they can be advised to make appropriate changes to their lifestyle. We have found that physical activity plays an important role in diabetes development. Lower levels of physical activity were found to correlate with more weight, higher levels of LDL (the “bad” cholesterol), as well as hypertension. Thus, everyone needs to include higher levels of physical activity in their daily routines and avoid associated complications. Second, LDL has been found to have an association with HbA1c and FPG and the LDL levels also increase as a person progresses in age. In a similar fashion, HDL levels are associated with FPG. Thus, it is important that people perform regular LDL/HDL tests and to control their levels in case it increases with time.

In our current work, we only had access to the hypertension feature of a patient as being Yes or No. However in practice, the patient’s blood pressure is recorded as diastolic and systolic values. Similarly, temperature, vision, waist size, etc. are some other features that can be recorded with commonly available instruments in every clinic. Thus, just like age and weight play an indirect role in diabetes prediction, these and other features may also play a certain role in diabetes prediction and should be recorded for each patient to improve the diagnostic process. Finally, the government could enforce the pre-screening of diabetes based on age, weight, physical activity levels, and the presence of hypertension. These factors do not require any specialized tests or equipment and can be checked in any clinic, even in rural areas. By controlling these basic factors, a large segment of the population can be averted from developing early diabetes, a problem that has a large economic and social burden in many countries including Saudi Arabia. It is also important that accurate recording of physiological data should be enforced by hospitals and local clinics for any visiting patients for better opportunities to diagnose patients-at-risk. The data should be recorded in the patient’s EHR so it can be used via a similar analysis on a larger scale to produce better analysis in the future.

*6.3. Limitations of Work*

We can also identify a few limitations within our work. First, as data availability is an important issue in health science research, although our data concerned 3000 patients, the final size of data was very small. The performance accuracy of a classification task mainly depends on the availability of large amounts of data [62] and with large data, we may have better insights. Unfortunately, our final dataset had only 199 records and after removing the missing values found for LDL and HDL features, we had only 162 records with complete feature values. With such small-scale data, there are limited options to test the available classifiers as well as the configuration of their various hyperparameters. That

is why we did not invest time in further optimizing our classifiers for the given small dataset. With more data, better classifiers can be trained, evaluated, and optimized.

Second, the data were obtained in the context of Saudi Arabia. It would be interesting to test our approach to similar datasets from other countries/regions of the world. Third, in our current work, we used common machine learning classifiers. After establishing the feature importance of various features, we could even utilize black-box approaches like machine learning or deep learning and achieve state-of-the-art performance evaluation results [15–17]. This is one of our near-future goals.

## 7. Conclusions

The prevalence of diabetes is not only a burden for the governments in terms of the associated expenditures, but it is also a lifelong strain on diabetic patients. HbA1c and FPG are two important features for diabetes classification. With a dataset having both these important features for diabetes analysis, we constructed two separate datasets that classified an instance into diabetic or non-diabetic class. We found that HbA1c used as a feature resulted in better performance (accuracy, precision, and recall) as compared to FPG. Moreover, we also identified several other features like hypertension, weight, and physical activity levels that had an indirect role in diabetic prediction. The LDL/HDL tests were also found to be correlated with diabetic conditions.

With data from other countries, our approach could be generalized, which may have important implications in the healthcare community. The prescreening of diabetes could be rapid, people could be more aware and educated about their lifestyles, and government expenditures could be reduced alongside the decrease in the significant burden on hospitals due to the prevalence of diabetes. With the ability to predict the onset of diabetes, necessary steps can be taken to avoid the diabetic stage of millions of people who are undiagnosed due to limited resources and lack of awareness. This can not only improve a person's quality of life but also result in a positive impact on the healthcare system. Several recommendations have been proposed in this article in this regard.

**Author Contributions:** Conceptualization, H.F.A., H.A. and H.M.; methodology, H.M. and H.A.; validation, H.M. and M.S.; formal analysis, H.M., H.F.A., and M.S.; writing—original draft preparation, H.M., H.A., A.A.; writing—review and editing, H.M., H.A., H.F.A., M.S. and A.A. All authors have read and agreed to the submitted version of the manuscript.

**Funding:** This research was funded by the Deputyship for Research & Innovation, Ministry of Education in Saudi Arabia through the project number IFT20129.

**Institutional Review Board Statement:** Ethical review and approval were not required for this study, as it did not involve actual humans directly or indirectly in the study. The electronic health records obtained were provided anonymized by the source institutions and the study did not modify or applied any changes to the data. The research only analyzed a set of data without any referral to any human.

**Informed Consent Statement:** Anonymized data without any reference to any patient was obtained and used in this research. The data cannot be traced back to the patients, so informed consent was not needed in this research.

**Data Availability Statement:** The authors choose not to make the data available yet. It might be available in the future.

**Acknowledgments:** The authors extend their appreciation to the Deputyship for Research & Innovation, Ministry of Education in Saudi Arabia for funding this research work through project number IFT20129. The authors also acknowledge the Deanship of Scientific Research at King Faisal University for the financial support Institutional Financing Track 2020.

**Conflicts of Interest:** The authors declare no conflict of interest. The funders had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript, or in the decision to publish the results.

## Abbreviations

The following abbreviations are used in this manuscript:

EMR	Electronic Medical Record
SVM	Support Vector Machines
LR	Logistic Regression
DT	Decision Tree
RF	Random Forest
EVM	Ensemble Voting Model
DM	Diabetes Mellitus
T1D	Type 1 Diabetes
T2D	Type 2 Diabetes
HTN	Hypertension
PAL	Physical Activity Level

## References

- Saeedi, P.; Petersohn, I.; Salpea, P.; Malanda, B.; Karuranga, S.; Unwin, N.; Colagiuri, S.; Guariguata, L.; Motala, A.A.; Ogurtsova, K.; others. Global and regional diabetes prevalence estimates for 2019 and projections for 2030 and 2045: Results from the International Diabetes Federation Diabetes Atlas. *Diabetes Res. Clin. Pract.* **2019**, *157*, 107843.
- Cho, N.; Shaw, J.; Karuranga, S.; Huang, Y.; da Rocha Fernandes, J.; Ohlrogge, A.; Malanda, B. IDF Diabetes Atlas: Global estimates of diabetes prevalence for 2017 and projections for 2045. *Diabetes Res. Clin. Pract.* **2018**, *138*, 271–281.
- Al-Rubeaan, K.; Al-Manaa, H.A.; Khoja, T.; Ahmad, N.; Alsharqawi, A.; Siddiqui, K.; Alnaqeb, D.; Aburishheh, K.; Youssef, A.; Al-Batil, A.; Alotaibi, M.S.; Ghamdi, A.A. The Saudi Abnormal Glucose Metabolism and Diabetes Impact Study (SAUDI-DM). *Ann. Saudi Med.* **2014**, *34*, 465–475.
- AlMazroa, M. Cost of Diabetes in Saudi Arabia. *Iproceedings* **2018**, *4*, e10566.
- Alotaibi, A.; Perry, L.; Gholizadeh, L.; Al-Ganmi, A. Incidence and prevalence rates of diabetes mellitus in Saudi Arabia: An overview. *J. Epidemiol. Glob. Health* **2017**, *7*, 211–218.
- Saad, A.M.; Younes, Z.M.; Ahmed, H.; Brown, J.A.; Al Owesie, R.M.; Hassoun, A.A. Self-efficacy, self-care and glycemic control in Saudi Arabian patients with type 2 diabetes mellitus: A cross-sectional survey. *Diabetes Res. Clin. Pract.* **2018**, *137*, 28–36.
- Alsuliman, M.A.; Alotaibi, S.A.; Zhang, Q.; Durgampudi, P.K. A systematic review of factors associated with uncontrolled diabetes and meta-analysis of its prevalence in Saudi Arabia since 2006. *Diabetes/Metab. Res. Rev.* **2020**, DOI:0.1002
- Almutairi, E.; Abbod, M.; Itagaki, T. Mathematical Modelling of Diabetes Mellitus and Associated Risk Factors in Saudi Arabia. *Int. J. Simul.-Sci. Technol.* **2020**, Vol. 21, No. 2, p. 1–7.
- Syed, A.H.; Khan, T. Machine Learning-Based Application for Predicting Risk of Type 2 Diabetes Mellitus (T2DM) in Saudi Arabia: A Retrospective Cross-Sectional Study. *IEEE Access* **2020**, *8*, 199539–199561.
- Alomar, M.J.; Al-Ansari, K.R.; Hassan, N.A. Comparison of awareness of diabetes mellitus type II with treatment's outcome in term of direct cost in a hospital in Saudi Arabia. *World J. Diabetes* **2019**, *10*, 463.
- Nathan, D.; Buse, J.; Davidson, M.; Heine, R.; Holman, R.; Sherwin, R.; Zinman, B. Management of hyperglycaemia in type 2 diabetes: A consensus algorithm for the initiation and adjustment of therapy. *Diabetologia* **2006**, *49*, 1711–1721.
- Sacks, D. A1C Versus Glucose Testing: A Comparison. *Diabetes Care* **2011**, *34*, 518–523.
- World Health Organization. Definition and Diagnosis of Diabetes Mellitus and Intermediate Hyperglycemia: Report of a WHO/IDF Consultation. 2006. Available online: [https://apps.who.int/iris/bitstream/handle/10665/43588/9241594934\\_eng.pdf](https://apps.who.int/iris/bitstream/handle/10665/43588/9241594934_eng.pdf). Last Accessed: Jan 26, 2021.
- American Diabetes Association. 2. Classification and diagnosis of diabetes: Standards of medical care in diabetes—2019. *Diabetes Care* **2019**, *42*, S13–S28.
- Wang, Q.; Cao, W.; Guo, J.; Ren, J.; Cheng, Y.; Davis, D.N. DMP\_MI: An effective diabetes mellitus classification algorithm on imbalanced data with missing values. *IEEE Access* **2019**, *7*, 102232–102238.
- Kaur, P.; Kaur, R. Comparative Analysis of Classification Techniques for Diagnosis of Diabetes. In *Advances in Bioinformatics, Multimedia, and Electronics Circuits and Signals*; Springer: Berlin/Heidelberg, Germany, 2020; pp. 215–221.
- Devi, R.H.; Bai, A.; Nagarajan, N. A novel hybrid approach for diagnosing diabetes mellitus using farthest first and support vector machine algorithms. *Obes. Med.* **2020**, *17*, 100152.
- Abbas, H.; Alic, L.; Erraguntla, M.; Ji, J.; Abdul-Ghani, M.; Abbasi, Q.H.; Qaraqe, M. Predicting long-term Type 2 Diabetes with Support Vector Machine using Oral Glucose Tolerance Test. *bioRxiv* **2019**, DOI:<https://doi.org/10.1371/journal.pone.0219636>.
- Kadhm, M.S.; Ghindawi, I.W.; Mhawi, D.E. An Accurate Diabetes Prediction System Based on K-means Clustering and Proposed Classification Approach. *Int. J. Appl. Eng. Res.* **2018**, *13*, 4038–4041.
- Afzali, S.; Yildiz, O. An effective sample preparation method for diabetes prediction. *Int. Arab J. Inf. Technol.* **2018**, *15*, 968–973.
- Tuso, P. Prediabetes and lifestyle modification: Time to prevent a preventable disease. *Perm. J.* **2014**, *18*, 3, 88–93.



22. Huxley, R.; James, W.; Barzi, F.; Patel, J.; Lear, S.; Suriyawongpaisal, P.; Janus, E.; Caterson, I.; Zimmet, P.; Prabhakaran, D.; et al. Ethnic comparisons of the cross-sectional relationships between measures of body size with diabetes and hypertension. *Obes. Rev.* **2008**, *9*, 53–61.
23. Zhu, Y.; Hedderson, M.M.; Quesenberry, C.P.; Feng, J.; Ferrara, A. Liver enzymes in early to mid-pregnancy, insulin resistance, and gestational diabetes risk: A longitudinal analysis. *Front. Endocrinol.* **2018**, *9*, 581.
24. Lomonaco, R.; Leiva, E.G.; Bril, F.; Shrestha, S.; Mansour, L.; Budd, J.; Romero, J.P.; Schmidt, S.; Chang, K.L.; Samraj, G.; et al. Advanced Liver Fibrosis Is Common in Patients With Type 2 Diabetes Followed in the Outpatient Setting: The Need for Systematic Screening. *Diabetes Care* **2020**, *44*(2), 399–406.
25. Jaiswal, M.; Divers, J.; Dabelea, D.; Isom, S.; Bell, R.A.; Martin, C.L.; Pettitt, D.J.; Saydah, S.; Pihoker, C.; Standiford, D.A.; et al. Prevalence of and risk factors for diabetic peripheral neuropathy in youth with type 1 and type 2 diabetes: SEARCH for Diabetes in Youth Study. *Diabetes Care* **2017**, *40*, 1226–1232.
26. Rawshani, A.; Rawshani, A.; Franzén, S.; Sattar, N.; Eliasson, B.; Svensson, A.M.; Zethelius, B.; Miftaraj, M.; McGuire, D.K.; Rosengren, A.; et al. Risk factors, mortality, and cardiovascular outcomes in patients with type 2 diabetes. *N. Engl. J. Med.* **2018**, DOI: 10.1056/NEJMoa1800256.
27. Mendola, N.D.; Chen, T.C.; Gu, Q.; Eberhardt, M.S.; Saydah, S. *Prevalence of Total, Diagnosed, and Undiagnosed Diabetes among Adults: United States, 2013–2016*; NCHS Data Brief No. 319; US Department of Health and Human Services, Centers for Disease Control and Prevention, National Center for Health and Statistics, USA. 2018.
28. Daanouni, O.; Cherradi, B.; Tmiri, A. Type 2 diabetes mellitus prediction model based on machine learning approach. In Proceedings of the Third International Conference on Smart City Applications, Casablanca, Morocco, 2–4 October 2019; Springer: Berlin/Heidelberg, Germany, 2019; pp. 454–469.
29. Lai, H.; Huang, H.; Keshavjee, K.; Guergachi, A.; Gao, X. Predictive models for diabetes mellitus using machine learning techniques. *BMC Endocr. Disord.* **2019**, *19*, 1–9.
30. Alić, B.; Gurbeta, L.; Badnjevic, A. Machine learning techniques for classification of diabetes and cardiovascular diseases. In Proceedings of the 2017 6th Mediterranean Conference on Embedded Computing (MECO), Bar, Montenegro, 11–15 June 2017; pp. 1–4.
31. Uddin, S.; Khan, A.; Hossain, M.E.; Moni, M.A. Comparing different supervised machine learning algorithms for disease prediction. *BMC Med. Inform. Decis. Mak.* **2019**, *19*, 1–16.
32. Yu, W.; Liu, T.; Valdez, R.; Gwinn, M.; Khoury, M.J. Application of support vector machine modeling for prediction of common diseases: The case of diabetes and pre-diabetes. *BMC Med. Inform. Decis. Mak.* **2010**, *10*, 16.
33. Semerdjian, J.; Frank, S. An ensemble classifier for predicting the onset of type II diabetes. *arXiv* **2017**, arXiv:1708.07480.
34. Dinh, A.; Miertschin, S.; Young, A.; Mohanty, S. A data-driven approach to predicting diabetes and cardiovascular disease with machine learning. *BMC Med. Inform. Decis. Mak.* **2019**, *19*.
35. Martín-González, F.; González-Robledo, J.; Sánchez-Hernández, F.; Moreno-García, M.N. Success/Failure Prediction of Noninvasive Mechanical Ventilation in Intensive Care Units. *Methods Inf. Med.* **2016**, *55*, 234–241.
36. Akay, M.F. Support vector machines combined with feature selection for breast cancer diagnosis. *Expert Syst. Appl.* **2009**, *36*, 3240–3247.
37. Chen, H.L.; Yang, B.; Liu, J.; Liu, D.Y. A support vector machine classifier with rough set-based feature selection for breast cancer diagnosis. *Expert Syst. Appl.* **2011**, *38*, 9014–9022.
38. Liaqat, S.; Dashtipour, K.; Zahid, A.; Assaleh, K.; Arshad, K.; Ramzan, N. Detection of atrial fibrillation using a machine learning approach. *Information* **2020**, *11*, 549.
39. YA Amer, A.; Vranken, J.; Wouters, F.; Mesotten, D.; Vandervoort, P.; Storms, V.; Luca, S.; Vanrumste, B.; Aerts, J.M. Feature Engineering for ICU Mortality Prediction Based on Hourly to Bi-Hourly Measurements. *Appl. Sci.* **2019**, *9*, 3525.
40. Tomar, D.; Agarwal, S. Hybrid feature selection based weighted least squares twin support vector machine approach for diagnosing breast cancer, hepatitis, and diabetes. *Adv. Artif. Neural Syst.* **2015**, DOI:http://dx.doi.org/10.1155/2015/265637.
41. Balakrishnan, S.; Narayanaswamy, R.; Savarimuthu, N.; Samikannu, R. SVM ranking with backward search for feature selection in type II diabetes databases. In Proceedings of the 2008 IEEE International Conference on Systems, Man and Cybernetics, Singapore, 12–15 October 2008; pp. 2628–2633.
42. Ephzibah, E. Cost effective approach on feature selection using genetic algorithms and fuzzy logic for diabetes diagnosis. *arXiv* **2011**, arXiv:1103.0087.
43. Aslam, M.W.; Zhu, Z.; Nandi, A.K. Feature generation using genetic programming with comparative partner selection for diabetes classification. *Expert Syst. Appl.* **2013**, *40*, 5402–5412.
44. Rodríguez-Rodríguez, I.; Rodríguez, J.V.; González-Vidal, A.; Zamora, M.Á. Feature Selection for Blood Glucose Level Prediction in Type 1 Diabetes Mellitus by Using the Sequential Input Selection Algorithm (SISAL). *Symmetry* **2019**, *11*, 1164.
45. Butterworth, R.; Piatetsky-Shapiro, G.; Simovici, D.A. On feature selection through clustering. In Proceedings of the Fifth IEEE International Conference on Data Mining (ICDM'05), Houston, TX, USA, 27–30 November 2005.
46. Ienco, D.; Meo, R. Exploration and reduction of the feature space by hierarchical clustering. In Proceedings of the 2008 SIAM International Conference on Data Mining, Atlanta, GA, USA, 24–26 April 2008; pp. 577–587.
47. Dua, D.; Graff, C. UCI Machine Learning Repository. University of California, Irvine, School of Information and Computer Sciences 2017. Online: <http://archive.ics.uci.edu/ml>, last accessed Jan 26, 2021.



48. American Diabetes Association. Standards of medical care in diabetes—2018 abridged for primary care providers. *Clin. Diabetes A Publ. Am. Diabetes Assoc.* **2018**, *36*, 14.
49. Rodbard, H.; Jellinger, P.; Davidson, J.; Einhorn, D.; Garber, A.; Grunberger, G.; Handelsman, Y.; Horton, E.; Lebovitz, H.; Levy, P.; et al. Statement by an American Association of Clinical Endocrinologists/American College of Endocrinology consensus panel on type 2 diabetes mellitus: An algorithm for glycemic control. *Endocr. Pract.* **2009**, *15*, 540–559.
50. Zien, A.; Krämer, N.; Sonnenburg, S.; Rätsch, G. The feature importance ranking measure. In Proceedings of the Joint European Conference on Machine Learning and Knowledge Discovery in Databases, Bled, Slovenia, September 7–11 2009; Springer: Berlin/Heidelberg, Germany, 2009; pp. 694–709.
51. Hastie, T.; Tibshirani, R.; Friedman, J. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*; Springer Science & Business Media: Berlin/Heidelberg, Germany, 2009.
52. Altmann, A.; Toloşi, L.; Sander, O.; Lengauer, T. Permutation importance: A corrected feature importance measure. *Bioinformatics* **2010**, *26*, 1340–1347.
53. Parr, T.; Turgutlu, K.; Csiszar, C.; Howard, J. Beware Default Random Forest Importances. March 26, 2018. Online: <https://explained.ai/rf-importance/>, last accessed Jan 26, 2021.
54. Chen, X.w.; Jeong, J.C. Enhanced recursive feature elimination. In Proceedings of the Sixth International Conference on Machine Learning and Applications (ICMLA 2007), Cincinnati, OH, USA, 13–15 December 2007; pp. 429–435.
55. Parr, T.; Wilson, J.D.; Hamrick, J. Nonparametric Feature Impact and Importance. *arXiv* **2020**, arXiv:2006.04750.
56. Kuhn, M.; Johnson, K. *Applied Predictive Modeling*; Volume 26, Springer: Berlin/Heidelberg, Germany, 2013.
57. Tu, J.V. Advantages and disadvantages of using artificial neural networks versus logistic regression for predicting medical outcomes. *J. Clin. Epidemiol.* **1996**, *49*, 1225–1231.
58. Caruana, R.; Niculescu-Mizil, A. Data mining in metric space: An empirical analysis of supervised learning performance criteria. In Proceedings of the tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Seattle, WA, USA, 22–25 August 2004; pp. 69–78.
59. Ho-Pham, L.T.; Nguyen, U.D.; Tran, T.X.; Nguyen, T.V. Discordance in the diagnosis of diabetes: Comparison between HbA1c and fasting plasma glucose. *PLoS ONE* **2017**, *12*, e0182192.
60. Vangipurapu, J.; Stančáková, A.; Jauhiainen, R.; Kuusisto, J.; Laakso, M. Short adult stature predicts impaired  $\beta$ -cell function, insulin resistance, glycemia, and type 2 diabetes in Finnish men. *J. Clin. Endocrinol. Metab.* **2017**, *102*, 443–450.
61. Wittenbecher, C.; Kuxhaus, O.; Boeing, H.; Stefan, N.; Schulze, M.B. Associations of short stature and components of height with incidence of type 2 diabetes: Mediating effects of cardiometabolic risk factors. *Diabetologia* **2019**, *62*, 2211–2221.
62. Zhou, L.; Pan, S.; Wang, J.; Vasilakos, A.V. Machine learning on big data: Opportunities and challenges. *Neurocomputing* **2017**, *237*, 350–361.

Article

# Multifrequency Impedance Method Based on Neural Network for Root Canal Length Measurement

Xiaoyue Qiao , Zheng Zhang and Xin Chen \*

School of Electronic Information and Electrical Engineering, Shanghai Jiao Tong University, Shanghai 200240, China; xy1121@sjtu.edu.cn (X.Q.); zhangzheng123@sjtu.edu.cn (Z.Z.)

\* Correspondence: xchen.ie@sjtu.edu.cn

Received: 3 September 2020; Accepted: 20 October 2020; Published: 22 October 2020

**Abstract:** Root canal therapy is the most fundamental and effective approach for treating endodontics and periapicalitis. The length of the root canal must be accurately measured to clean the pathogenic substances in it. This study aims to present a multifrequency impedance method based on a neural network for root canal length measurement. A circuit system was designed which generates a current of frequencies from 100 Hz to 20 kHz in order to augment the data of impedance ratios with different combinations of frequencies. Several impedance ratios and other quantified characteristics, such as the type of tooth and file, were selected as features to train a neural network model that could predict the distance between the file and apical foramen. The model uses leave-one-out cross-validation, adopts the Adam optimizer and regularization, and has two hidden layers with nine and five nodes, respectively. The neural network-based multifrequency impedance method exhibits nearly 95% accuracy, compared with the dual-frequency impedance ratio method (which demonstrated no more than 85% accuracy in some situations). This method may eliminate the influence of human and environmental factors on measurement of the root canal length, thereby increasing measurement robustness.

**Keywords:** root canal measurement; multifrequency impedance; data augmentation; neural network

## 1. Introduction

At present, root canal therapy is the most effective treatment for pulpal and periapical diseases. The keys to the process include thoroughly cleaning the root canal system, removing the source of infection, and reducing the damage to the root tip and periapical tissue [1,2]. Thus, root canal therapy requires accurate measurement of the length of root canals [3]. Three approaches exist for measurement of root canal length: the hand feeling method, radiographic determination [4], and the use of an electronic apex locator (EAL). The EAL, which is most commonly used for measuring root canal length in clinical settings, has been developed over a long period using different methods [5].

Root canal measurement was initially based on a resistance model, according to the phenomenon that the direct current resistance between the apical foramen and oral mucosa is almost constant, although the ages, tooth types, and root canal shapes of patients differ widely. The performance of this model has generally been considered unsatisfactory, as the measurement does not reflect a simple resistance model but, rather, a complex model with capacitor characteristics [6,7]. Then, root canal measurement was carried out by a voltage gradient method, which was more accurate than before but still unstable [8,9]. A method based on dual-frequency impedance ratios was proposed in the 1990s. This method uses the relative quantity instead of absolute quantity to make the results universal and reduce the effect of the measuring environment on the results [10]. Several EALs using the dual-frequency impedance ratios method have demonstrated exceptional clinical performance [11,12]. The multifrequency impedance method has been employed, based on the success of dual-frequency

methods [13]; however, the EALs found on the market using the multifrequency impedance method cannot measure the length of root canal very precisely [14,15]. The reason for such low accuracy is that the measurement environment in the root canal is extremely complex with infective substances and infected biological tissue. Deep learning methods have superior ability to cope with challenging environmental noises, so it was considered appropriate to apply a neural network for canal root length measurement. Data augmentation is also employed, as the data are limited by the scarce signals at confined frequencies, which prevents the ability to train an excellent neural network.

Given the inadequacy of the current methodologies, this paper proposes a method for precisely locating the position of the apical foramen using a multifrequency impedance method based on a neural network, which is trained (after data augmentation) by distinct combinations of multifrequency signals generated from the designed circuit system. The method is experimentally verified, demonstrating that the proposed approach can promote accuracy and stability when measuring the length of root canals.

## 2. Material and Methods

The dual-impedance ratio method selects the impedance ratio of one high frequency and one low frequency as the criteria for locating the apical foramen; however, the level of accuracy is not ideal and the method may be greatly affected by the measurement conditions. The multifrequency impedance method proposed in this paper requires multiple impedance ratios of multiple frequency combinations and utilizes the relationships between them to increase the reliability of the results. A large number of impedance ratios of different frequency combinations are essential for high accuracy of measurement. With the data augmentation of impedance ratios, a nonlinear regression model with the distance between the file and the apical foramen as the prediction result is constructed based on deep learning methods. Neural networks are the most widely used type of deep learning model, which exhibit excellent performance in solving complex nonlinear regression problems [16,17]. Considering the delicate electrical properties of organisms, deep learning plays a major role in detecting and processing bio-electric signals [18,19]. In a concrete situation, the structure of the network can be targeted, designed, and optimized, making the model more flexible and efficient. It is crucial to select appropriate features to improve the multifrequency impedance method for root canal length measurement.

### 2.1. Multifrequency Impedance Measurement

All of the measurement instruments and measuring methods used for the experiments were designed and prepared to meet the research requirements. The impedance was measured on 21 extracted teeth that had been treated and cleaned using a variable frequency voltage signal generated by a programmable digital frequency synthesizer. The signal varied from 100 Hz to 20 kHz, and the impedances at different frequencies were collected using automatic data collection (ADC) and stored in a computer. Figure 1 presents a schematic of the whole measurement process.

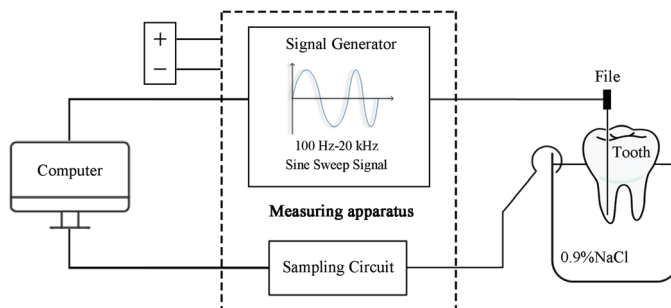


Figure 1. Diagram of the measurement process.

Figure 2 illustrates the physical diagram of the measurement process and its constituent parts. The 21 tooth samples were single root canal teeth from 21 adults of different ages, including incisors, canines, and molars. The teeth were preprocessed in vitro by soaking them in a 2.5% sodium hypochlorite solution and removing the periodontal tissues and attachments, such as dental calculus. The root canal was washed after opening and removing the pulp. The crown of the tooth was polished as a reference plane. A tooth fixation device was used to keep the relative position of the tooth constant during the experiment, ensuring the consistency of the data. A precise file translation device was used to strictly control the change of distance by increments of 0.5 mm with an accuracy of 1 μm. One of the files of types #15, #25, and #40 was fixed on the file translation to be located in the root canal. The root canal length measurement prototype was designed to generate sine sweep signals with multiple frequencies and select the impedance values. The impedance ratios of different frequency combinations could then be obtained through data processing in the prototype.

For each tooth, we changed the file distance from the root tip by +5 mm to −1 mm (where + denotes that the file does not reach the apical foramen and − denotes that the file exceeds the apical foramen).

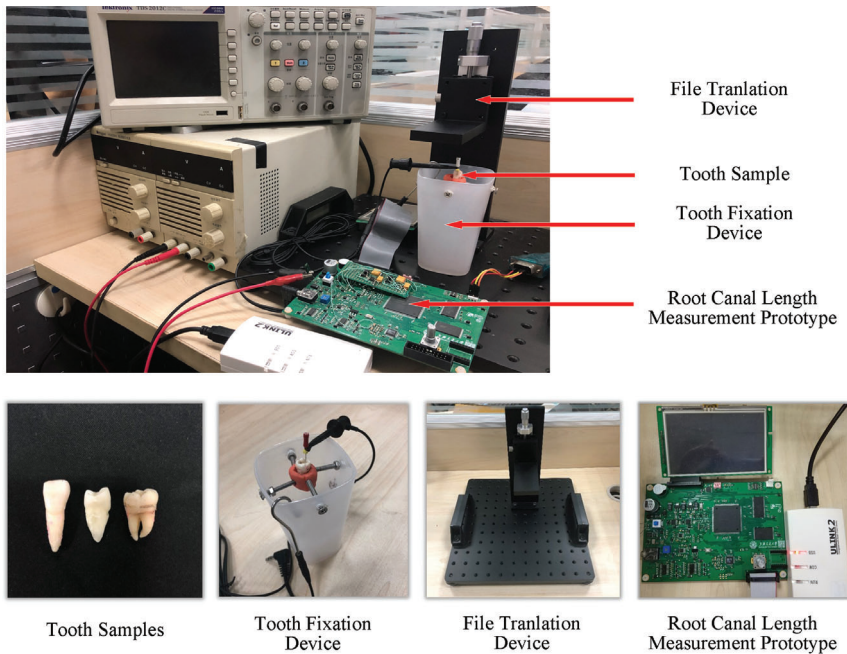


Figure 2. Experimental equipment.

Two theoretical tests were verified during this process: the impedance decreased with increasing frequencies when the file was in the same position; the closer the file was to the apical foramen, the smaller the impedance ratio, and the larger the frequency difference, the smaller the ratio. The detailed verification is explained in the following results section (pre-verification).

## 2.2. Data Augmentation

The impedance ratio was the software-processed output of the circuit system, as well as the critical data in root canal length measurement. Its expression is as follows:

$$\text{Impedance Ratio} = \frac{Z(f_{high})}{Z(f_{low})}, \quad (1)$$

where  $Z(f_{high})$  and  $Z(f_{low})$  are the impedances with the signal at high and low frequency separately. In the circuit system, the ratio of impedance was calibrated with the impedance obtained by the detection of signals through the root canal. Figure 3 shows the time domain diagram of an input signal and the signal after it had passed through the root canal.

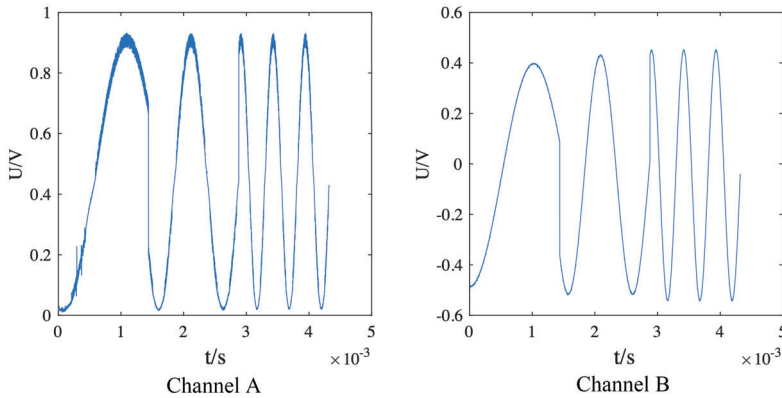


Figure 3. Time domain diagram of the acquisition signal.

The signal input to channel A then passed through the measured root canal and was detected in channel B. The frequency of the signal in Figure 3 exhibits a periodic variation with time (i.e., in the time domain). The energy of the signal in the frequency domain after carrying out a fast Fourier transform was uniformly distributed over this particular frequency. Other signals at different frequencies all exhibited the same time and frequency domain characteristics. It was verified that the circuit system could generate sine sweep signals meeting the measurement requirements. The output sine sweep signal was processed to deduce the impedance of the root canal with this signal at a particular frequency.

The circuit system emitted the sine sweep signals at different frequencies, corresponding to different impedance values at the frequency. Two impedance values were combined together to calculate the impedance ratio, which was used as the deep learning training data.

The goal of the multifrequency impedance method based on deep learning is to train an advanced neural network; therefore, the data had to be augmented before their substitution into the model. The impedance ratio related to the root canal length is the ratio of impedance with a high and a low frequency signal. The circuit can generate a current of frequencies from 100 Hz to 20 kHz, in which an arbitrary high frequency signal and an arbitrary low frequency signal were combined to obtain the impedance ratio, as shown in Figure 4.

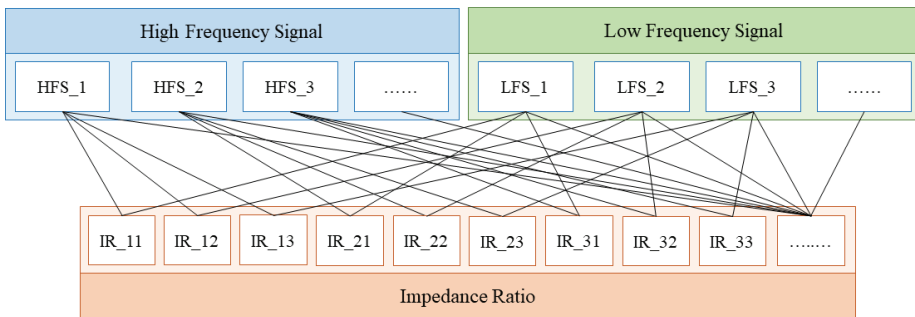


Figure 4. Diagram of the data augmentation.

The multifrequency method utilizes the signals of more frequencies than the traditional dual-frequency method. The impedance ratio data set was augmented by using arbitrary combinations of high and low frequency signals. An example of data augmentation is shown in Figure 5.

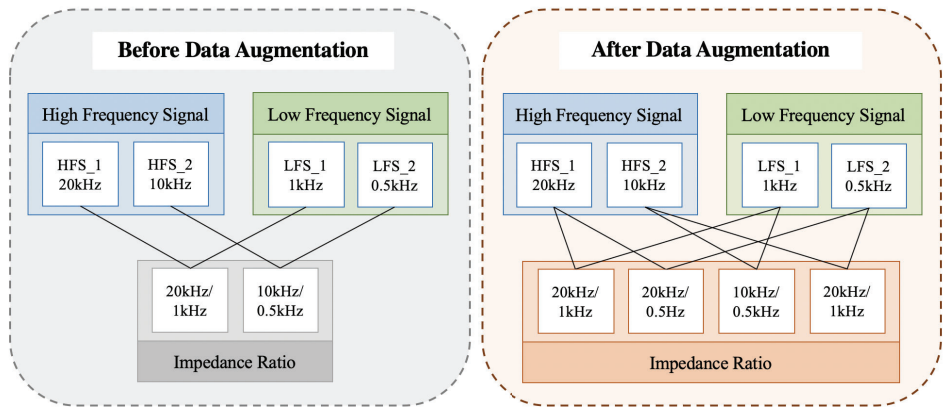


Figure 5. An example of data augmentation.

For the two high frequency signals at 20 kHz and 10 kHz and the two low frequency signals at 0.5 kHz and 1 kHz, the number of the impedance ratio can rise from two to four. It is obvious that the number of data can be considerably increased through data augmentation. Moreover, the combinations containing all the signals had an utmost use in the measurements in order to avoid exceptions. The accuracy of the proposed multifrequency impedance method based on deep learning could be improved significantly by training the neural network with the augmented data.

### 2.3. Feature Selection

The data were augmented using combinations of multiple frequencies. The accuracy of root canal length measurement could be increased after the sufficient impedance ratios were available. Features were selected according to the divergence and correlation between features and goals [20]. Filtering is one of the most commonly used methods for feature selection in deep learning in order to develop a rule to measure each feature and sort all features by their importance with respect to the target attribute. The first step was to calculate the variance of each feature, where features with variance below the threshold were deleted. The next step was to calculate the correlation coefficient of each remaining feature to the label. In addition, the features for the training model also required the numerical characteristics of the measuring conditions, such as tooth and file types. We used one-hot encoding to convert these non-numerical attributes into numerical features, as shown in Table 1.

Table 1. One-hot encoding for tooth types.

Tooth Type	Incisor	Canine	Molar
Sample 1	1	0	0
Sample 2	0	1	0
Sample 3	0	0	1

Among the large amount of augmented data (i.e., impedance ratios), based on the model performance and convenience of calculation, the top ten groups of impedance ratios—as ranked by the correlation coefficient—were selected (5 kHz/0.5 kHz, 8 kHz/0.5 kHz, 10 kHz/0.5 kHz, 12 kHz/0.5 kHz, 15 kHz/0.5 kHz, 20 kHz/0.5 kHz, 8 kHz/1 kHz, 10 kHz/1 kHz, 15 kHz/1 kHz, and 20 kHz/1 kHz), together with tooth type and file type, as the features for the model.

### 2.4. Neural Network Model

The neural network model used in this study is presented in Figure 6. The input layer takes the selected features as inputs, the output of the output layer is the distance between the file and apical foramen predicted by the model, and the hidden layer is used to enhance the nonlinearity of the model. The activation function used is the sigmoid function.

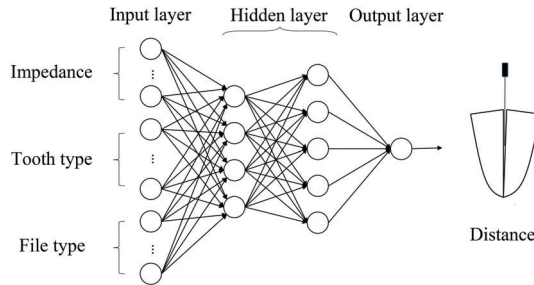


Figure 6. Schematic of neural network structure.

One of the Momentum, Adam [21], and SGD optimizers is selected as an optimum optimizer to accelerate the training of a neural network model. Considering that the small data set used increased the risk of overfitting, some noise was added to the training set during the training phase [22] and regularization was applied in the training process to enhance the generalization ability [23].

## 3. Results

### 3.1. Pre-Verification

Before the training of the neural network for root canal length measurement, the experimental results were discussed through the relationships between impedance values and different variables in order to determine that the theories about the impedance characteristics of teeth are correct, which can verify the feasibility of the method in this study.

#### 3.1.1. Impedance Verification

The impedance of the root canal is related to the tooth type, input signal frequency, and the distance from file to apical foramen. We took these three factors in turn as independent variables in order to observe their effects on the dependent variable of impedance.

First, some different tooth samples were selected to observe the root canal impedance.

Figure 7 illustrates the impedances of four kinds of selected teeth with several input signals. The selected teeth differed in terms of type and age, but the impedance trends of each tooth type were similar. The impedance of all the teeth decreased with the increase of frequency, thus guaranteeing that the impedance ratio method is reasonable.

For these four selected teeth, the relationships between the impedance of different teeth and the position of file are presented in Figure 8. The frequency was fixed.

It was found that the impedance changed smoothly when the file was far away from the apical foramen, while the impedance decreased rapidly when the file was close to the apical foramen. In particular, the impedance varied the most within 1 mm of the file from the apical foramen. Moreover, when the distance from apex was 0 mm (i.e., the file was at the position of the apical foramen), there was not much difference among the impedances of different tooth samples. The results in Figure 8 provide a theoretical basis for the multifrequency impedance ratio method for root canal length measurement.



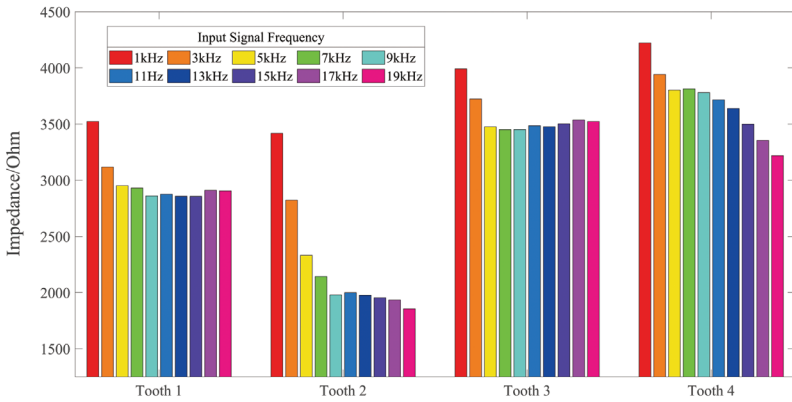


Figure 7. The impedance of different input signal frequencies varying with the tooth type.

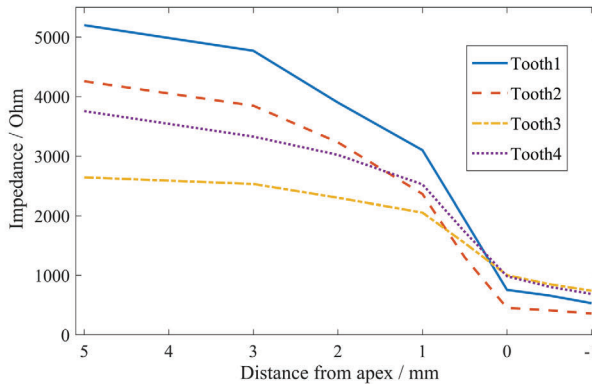
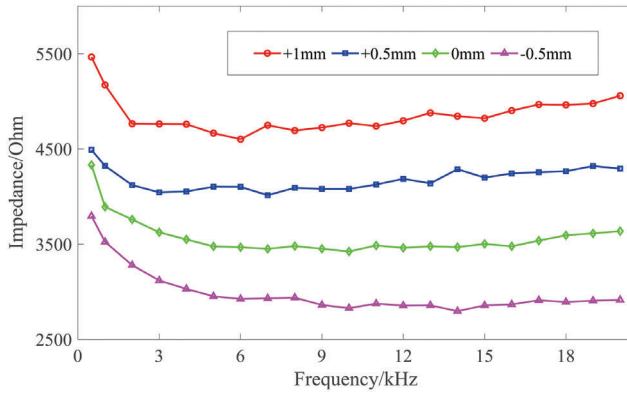


Figure 8. The impedance of different teeth varying with the distance between the file and the apical foramen.

In the next step, the tooth type was kept constant. The impedance of the file at various positions from the apical foramen varying with frequency was explored.

As shown in Figure 9, the impedance decreased as the frequency increased, regardless of the distance between the file and apical foramen. These phenomena verified the correctness of the early methods for root canal length measurement, as well as the reliability of the proposed measurement system.

When the frequency increased to approximately 20 kHz, the impedance almost stopped increasing. Therefore, setting the highest frequency as 20 kHz barely affects the subsequent analysis of experimental results when the impedance ratios of different frequency combinations were calculated. It was considered sufficient to conclude the experiments with the frequency ranging up to 20 kHz.

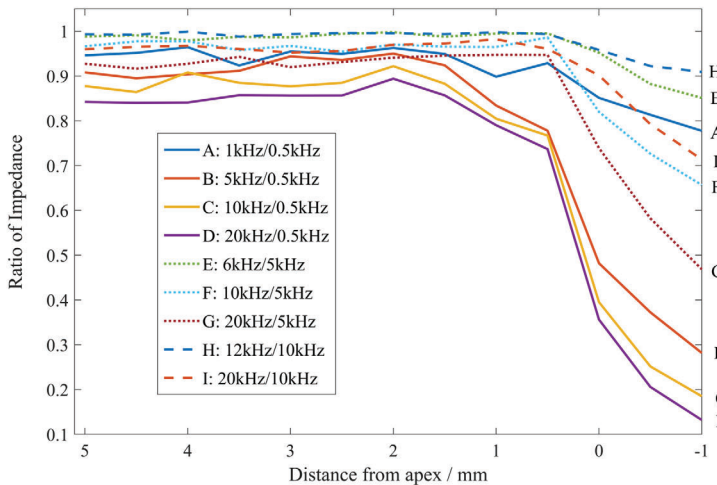


**Figure 9.** The impedance of the file at different position from apical foramen varying with frequency.

### 3.1.2. Frequency Ratio Verification

The impedance ratios were calculated with multiple combinations of frequencies in order to find out the optimal frequency ratio. It is possible to compare the performance in these subsequent experiments with the neural network-based multifrequency impedance method for root canal length measurement. In addition, frequency ratio verification provided a reference for the following feature selection and evaluation.

The impedance ratio results, according to the frequency combinations, are illustrated in Figure 10. The tooth type and the file were kept constant.



**Figure 10.** The impedance ratios of nine groups of frequency combinations varying with the distance between the file and the apical foramen (molar, #15).

In Figure 10, the impedance ratios did not change significantly when the file was far from the apical foramen. When it was close to the apical foramen (especially when the distance was less than 1 mm), the impedance ratio dropped rapidly, while the gradient was steepest at the apical foramen. Moreover, the larger the difference between the high and low frequencies, the more significantly the impedance decreased when the file was close to the apical foramen; in contrast, it was not beneficial

to determine the position of the file. The data for Figure 10 were obtained under the conditions of a molar tooth type and file #15; the curve trend was similar for other conditions.

The aforementioned phenomena correspond with feature selection. The impedance ratios selected as features with large variances were the impedance ratios of substantially divergent high and low frequencies.

### 3.2. Neural Network Training

The sample set was taken from numerous measurements of 21 teeth with multifrequency signals combinations. Due to the lack of tooth samples, leave-one-out (LOO) cross-validation was used to evaluate the performance of the neural network and to prevent overfitting [24]. For the 21 tooth samples, the LOO-based validation was performed with 21 iterations. In each iteration, the neural network was trained with the data set of 20 samples and tested on the remaining sample. According to the loss curve in Figure 11, a suitable optimizer was selected. The mean values of accuracy of the training and test sets after all iterations were calculated as indicators to assess the generalizability of the model. The performance of the neural network model with different structures using the LOO method is depicted in Figure 12. It should be noted that a large number of structures with different layers and nodes were verified, while a few of these results were selected to display. A highly effective model should possess both low bias and low variance. High performance on the training set reflects low bias but can cause overfitting, which suggests high variance. Therefore, point E represented the best structure.

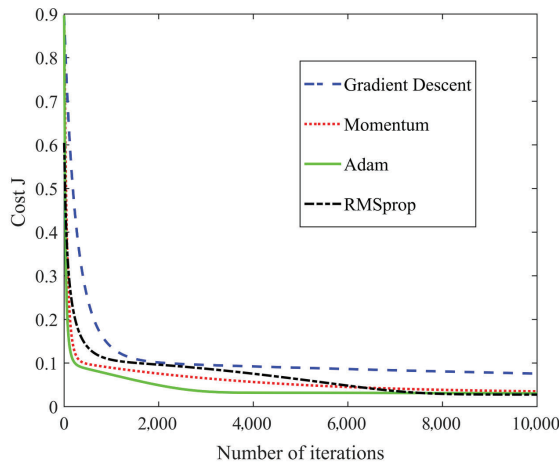


Figure 11. Loss curves with different optimizers.

As evident from Figures 11 and 12, the neural network model eventually developed had two hidden layers, where the numbers of nodes in the hidden layers were nine and five, respectively. A neural network with no more than three layers was sufficient for the uncomplicated data set of impedance, tooth type, and file type, while it was identified that the performance with three layers was not good enough (especially point K). More layers can make the training process more complicated and can prevent the model from converging, leading to overfitting. The optimization method used for training was Adam; furthermore, regularization was added.

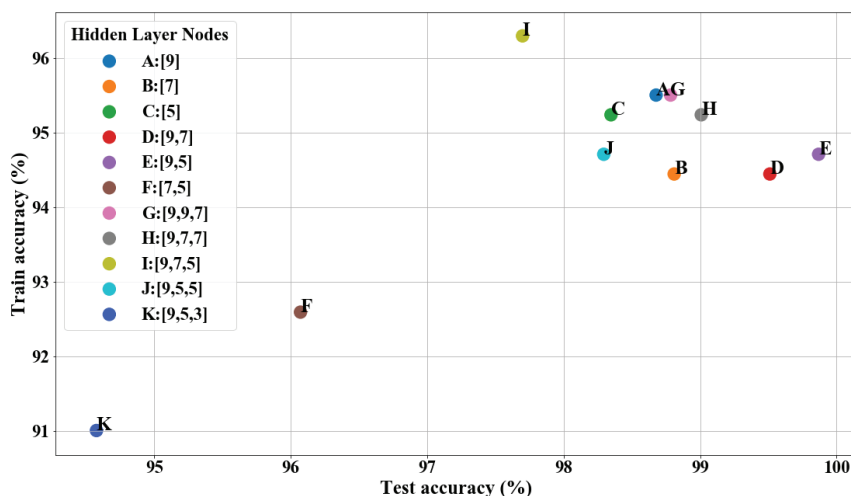


Figure 12. Performances of neural network models with different structures.

The 21 teeth were divided into three groups, according to the types of tooth and file: Group 1—molar, #15(6); Group 2—molar, #25(6); and Group 3—canine, #15(9). Table 2 presents a comparison of the performance of the dual-impedance ratio method and the neural network-based multifrequency method. Four pairs of frequencies were selected for the impedance ratios: 5 kHz/0.5 kHz, 10 kHz/0.5 kHz, 10 kHz/1 kHz, and 20 kHz/1 kHz.

For the dual-frequency impedance ratio method, using the average impedance ratio of 21 teeth to determine the apical foramen was not appropriate for Group 1. This was because there were three teeth in Group 1 for which the dual-frequency impedance ratio at the 0 mm position was quite different from the average impedance ratio. By contrast, the neural network-based multifrequency method could solve this problem, had a high accuracy rate, was less affected by changes in tooth and file type, and exhibited decent robustness.

Table 2. Performances of the impedance ratio method and neural network-based multifrequency method.

Frequency Combination (kHz)	Group 1	Group 2	Group 3	Total
5/0.5	66.67%	83.33%	100.00%	85.71%
10/0.5	50.00%	83.33%	100.00%	80.95%
10/1	50.00%	83.33%	89.89%	76.19%
Multifrequency	83.33%	100.00%	100.00%	95.24%

The experimental results indicated that the proposed measurement method is relatively robust and improved the effects of measuring factors on the results. The experiments in this study may not have been perfect, however. Further improvements in accuracy can be considered, based on the following aspects: improving the neural network structure, using different judgment strategies, using different optimization methods and, most importantly, expanding the data set.

### 3.3. Discussions Compared with EALs

The development of EALs has been a long and continual process, which seems to lag behind the rapid development of modern medical technology. After the impedance ratio method was established, only a few studies researched means by which to select two proper frequencies to enhance the performance of third-generation EALs [25–27]; however, no landmark improvement has emerged in the field of root canal length measurement until the present study. Although the measurements of

fourth-generation EALs seem to be more accurate, their actual performance does not present a great improvement, compared with the benchmark of the third-generation product—Root ZX—according to product comparison experiments [28,29]. In addition, in actual surveys, many dentists have provided feedback that they are more inclined to use third-generation products (e.g., Root ZX) as, in clinical settings, the third- and fourth-generation EALs differ little in accuracy; furthermore, the older models are more stable and cheaper. Root ZX is an excellent product, but it is not perfect; doctors often must use radiography as an aid to obtain accurate results when employing RootZX [30].

Obviously, much room still exists for improvement in the development of EALs. Machine learning is a popular subject, which has been widely used in the medical field and has promoted the rapid development of medical technology [31]. The method proposed in this paper combines the multifrequency impedance ratio method and neural networks. In fact, an EAL can be regarded as a prediction system or nonlinear regression model, considering that the measuring conditions have a critical influence on the results. Therefore, the impedance ratios remaining after feature selection and the numerical measuring factors can be used as the features. Neural networks possess great advantages in formulating such prediction models.

The experimental results indicated that the proposed measurement method is relatively robust and can improve the effects of measuring factors on the results. The experiments in this study may not have been perfect, however. Further improvements in accuracy can be considered based on the following aspects: improving the neural network structure, using different judgment strategies, using different optimization methods and, most importantly, expanding the data set.

#### 4. Conclusions

The method proposed in this paper combined the multifrequency impedance ratio method and neural networks. To increase the accuracy of the model, the impedance ratio data were augmented with different combinations of currents at various frequencies generated by the designed circuit system. The pre-verification of impedance was performed to provide theoretical support for training the neural network. Impedance, tooth type, and file type were selected as features in the model. Leave-one-out cross-validation was used during the training process due to the limited tooth samples. An optimal neural network was determined according to the performances of neural network models with different structures. Compared with the dual-frequency impedance ratio method, the proposed approach can reduce the influence of measuring factors on the measurement results, increase the measurement accuracy, and enhance the robustness.

**Author Contributions:** Conceptualization, X.Q., Z.Z., and X.C.; methodology, Z.Z. and X.C.; software, Z.Z.; validation, Z.Z.; formal analysis, X.Q. and Z.Z.; investigation, X.Q.; resources, X.C.; data curation, X.Q.; writing—original draft preparation, Z.Z.; writing—review and editing, X.Q.; supervision, X.C.; project administration, X.C.; funding acquisition, X.C. All authors have read and agreed to the published version of the manuscript.

**Funding:** This research was funded by the National Key Research and Development Program of China (No. 2019YFF0216401) and the Major projects of Science and Technology Commission of Shanghai (No. 17JC1400800).

**Conflicts of Interest:** The authors declare no conflict of interest.

#### References

1. Gordon, M.; Chandler, N.T. Electronic apex locators. *Int. Endod. J.* **2004**, *37*, 425–437.
2. Minetti, E.; Palermo, A.; Ferrante, F.; Schmitz, J.H.; Lung Ho, H.K.; Hann, D.; Ng, S.; Giacometti, E.; Gambardella, U.; Contessi, M.; et al. Autologous Tooth Graft after Endodontical Treated Used for Socket Preservation: A Multicenter Clinical Study. *Appl. Sci.* **2019**, *9*, 5396.
3. Razumova, S.; Brago, A.; Howijieh, A.; Barakat, H.; Kozlova, Y.; Baykulova, M. Evaluation of Cross-Sectional Root Canal Shape and Presentation of New Classification of Its Changes Using Cone-Beam Computed Tomography Scanning. *Appl. Sci.* **2020**, *10*, 4495.

4. Lee, J.; Lee, S.H.; Hong, J.R.; Kum K.Y.; Oh, S.; Al-Ghamdi, A.S.; Al-Ghamdi, F.A.; Mandorah, A.O.; Jang, J.H.; Chang, S.W. Three-Dimensional Analysis of Root Anatomy and Root Canal Curvature in Mandibular Incisors Using Micro-Computed Tomography with Novel Software. *Appl. Sci.* **2020**, *10*, 4385.
5. Yildirim, C.; Aktan, A.M.; Karataslioglu, E.; Aksoy, F.; Isman, O.; Culha, E. Performance of the Working Length Determination using Cone Beam Computed Tomography, Radiography and Electronic Apex Locator, in Comparisons to Actual Length. *Iran. J. Radiol.* **2017**, *14*, 1.
6. Marjanović, T.; Lacković, I.; Stare, Z. Comparison of electrical equivalent circuits of human tooth used for measuring the root canal length. *Automatika* **2011**, *52*, 39–48.
7. Meredith, N.; Gulabivala, K. Electrical impedance measurements of root canal length. *Dent. Traumatol.* **1997**, *13*, 126–131.
8. Ushiyama, J. New principle and method for measuring the root canal length. *J. Endod.* **1983**, *9*, 97–104.
9. Kobayashi, C. Electronic canal length measurement. *Oral Surg. Oral Med. Oral Pathol. Oral Radiol. Endod.* **1995**, *79*, 226–231.
10. Kobayashi, C.; Suda, H. New electronic canal measuring device based on the ratio method. *J. Endod.* **1994**, *20*, 111–114.
11. Ali, R.; Okechukwu, N.C.; Brunton, P.; Nattress, B. An overview of electronic apex locators: Part 2. *Br. Dent. J.* **2013**, *214*, 227–231.
12. Üstün, Y.; Aslan, T.; Şekerci, A.E.; Sağsen, B. Evaluation of the reliability of cone-beam computed tomography scanning and electronic apex locator measurements in working length determination of teeth with large periapical lesions. *J. Endod.* **2016**, *42*, 1334–1337.
13. Nekoofar, M.H.; Ghandi, M.M.; Hayes, S.J.; Dummer, P.M. The fundamental operating principles of electronic root canal length measurement devices. *Int. Endod. J.* **2016**, *37*, 595–609.
14. Stober, E.K.; Duran-Sindreu, F.; Mercade, M.; Vera, J.; Bueno, R.; Roig, M. An evaluation of root ZX and iPex apex locators: an in vivo study. *J. Endod.* **2011**, *37*, 608–610.
15. Welk, A.R.; Baumgartner, J.C.; Marshall, J.G. An in vivo comparison of two frequency-based electronic apex locators. *J. Endod.* **2003**, *29*, 497–500.
16. Specht, D.F. A general regression neural network. *IEEE Trans. Neural Netw.* **1991**, *2*, 568–576.
17. Goldberg, Y. Neural network methods for natural language processing. *Synth. Lect. Hum. Lang. Technol.* **2017**, *10*, 1–309.
18. Duan, F.; Dai, L. Recognizing the gradual changes in sEMG characteristics based on incremental learning of wavelet neural network ensemble. *IEEE Trans. Ind. Electron.* **2017**, *64*, 4276–4286.
19. Zhang, Z.; Duan, F.; Sole-Casals, J.; Dinares-Ferran, J.; Cichocki, A.; Yang, Z.; Sun, Z. A novel deep learning approach with data augmentation to classify motor imagery signals. *IEEE Access* **2019**, *7*, 15945–15954.
20. Guyon, I.; Elisseeff, A. An introduction to variable and feature selection. *J. Mach. Learn. Res.* **2003**, *3*, 1157–1182.
21. Ruder, S. An overview of gradient descent optimization algorithms. *arXiv* **2016**, arXiv:1609.04747.
22. Tanner, M.A.; Wong, W.H. The calculation of posterior distributions by data augmentation. *J. Am. Stat. Assoc.* **1987**, *82*, 528–540.
23. Zou, H.; Hastie, T. Regularization and variable selection via the elastic net. *J. R. Stat. Soc. Ser. B Stat. Methodol.* **2015**, *67*, 301–320.
24. Wong, T.T. Performance evaluation of classification algorithms by k-fold and leave-one-out cross validation. *Pattern Recognit.* **2015**, *48*, 2839–2846.
25. Jan, J.; Krizaj, D. Accuracy of root canal length determination with the impedance ratio method. *Int. Endod. J.* **2009**, *42*, 819–826.
26. Kim, D.W.; Nam, K.C.; Lee, S.J. Development of a frequency-dependent-type apex locator with automatic compensation. *Crit. Rev.™ Biomed. Eng.* **2000**, *28*, 473–479 doi:10.1615/critrevbiomedeng.v28.i34.200.
27. Nam, K.C.; Kim, S.C.; Lee, S.J.; Kim, Y.J.; Kim, N.G.; Kim, D.W. Root canal length measurement in teeth with electrolyte compensation. *Med Biol. Eng. Comput.* **2002**, *40*, 200.
28. Martins, J.N.; Marques, D.; Mata, A.; Carames, J. Clinical efficacy of electronic apex locators: Systematic review. *J. Endod.* **2014**, *40*, 759–777.
29. Vasconcelos, B.C.; Bueno Mde, M.; Luna-Cruz, S.M.; Duarte, M.A.; Fernandes, C.A. Accuracy of five electronic foramen locators with different operating systems: An ex vivo study. *J. Appl. Oral Sci.* **2013**, *21*, 132–137.

30. Nawab, S.; Rana, M.J.A.; Yar, A. Comparative evaluation of working length with digital radiography and third generation electronic apex locator. *Pak. Oral Dent. J.* **2016**, *36*, 308–311.
31. Deo, R.C. Machine Learning in Medicine. *Circulation* **2015**, *132*, 1920–1930.

**Publisher’s Note:** MDPI stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.



© 2020 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<http://creativecommons.org/licenses/by/4.0/>).





Article

# The INSESS-COVID19 Project. Evaluating the Impact of the COVID19 in Social Vulnerability While Preserving Privacy of Participants from Minority Subpopulations

Karina Gibert \* and Xavier Angerri

Intelligent Data Science and Artificial Intelligence Research Center and Institut de Ciència i Tecnologia de la Sostenibilitat, Universitat Politècnica de Catalunya—BarcelonaTech, 08001 Barcelona, Spain; xavier.angerri@upc.edu

\* Correspondence: karina.gibert@upc.edu

**Featured Application:** The results of this research have been delivered to the General Director of Social Services and the General Director of Equity from the Catalan Government and will be used to support the new policies and strategies on Social Services in the post-COVID19 period.

**Abstract:** In this paper, the results of the project INSESS-COVID19 are presented, as part of a special call owing to help in the COVID19 crisis in Catalonia. The technological infrastructure and methodology developed in this project allows the quick screening of a territory for a quick a reliable diagnosis in front of an unexpected situation by providing relevant decisional information to support informed decision-making and strategy and policy design. One of the challenges of the project was to extract valuable information from direct participatory processes where specific target profiles of citizens are consulted and to distribute the participation along the whole territory. Having a lot of variables with a moderate number of citizens involved (in this case about 1000) implies the risk of violating statistical secrecy when multivariate relationships are analyzed, thus putting in risk the anonymity of the participants as well as their safety when vulnerable populations are involved, as is the case of INSESS-COVID19. In this paper, the entire data-driven methodology developed in the project is presented and the dealing of the small subgroups of population for statistical secrecy preserving described. The methodology is reusable with any other underlying questionnaire as the data science and reporting parts are totally automatized.

**Keywords:** data science; intelligent decision support; social vulnerability; gender-gap; digital-gap; COVID19; policy-making support



**Citation:** Gibert, K.; Angerri, X. The INSESS-COVID19 Project. Evaluating the Impact of the COVID19 in Social Vulnerability While Preserving Privacy of Participants from Minority Subpopulations. *Appl. Sci.* **2021**, *11*, 3110. <https://doi.org/10.3390/app11073110>

Academic Editor: Jordi Solé-Casals

Received: 18 January 2021

Accepted: 9 March 2021

Published: 31 March 2021

**Publisher's Note:** MDPI stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.



**Copyright:** © 2021 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

## 1. Introduction

The consequences of the crisis caused by COVID19 have been devastating from a sanitary point of view, but they will presumably be also devastating from an economic and social point of view. The COVID19 generated a situation never seen before and at the time of starting our research, in April 2020, we were convinced that new social needs would emerge, and it would be urgent to identify them as soon as possible to properly address them.

Most of the research done in the field of COVID19 is focusing on the prediction of the infection rates in the population, survival rates, propagation of the disease, or diagnosis, like in [1]; indeed, most of the research in COVID19 topics is done under a health approach. However, the project INSESS-COVID19 was born with the aim to focus on Social Services, largely forgotten in the management of the pandemics, although being a field with a strong need of including data as an asset for management and improvement of the Social Services system itself as well as for improvement of services to citizens.

The INSESS-COVID19 project (namely Identification of Emerging Social Needs as a consequence of COVID19 and effect on the Social Services of the territory), is one of the 21 proposals funded by the Special Call on COVID19 Research launched in April 2020 by the Centre for Cooperation in Development of the Universitat Politècnica de Catalunya. INSESS-COVID19 is a prospective study to identify the social vulnerabilities of the Catalan population and to provide elements to support decision-making to the 107 Basic Areas of Social Services (BASS) of Catalonia and to the Social Services Department from the Catalan government. The BASS will have to face all these new vulnerabilities and require decision support tools to be able to manage the incoming overflow.

INSESS-COVID19 uses an innovative approach based on mechanisms for rapid data collection from an entire territory, based on participatory processes where citizens and experts in social services can contribute at different levels. The project uses a mixed methodology that combines data science techniques, knowledge management and artificial intelligence, which has allowed contributing to provide data/knowledge-driven outcomes useful to policymaking in the matter of Social Services in Catalonia [2]. The technological tool developed in INSESS-COVID19 proves the feasibility of quickly getting data direct from citizens and making a rapid diagnosis of territory whenever needed. The methodology proposed in the project, and the technology developed to implement it is general, being as well valid, not only in Social Services, but in any governmental or business area. The INSESS-COVID19 proposal allows overcoming the limitations of the most classic information systems in relation to decision support [3] need in front of an unexpected situation, as it allows to obtain direct information from the source (citizens in our case) whenever required, even if the ordinary information system do not contain it.

According to the GDPR (General Data Protection Regulation) [4] GDPR law, privacy of citizens participating in the project must be guaranteed and is critical to create the trustworthy climate that allows citizens to openly confess their vulnerabilities in a protective way, with the certainty that disclosing vulnerabilities (like being illegally in the country for example) would not have any direct consequence for him/her. Provided that the data collection process regards a big number of variables (195), there are 945 BASS organized in eight administrative bigger areas (named Vegueries), and grouped by four provinces, the risk of getting very small groups of citizens following a certain pattern of vulnerability is high, this raising the risk of violation of the anonymization principle in the practical application of the proposal.

The main challenge is to extract as much relevant decisional information from the dataset by preserving overall the privacy of the participant citizens. Privacy and human rights oversight are two of the main principles recommended in the guidelines for Ethic in AI provided by the European Commission in May 2019. Thus, this paper provides a methodological proposal to guarantee the participant privacy in the publication of results. Considering that this project includes some vulnerable citizen's profiles, like illegal foreigners, victims of domestic violence, or mental health patients, the preservation of privacy of all participants is crucial for their safety.

The project is a close collaboration between Intelligent Data Science and Artificial Intelligence research center at UPC and the iSocial Foundation, being Karina Gibert (IDEAI-UPC) and Toni Codina (iSocial) the main researchers of the project. When this study started by last May 2020, the general expectation was that pandemics lockdown would finish by July and de-escalation would start then, so that we could focus on analyzing the collected data and contributing to build this new normality mentioned everywhere. Nothing further from reality. The pandemic is still among us nowadays, as is the state of alarm, and the situation is still far from stabilizing. The new outbreaks from last July had a strong impact on the project work plan. The collapsed Social Services were not able to be involved in research projects, of course, and the organization of the face-to-face workshops originally planned in the project became unfeasible with the containment measures again enacted. The INSESS-COVID19 team put all their energy into rethinking the design of the data collection process, in order to enable citizen participation, at minimum cost for the BASS.

The data collection period, originally planned by June and half July was extended as much as possible, until last 6 December 2020. The data analyzed in this paper were collected between end June and 6 December 2020. Four months throughout entire Catalonia, with the invaluable collaboration of an important part of the 107 BASS, where social services professional staff made the effort of finding moments to collaborate with the project and contacting the participant citizens, in spite of being in a very complex overflow situation.

In the next sections, the different elements developed in the project are presented, as well as the methodological proposal to deal with small data. Real results from the questionnaire and some results resulting from the automatic analysis are shown.

## 2. Materials and Methods

### 2.1. State of the Art

Before building the INSESS-COVID19 instrument, different related studies had been consulted. In Table 1, some of the works are listed.

**Table 1.** State of the Art main references.

Title	Promoter Entity	Link
Survey on the impact of COVID-19. 2020	CEO (Centre d'Estudis d'Opinió)	<a href="http://ceo.gencat.cat/ca/estudis/registre-estudis-dopinio/estudis-dopinio-ceo/societat/detall/index.html?id=7588">http://ceo.gencat.cat/ca/estudis/registre-estudis-dopinio/estudis-dopinio-ceo/societat/detall/index.html?id=7588</a> (accessed on 31 March 2021)
Survey on time uses in lockdown. 2020	CEO	<a href="http://ceo.gencat.cat/ca/estudis/registre-estudis-dopinio/estudis-dopinio-ceo/societat/detall/index.html?id=7608">http://ceo.gencat.cat/ca/estudis/registre-estudis-dopinio/estudis-dopinio-ceo/societat/detall/index.html?id=7608</a> (accessed on 31 March 2021)
Special Barometer May 2020	CIS (Centro de Investigaciones Sociológicas)	<a href="http://www.cis.es/cis/opencms/ES/NoticiasNovedades/InfoCIS/2020/Documentacion_3281.html">http://www.cis.es/cis/opencms/ES/NoticiasNovedades/InfoCIS/2020/Documentacion_3281.html</a> (accessed on 31 March 2021)
Covid 19 Impact Survey	Dr. Nuria Oliver, commissioned for AI and COVID-19. Generalitat Valenciana	<a href="https://covid19impactsurvey.org/">https://covid19impactsurvey.org/</a> (accessed on 31 March 2021)
Gestioemocional.cat	Health department, Generalitat de Catalunya.	<a href="https://gestioemocional.catsalut.cat/">https://gestioemocional.catsalut.cat/</a> (accessed on 31 March 2021)
Social Service Survey	ACM (Associació Catalana de Municipis)	<a href="https://docs.google.com/forms/d/e/1FAIpQLSe7MBgTSeA4NtfWlzWM3yDtdsVUXHX118r-FvwHXLimgvKVCA/viewform">https://docs.google.com/forms/d/e/1FAIpQLSe7MBgTSeA4NtfWlzWM3yDtdsVUXHX118r-FvwHXLimgvKVCA/viewform</a> (accessed on 31 March 2021)
Encuesta Condiciones de Vida	INE (National Statistics Institute)	<a href="https://www.ine.es/dyngs/INEbase/es/operacion.htm?c=Estadistica_C&amp;cid=1254736176807&amp;menu=ultiDatos&amp;idp=1254735976608">https://www.ine.es/dyngs/INEbase/es/operacion.htm?c=Estadistica_C&amp;cid=1254736176807&amp;menu=ultiDatos&amp;idp=1254735976608</a> (accessed on 31 March 2021)

### 2.2. INSESS-COVID19 Methodology

The project proposes an innovative methodology to reach the goals. The novelty regards three different issues:

- The technological solution provided to collect data from citizens in short time
- The methodological solution for automatic analysis of collected data
- The methodological solution proposed for reaching the citizens to involve in the analysis

The main steps of the proposal are listed below. In the next subsections, details on each step are provided and the novelty highlighted where it is.

#### A. Analysis of the phenomenon and design of observation tools

Before starting with the technical part of data management, the proposed methodology suggests starting by understanding the structure of the target ecosystem. From this analysis, a clear idea about the sample design will appear on the one hand, and the kind of questions required from participants as well. In addition, the ways in which data will be collected require attention.

- a. Analysis of the target ecosystem
  - b. Identification of target subpopulations and profiles
  - c. Robustness with regards to the moment of answering
  - d. Construction of the impact-oriented questionnaire
  - e. Design of technological infrastructure
  - f. Workshop design 3
- B. Collection of territorial information and data analysis 5

The proposed sequence of steps to perform the analysis is inspired in the traditional KDD procedure (KDD: Knowledge discovery from data). In our case, we introduce a specific proposal for the operativization of the very last step of Knowledge Production proposed by Fayyad [5], from which a significant lack of literature exists even nowadays, and which is aligned with the emergent field of Explainable AI [6].

1. Data collection methodology
2. Data pre-processing
3. Descriptive and territorial analysis
  - 3.1. Multivariate variables
  - 3.2. Temporal variables
  - 3.3. Open questions analysis through Natural language processing methods
4. Intelligent Multivariate Analysis
5. Pattern identification across BASS
6. Artificial Intelligence-based Conceptualization

As it will be seen along the paper, the questionnaire includes variables with complex structures and some of them express along several columns or not in the DB. Dealing with this situation requires the development of some new methodological components that will be detailed along the paper.

1. Definition of a typology of complex variables
  2. Design of automatic analysis procedures 10
  3. Design of specific visual and analytical tools for complex type of variables
  4. Statistical secrecy preservation 8
  5. Privacy issues 7
  6. Treatment of the statistical error 6
  7. Metainformation model 11
  8. Automatic reporting 12
  9. Implementation
- C. Results interpretation, diagnostics and final recommendations

In the following, details on all steps are provided.

### 2.3. Analysis of the Target Ecosystem

We propose that this part includes three aspects:

1. Understanding of the structure of the target domain. For the case of INSESS-COVID19, this requires understanding the structure of Social Services in Catalonia. How are they organized, which public administrations have competences in the different kind of services, the set of available services and so on.
2. Also, a review of current sources of official statistics about the target domain that can be used as a reference for the analysis is required. Literature review is useful. Not in the academic sense, but finding official reports describing the target domain (in this case, official statistics and surveys followed during pandemics)

This analysis, conducted together with the domain experts, will result in a clear identification of the kind of participants required and the kind of information required from them and will provide the inputs for the decisions taken in next steps.

#### 2.4. Identification of Target Subpopulations and Profiles 4

As said before, after a deep understanding of the list of available Social Services offered in primary social care system, a list of 20 target profiles and the corresponding inclusion criteria were defined together with the Social Services professionals, from both government, city councils and regional councils (consells comarcals). The proposed profiles point out to segments of population a priori expected to be significantly damaged by the pandemics:

1. Single-parent families
2. Young people unemployed
3. Unemployed over 50 years old
4. Citizens coming from abroad in irregular situations
5. Ex-tutelage people and underage alone
6. Poor workers (very low salaries)
7. Poor workers (temporal and discontinuous permanence)
8. Poor workers (submerged economy)
9. People under EERTO or dismissed
10. Autonomous workers and small entrepreneurs under bankruptcy
11. Dependent elderly
12. Elderly leaving alone
13. People with disability (physical, sensorial or emotional)
14. Informal care-givers
15. People with mental disorders
16. People from LGTBI community under vulnerability
17. People with addictions to alcohol, drugs or conducts
18. Women victims of male violence
19. People without home or leaving in infra-homes
20. Professionals from Essential Social Services and Health

#### 2.5. Construction of the Impact-Oriented Questionnaire

After an extensive analysis of the conceptual framework, a conceptualization of the target areas of life to be studied was agreed with the experts. Among all the instruments, surveys and reports analyzed, the reference conceptual model was the SSM.cat model [7], an instrument to compute the social vulnerability adopted by the Catalan government to be part of the new Social Services system (e-Social), planned as the kernel of the digital transformation of Social Services targeted in the Strategic Plan of Social Services of Catalonia [8] and very much aligned with the current structure of primary care Social Services in Catalonia. The process by which this reference model was selected is new as it is based on a systematic review of the State of the Art, including the elaboration of a taxonomy of indicators, grouped by themes, and the description of the reviewed surveys in terms of the number of variables (and topics) related to every theme, the expert-based evaluation of the utility of these questions regarding the goals of the study, and the design of the thematic blocs and sequence according to that.

The SSM.cat model was inspired by the Dutch version of the Self Sufficiency Matrix model, developed by the University of Amsterdam [9], which in turn is an adaptation from the original Self Sufficiency Matrix developed by Diana Pearce for Wider Opportunities for Women as part of the State Organizing Project for Family Economic Self-Sufficiency [10,11].

Inspired in SSM.cat, INSESS-COVID19 assesses social vulnerability from the following 11 areas of daily life:

- Incomes
- Daily activities
- Home
- Domestic relationships
- Mental health
- Physical health

- Substance abuse
- Daily activities skills
- Social network
- Community participation
- Legal framework

The INSESS-COVID19 questionnaire has been developed by focusing questions on these areas. Each area can contain a different number of questions, mainly oriented to bring to the fore not only social vulnerability but also the impact of the COVID19 in this vulnerability. The result is a questionnaire with 21 blocks that generate up to 195 items, of different structures, according to the type of questions. Figure 1 shows the global structure of the survey.

#### 2.6. Validation of Questionnaire and Profiles

The questionnaire and sample design outcoming from the first phase of the methodology are extensively validated through several rounds of experts.

1. Two experts of the Advisory board of the project specialized in innovation for Social Services analyzed both que list of questions and the set of target profiles defined and provided positive feedback and some suggestion to improve writing to reduce ambiguities
2. The updated versions of both target profiles and questionnaire were submitted to the Social Services Commission of the Catalan Federation of Municipalities and a workshop was celebrated to evaluate the proposal that was fully accepted by the experts
3. Technical staff of the Social Services Department of the Catalan Government also reviewed the materials with successful feedback
4. Practitioners on Social Services checked the materials with a positive feedback as well and small amendments on ambiguity of the writings

None of them detected any missing profile in the sample design or question in the questionnaire and some highlighted the interest of some profiles or questions appeared as a consequence of the systematic review proposed in the paper that would have not been included from a more traditional expert-based approach (like Delphi or focus-groups).

#### 2.7. Robustness of Data Collection Moment by Design

The INSESS-COVID10 instrument introduces an innovative structure in the questionnaire, intended to allow a long period of data collection while preserving the comparability of the data collected. This is a very relevant characteristic of the questionnaire that allows extensions of the data collection period in such a way that keeps the property of considering data together for the analysis. This provides an important advantage in front of small samples, as providing longer period for data collection valid sample can increase without limit in the validity of previously collected data.

The proposal made in our work is that all questions from the questionnaire are divided in two categories:

- Static: Characteristic that keeps static along the entire study period (age, sex, place of birth, etc.)
- Dynamic: Characteristic that might change value along the study period

The proposal is to require answers in some fixed moments along time for all Dynamic questions in the questionnaire (Figure 2). The methodology is general, but for the case of INSESS-COVID19, it was decided to fix three moments of inquire: Pre-pandemics (January 2020), post-pandemics (July 2020) and expectations for the future (January 2021).



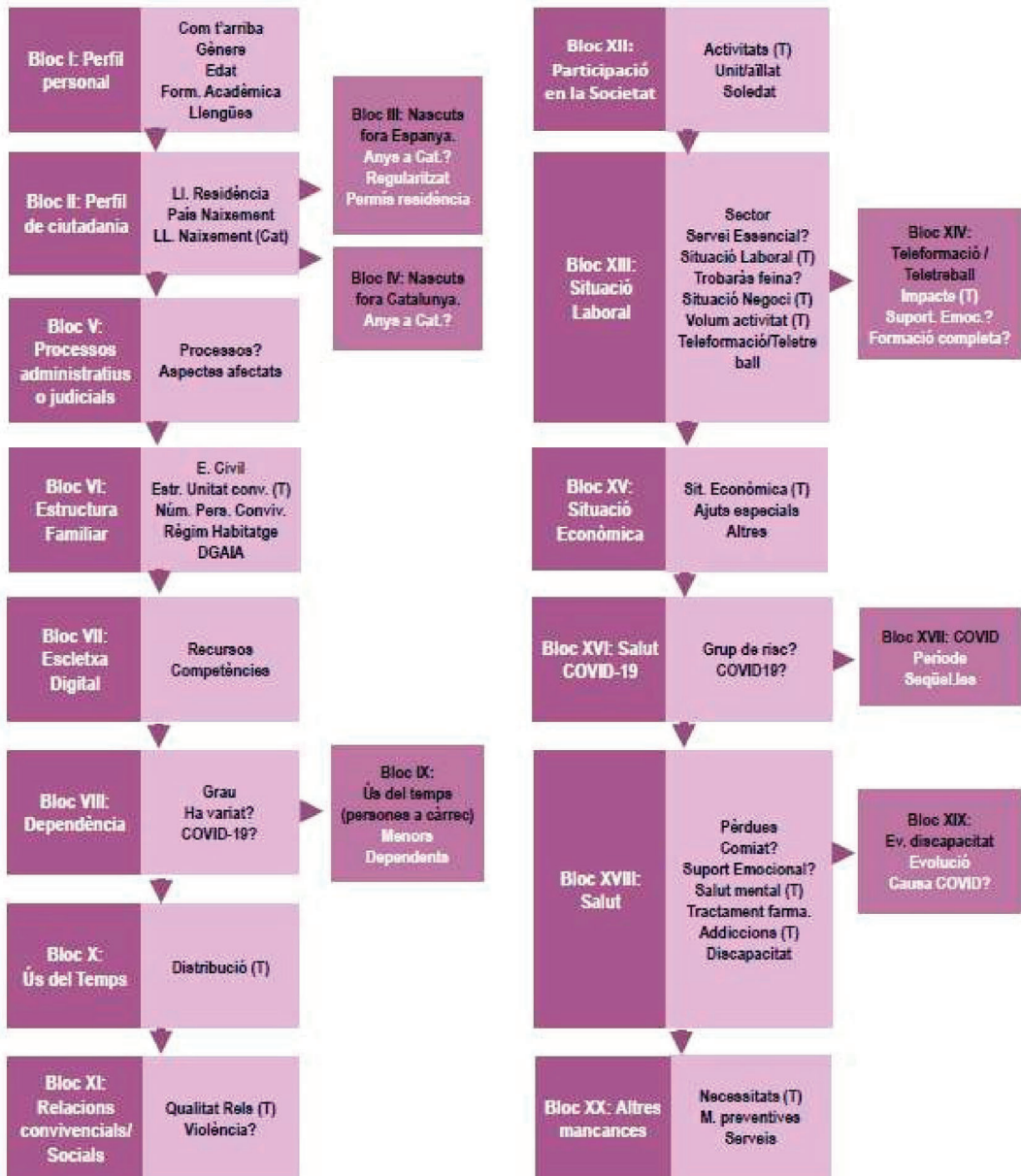
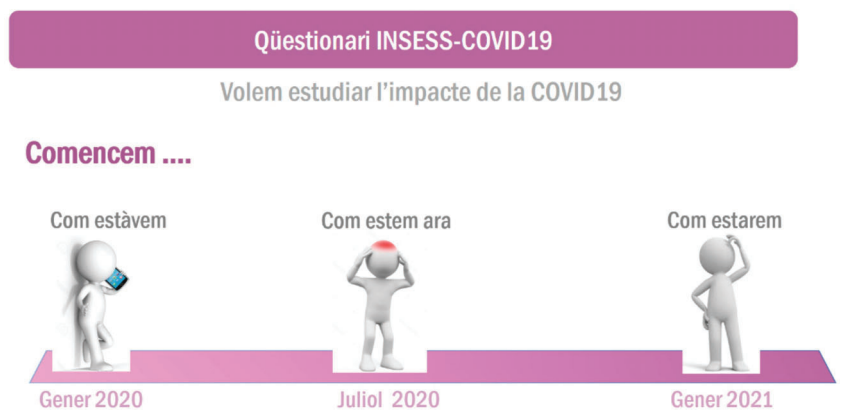


Figure 1. Structure of the INSESS-COVID19 questionnaire.

Introducing this design in the questionnaire has the property that the study gains robustness with respect to the specific date in which the citizen is participating into the project. The questionnaire asks about situations/perceptions in this three fixed time points, so the data collection process can be as long as required and data still permits the analysis of the dynamics of the phenomenon. Answers of persons participating in July, or August, or October, still provide information about the situation of the person in January 2020, July 2020 and January 2021, so they can be analyzed together. Considering the critical situation of the BASS during the 1st wave of pandemics, this solution is overcoming the limitation that most of the territory could not dedicate time to the study in June–July, and without introducing this kind of design the viability of the project would not survive.

The impact of the 1st wave of the pandemics becomes measurable through the differences between July and January 2020.

The consequence is that this design introduces packs of variables in the questionnaire, which are not anymore independent, and specific procedures to analyze them in the correct way will be required. These are introduced later in the paper.



**Figure 2.** The three fixed time-stamps of the INSESS-COVID19 questionnaire.

### 2.8. Technological Infrastructure

The methodology includes the design of the technological infrastructure that allows easy and secure access to the questionnaire to the citizens that will participate into the study and provides the pipeline to generate the final automatic reports based on the data collected in these questionnaires. Figure 3 displays the overview. A server in the cloud compliant with all GDPR is hosting the digital questionnaire. The access to the questionnaire is made through a website that requires authentication and it can be reached with either a cell phone, tablet or PC (personal computer). Data collected in the questionnaire is downloaded (even periodically) to be automatically processed through R and KCLASS [12] scripts and a well edited working report is automatically produced in Word, where the results of the analysis are displayed and formatted as a final document. The web is also hosting a view for the BASS staff with support documents to organize the workshops.

After implementation and deployment, technical validation of scripts, server performance, web functionality, and availability of all required materials was performed.

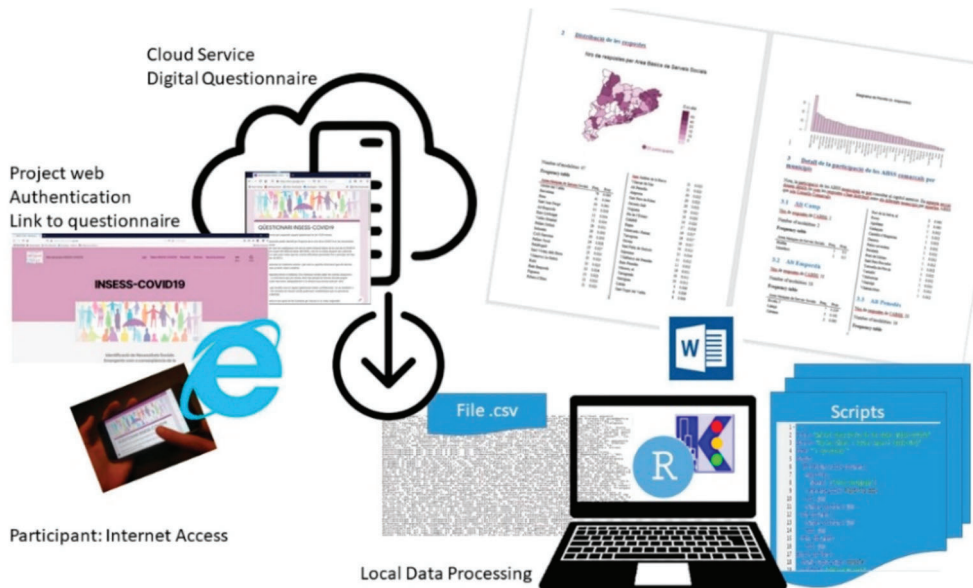


Figure 3. Technological infrastructure of INSESS-COVID19 system.

### 2.9. The INSESS-COVID19 Workshops

Originally, the project planned face-to-face workshops with the citizenship and part of it consisted in filling-in the INSESS-COVID19 questionnaire. The main advantages of such a design are:

- Digital gap issues of vulnerable population are assisted during the workshop and successful participation is guaranteed
- The data collection time is constrained. In 2 h workshop all answers for a given BASS are collected
- Missing data is reduced.
- Misunderstanding of questions is reduced, or eliminated

The main limitation of this design is to require coincidence in time and space and requires logistics and specific rooms offered by the BASS for the workshop celebration.

With the circumstances of the prolonged pandemics in successive outbreaks, an on-line, delocalized in time and space, version of the workshop was activated. The contextualization of the activity was pre-recorded in videos, uploaded to web, so that each participant has to enter the web, follow the videos (10 min) and answer the questionnaire, all available in a private web area. In this modality, the properties of the workshop are:

- No need to offer a specific room to celebrate the workshop
- No need to fix a day and time to celebrate the workshop
- Time for data collection needs to be a longer period. Extra follow-up of participants is required to guarantee the delivery of questionnaires on time
- Missing data can increase
- Misunderstanding of questions is still reduced through videos, but no interaction is available, so it might be not totally eliminated
- Specific long-term human support is required to solve the digital gap issues of vulnerable population. The BASS has to offer a person to this purpose.

The main aim of the mini-videos is to guarantee that all participants have the same understanding of the questions and know the main goals of the project, thus still helping to reduce both the misinterpretations of the questions.

The project considered four workshop modalities (Figure 4):

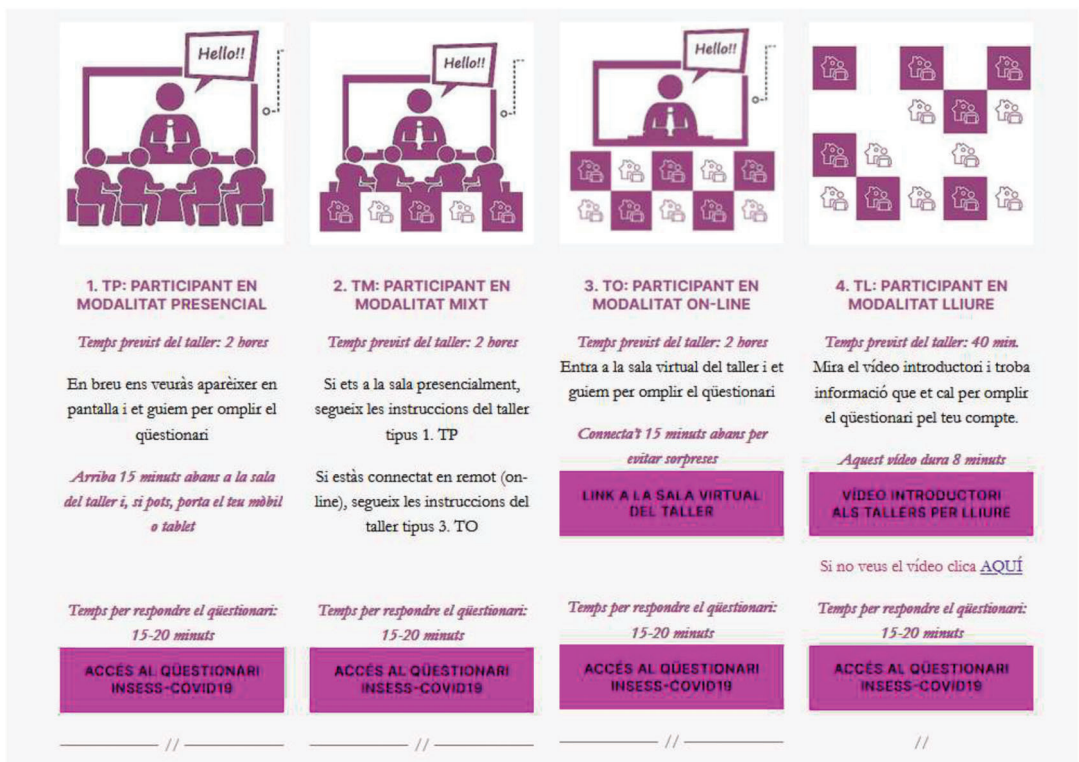


Figure 4. Workshop modalities.

- INSESS-COVID19 Face-to-face Workshop (Figure 5a): All people who are participating in the workshop meet together in a site provided by the BASS. INSESS-COVID19 team join through videoconference to lead the workshop.
- INSESS-COVID19 Mixt Workshop (Figure 5b): Some participants are located in the room assigned by the BASS for workshop. Also, some participants join the workshop through videoconference. The INSESS-COVID19 team join through videoconference.
- INSESS-COVID19 Online Workshop (Figure 6a): All participants join the workshop through videoconference. INSESS-COVID19 team join by means of videoconference to lead the workshop.
- INSESS-COVID19 Free Workshop (Figure 6b): All participants do the workshop online in its own schedule and from where they prefer (home, working place, etc.). The BASS professionals send to all participants the accesses to the website and videos of the project, to be seen before starting the workshop. The workshop consists on watching the proposed videos where the project is presented and after that, answering the questionnaire.

Other modalities: Along the data collection process, some BASS used creative mechanisms to involve citizenship into the study:

- Rubí decided to organize a *quasi-face-to-face workshop* on his own, by using the electronic materials available for the Free Workshop on the project website.
- Mollet del Vallès, convened the Culture department to hold a *quasi-face-to-face workshop* in an open day in order to involve more citizens and provided information for 80 citizens.
- Reus was using a distributed network strategy, so each of the specialists had to found only two or three participants and a phone interview was followed to fill in the questionnaire
- Cervemakers and the Institut de Cervelló, proposed participation in INSESS-COVID19 as a volunteer activity for 4th ESO students and they have also collaborated as agents of the project by monitoring citizen participation.

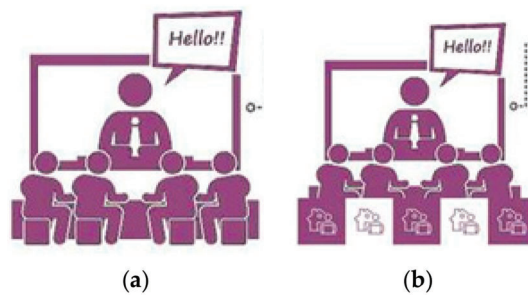


Figure 5. (a) Face-to-face modality (b) Mixt modality.

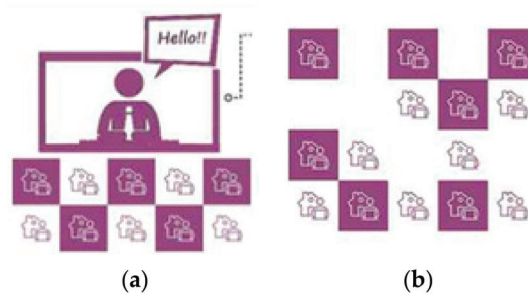


Figure 6. (a) Mixt modality; (b) Free modality.

### 2.10. Validation of Workshop Design and Technological Infrastructure

On 2 July, two pilots were conducted:

1. The BASS Castell-Platja d’Aro found 20 people who met some of the requested profiles and called for the workshop in a place provided by Social Services. Due to pandemics, INSESS-COVID 19 team joined the meeting remotely, through videoconference, and introduced the project, gave the context and all the instructions and answered all doubts about the questionnaire to the participants. After the 2 h workshop, all the 20 responses were already uploaded in the INSESS-COVID19 server. None of the questions was misinterpreted and all answers were provided.
2. The second pilot took place in the BASS la Noguera. In this case, trying to bridge the digital gap, social services professionals selected the 20 participants, and gave a phone call to pass the questionnaire; the professional was transcribing the citizen answers into the INSESS-COVID19 server. The time required to collect answers from the 20 participants took more than 2 months. None of the respondents misinterpreted any question.

### 2.11. Data Collection Methodology

According to the official statistics from last Third Sector Barometer [13], the vulnerable population from Catalonia is 1,584,000 people. The sample size can be determined



under the approach of infinite population, as the asymptote of the sample error under the finite population approach is reached around 1,000,000 population. According to classical expressions [14], a sample of 1067 citizens participating into the project would provide a sample error of 0.03 at a 0.95 confidence level.

Taking into account that the BAS were in an overflow crisis because of the pandemics, we assumed that about a 20% of them would not be able to engage the project, so, we determined that the network of 107 BASS from all territory would be asked to find 20 citizens each, by following a minimum of 10 of the target profiles. Social Services professionals for each BASS were selecting 20 citizens from a subset of profiles that properly represented the main problematics occurring in their geographical areas. Expertise of Social Services teams was on play at this step in a two-stage sample design, in a combination between collaborative co-creation methodologies and classical multiple stage sampling strategies.

The selected citizens were invited to participate in the project by following the INSESS-COVID19 workshops in any of its forms. The INSESS-COVID19 questionnaire was opened from 17th July 2020 and has been continuously collecting data until 6th December 2020. On 7th December 2020, (01:00 am) 971 answers were collected in a database containing 195 variables and downloaded for automatic analysis.

### 2.12. Typology of Variables in INSESS-COVID19 Questionnaire and Analysis Proposed

The INSESS-COVID19 questionnaire combines variables of different types and structures, which require different kind of analysis. Figure 7 lists the different types of variables considered in the questionnaire, the form that the question has in the digital questionnaire, the internal format generated at the level of the background database where data is represented, and the combination of graphical and numerical tools used for the basic descriptive analysis. In addition, an example of the INSESS-COVID19 questionnaire for each case.

This typology is one of the contributions of the paper and provides some complex variables that enquire to a certain issue and generate more than one column in the background dataset. According to the type of the question, the nature of the information collected and the way in which this information is represented in the database, this is directly affecting the way to visualize this data and the statistical procedure associated. This produced, in consequence, the creation of some new procedures to analyze these complex datasets and some new visualization tools.

- *Multivalued variables*: A multivalued variable is a qualitative variable  $X$  with  $S$  possible modalities  $D = \{m_1, m_2, \dots, m_S\}$  so that an individual can simultaneously take several values from  $D$ . This is the case for example of  $X = \text{"ICT available"}$ ,  $D = \{\text{PC, tablet, cellphone}\}$ . The values of  $X$ ,  $x_i \in P(D)$ , so that an individual can simultaneously have PC and cellphone and so. The concept of multivalued variable is not new. However, in this work specific descriptive tools to better extract information from these kinds of variables is introduced.
- *Temporal basic variables* Temporal variable  $(X, T)$  defines as a qualitative variable (nominal, ordinal, binary or Likert) that is measured  $T$  times along time, providing  $T$  columns in the dataset as temporal replicas of  $X$ . We can denote these replicas as  $X_{t1}, X_{t2}, \dots, X_{tT}$ . In the Figure 7 they are denoted by LikertXtime or NominaXtime
- *TQQ variable*: Temporal Qualified Qualitative  $(X, T, Q)$  is a qualitative variable  $X$  with  $S$  modalities in  $D$ , replicated  $T$  times. For each modality  $m_s$  a value (from  $Q$ ) indicating the qualification  $f_{ms}$  is given ( $Q$  is a Likert or ordinal set of values). As an example, variable  $X = \text{"participation in the society"}$ , is taking four modalities  $D = \{\text{Associations, Networks, Voluntariety, Others}\}$ , which indicates the kind of participative actions that the person follows. The variable is replicated  $T = 3$  times ( $t_1 = \text{January 2020}$ ,  $t_2 = \text{July 2020}$ ,  $t_3 = \text{January 2021}$ ). Each  $X_t$  is, in turn a set of Likerts, such that for each modality of  $X$  in  $t_1$  we have a Likert qualifying the degree of participation of the person in it.  $Q = \{\text{Molt (A lot, Una mica (Some), Gens (None), NC (missing answer)}\}$ . At the level of internal representation, each TQQ variable provides  $\text{card}(D) \times T$  Likert columns each taking values in  $Q$  that require a joint description.

Type	Form of question	DB structure	Graphical tools						Numerical tools																	
			Univariate		Multivariate	Bivariate	Tri-var	4-var	Univariate		Multivariate	Bivariate	Bivariate													
			Pie chart	Barplot	Histogram	Boxplot	WordCloud	Marginal Pie chart	Marginal Barplot	Multiple Barplots	Grid of plots	Trajectory maps	Stacked Barplots	Multiple Stacked Barplots	Frequency Table	Extended Summary	Standard Error	95% CI	Frequency table	Multivariate	Contingency table	Cross table	Cross table	Contingency table	Contingency table	
Numerica	Integer field	One numerical column																								
Likert or Ordinal	Simple choice response	One alphanumeric column	x																							
Nominal	Simple choice response	One alphanumeric column	x																							
Nominal Multivalued	Multiple choice response	One column with lists of values separated by ".,."																								
OrdinalXtime	Simple choice grid	One alphanumeric column per timeStamp																								
OrdinalXordinalXtime	Several Packs of several Ordinal	Several packs of several columns																								
Open question	Open textual window to be edited by the respondent	Textual column with complete text																								

Figure 7. Typology of questions, variables, data structures and analysis tools.



2.13. Design of Specific Visual and Analytical Tools for Complex Type of Variables

Indeed, some of the tools used are very basic, but others have been developed ex professo in this project and open the door to enlarge the knowledge provided in the first descriptive analysis of any database, given that the type of variables are properly conceptualized prior to the analysis itself. In the following, the description of the new advanced descriptive tools is proposed.

Each of these tools have been properly validated before including in the procedures used to analyze the project data. First, the proposal was validated with stakeholders of the report to see if they appreciated useful information given by the tool. Then, technical validation of scripts implementing them was performed. Finally, interpretability of the results was used as final validation criteria when the entire project report was submitted to final stakeholders.

2.13.1. Extended 5-Number Summary

Being  $X$  a numerical variable  $(x_1, \dots, x_n)$ , the 5-Number Summary [15] is a set of 5 sufficient robust statistics used to describe numerical variables (See Table 2). It is composed by Minimum, Q1, Median, Q3 and Maximum. In our version, we extend it by adding Mean, Quasi-standard deviation and Variation Coefficient, so that information about symmetry of the variable and relevance of variance can also be evaluated.

Table 2. Extended 5-Number Summary table.

Min	Q1	Median	Mean	Q3	Max	StDev (s)	CV
$\min(X)$	$x \text{ tq card}(X \leq x) = 0.25n$	$x \text{ tq card}(X \leq x) = 0.5n$	$\frac{\sum_{i=1}^n x_i}{n}$	$x \text{ tq card}(X \leq x) = 0.75n$	$\max(X)$	$\frac{\sum_{i=1}^n (x_i - \bar{x})^2}{n-1}$	$\frac{s}{\bar{x}}$

2.13.2. Extended Frequency Table

Being  $X$  a nominal variable the Extended Frequency table (Table 3) extends the traditional one with the standard error, computed according to expressions described in this paper and the pooled standard deviation of all modalities together as a goodness indicator of the question as a whole. For nominal qualitative variables, the modalities are presented in descending order, in a Pareto style, so that the most frequent modalities appear in the top of the table. For Likert variables, the original order of the modalities is presented.

Table 3. Extended Frequency table of Gender.

P2. Gènere	Freq.	Prop.	Std. Err
2. Dona (Female)	655	0.675	0.0152
1. Home (Male)	307	0.316	0.0148
3. NoBinary	5	0.005	0.0032
4. NC	4	0.004	0.0000

95% CI error:  $\pm 5 \times 10^{-4}$ ; Std. Error of the question: 0.0107.

2.13.3. Marginal Bar Plot, Pie or Frequency Table

The multivalued variables provide multivalued responses composed by subsets of modalities. This is the case for example of digital devices used by a person (they can be multiple, right? Cell phone, tablet, pc, laptop ...). Being  $X$  a multivalued variable  $(x_1, \dots, x_n)$ , where  $x_i$  is a list of modalities separated by “;”. The frequencies of each single modality of the variable are not available by direct analysis.

The marginal bar plot, as in Figure 8, apparently looks like the classical bar plot, but it is built over a multivalued variable. This means that a single individual might be represented in several bars simultaneously. Consequently, the corresponding proportions column have a total overcoming 100%. So that the marginal frequency table has a similar aspect to the frequency table but represent proportions that sum up over 100%. The same happens with the pie. All of them represent the marginal counts or proportions of the

(eventual) dummies representing each of the modalities of the variable, independently of how this variable is internally represented in the data base (as a single column of lists of values in the cells, or as a set of dummies, one per modality). Figure 8 shows the area of the life impacted by unsolved processes. The same person can have several areas impacted simultaneously, like civil status (divorce process for example) and economy and family.

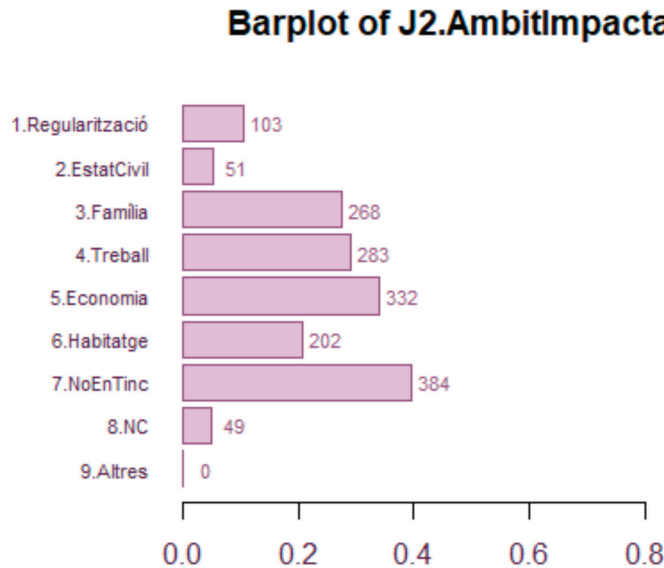


Figure 8. Marginal bar plot of J2 question.

#### 2.13.4. Multivalued Frequency Table

As nominal multivalued variables are represented by columns with lists of modalities in the cells, we propose the multivalued frequency table (Table 4) to analyze the bags of modalities selected by respondents. In the multivalued frequency table, all the subsets of modalities provided as answers are displayed with their corresponding counts and frequencies. This in fact represents a subset of the empirical joint probability distribution of the variable. To preserve the statistical secrecy combinations are published only for frequencies greater than three. The number of hidden combinations is also reported at the end, as well as the uncertainty metrics. These variables are implemented through multiple-choice questions in the questionnaire. When collapsed in bags of modalities their weight in the analysis keep as one variable. When represented as dummy variables, as in the traditional way, they can bias the analysis as they increase dimensionality of data set unnecessarily.

#### 2.13.5. Trajectory Graph

Originally introduced in [16,17], it consists of a two-dimensional plot with the modalities of the target qualitative variable (sorted or depending if it is nominal, or Likert or ordinal). Time is represented in the X-axis and it is discrete. In the INSESS-COVID19 questionnaire, this tool is used to represent all the temporal basic variables. Those corresponding to Dynamic characteristics and measured at the three time stamps presented before: January 2020, July 2020 and January 2021. For each individual the nodes representing their choices along time are linked with an edge. Edges of same colour represents same trajectory of the individuals. The thickness of the trajectories represents the proportion of respondents following that pattern. Trajectory graphs represent in a single tool packs of 3 different columns in data file corresponding to same variable X measured in 3 timestamps

$X_{T1}$ ,  $X_{T2}$ ,  $X_{T3}$ ; where each  $X_{Ti}$  is a replica of  $X$  showing the value along time. Trajectory Graphs teach which individuals evolve in similar ways. They give an opportunity to identify temporal patterns and further find which variables distinguish them. This interpretative analysis generates hypotheses about which factors are associated to negative evolutions or harmful for individuals. The tool is transversal, and it has been used in [16] to identify causes of functional impairment in neurological patients with spinal cord injury during the process of social inclusion after discharge. In [17] it was used to understand the patters of evolution of the operation mode of wastewater treatment plants daily. Here we apply to discover the main trends of temporal evolution of the main variables from INSESS-COVID19 questionnaire one by one.

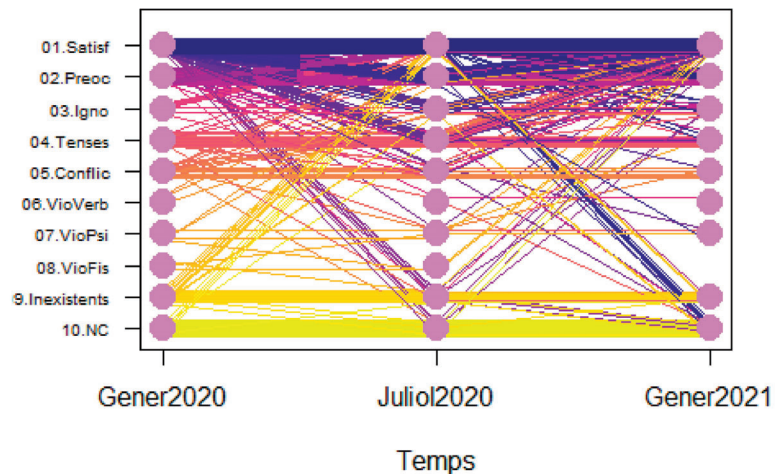
**Table 4.** Multivalued frequency table.

J2. AmbitImpactat	Freq.	Prop.	Std. Err
7. NoEnTinc	378	0.389	0.0155
3. Família	56	0.058	0.0077
4. Treball; 5. Economia	51	0.053	0.0071
8. NC	47	0.048	0.0071
3. Família; 4. Treball; 5. E-conomia; 6. Habitatge	45	0.046	0.0071
5. Economia	39	0.040	0.0063
3. Família; 4. Treball; 5. Economia	32	0.033	0.0055
4. Treball; 5. Economia; 6. Habitatge	30	0.031	0.0055
1. Regularització	27	0.028	0.0055
4. Treball	27	0.028	0.0055
3. Família; 5. Economia	24	0.025	0.0055
2. EstatCivil; 3. Famí-lia; 4. Treball; 5. Economia; 6. Habitatge	16	0.016	0.0045
1. Regularització; 3. Fami-lia; 4. Treball; 5. Economia;	14	0.014	0.0032
6. Habitatge	14	0.014	0.0032
6. Habitatge	14	0.014	0.0032
3. Família; 5. Economia; 6. Habitatge	11	0.011	0.0032
5. Economia; 6. Habitatge	11	0.011	0.0032
3. Família; 4. Treball	8	0.008	0.0032
1. Regularització; 4. Treball; 5. Economia; 6. Habitatge	7	0.007	0.0032
1. Regularització; 3. Fa-mília	6	0.006	0.0032
1. Regularització; 4. Treball; 5. Economia	6	0.006	0.0032
4.Treball; 6. Habitatge	6	0.006	0.0032
1.Regularització; 4. Treball	5	0.005	0.0032
2. EstatCivil; 3. Famí-lia; 5. Economia; 6. Habitatge	5	0.005	0.0032
1. Regularització; 2. EstatCivil; 3. Família; 4. Treball;	4	0.004	0.0000
5. Economia; 6. Habitatge	4	0.004	0.0000
1. Regularització; 3. Família; 4. Treball; 5. Economia	4	0.004	0.0000
1. Regularització; 3. Família; 5. Economia	4	0.004	0.0000
2. EstatCivil	4	0.004	0.0000
3. Família; 4. Treball; 6. Habitatge	4	0.004	0.0000
3.Família; 6. Habitatge	4	0.004	0.0000
1. Regularització; 3. Família; 5. Economia; 6. Habitatge	3	0.003	0.0000
1. Regularització; 6. Habitatge	3	0.003	0.0000
2. EstatCivil; 3. Família	3	0.003	0.0000
2. EstatCivil; 5. Economia	3	0.003	0.0000
Salut	3	0.003	0.0000

Using Trajectory Graphs in R is another contribution of this paper, this being the first time that it is implemented in R to be automatically represented in automatic reporting. Figure 9 shows the trajectory graph for the variable convivial relationships.

The variable Quality of convivial relationships is ordinal and can take 10 different modalities (from 01. Satif (Satisfactory) to 9. Inexistent and 10. NC (missing)). This variable has been measured by three timestamps in the questionnaire. A line in the graph represents each respondent. All respondents following same temporal path are shown with same line color.

## Evolució de R1.RelUConv al llarg del temps



**Figure 9.** Trajectory map of question R1.

The interpretative power of this tool for non-technical-skilled users is enormous: Horizontal bands mean stability. Whenever the modalities of the target variable (X variable) are sorted top-down from better to worse, the “V” and “^” patterns mean instability found after pandemics 1st wave (in July 2020) in opposite senses. While “V” pattern means worsening and retrieving, the “^” pattern means improvement after pandemics and bad hopes in January 2021. Of course, the trajectory map can be generalized to more timestamps and any kind of qualitative variable. It is useful to understand the dynamics of a group of individuals along time. Another contribution of this research is that an efficient algorithm was designed so that the combinatorial nature of the trajectories can be managed and computed in very short CPU times.

The “V^” pattern is a special pattern identified for the first time during this research. It corresponds to a double dynamics in the same process (in this case, the pandemics), where part of the individuals follow a “V” pattern (the pandemics worsen their situation and they expect to recover by the beginning of 2021) whereas another segment of individuals follow the opposite pattern “^” (they were in bad conditions before the pandemics and the pandemics connected with people, better emotional conditions etc., while they expect to come back to the original situation by the beginning of 2021).

### 2.13.6. Trajectory Frequency Table

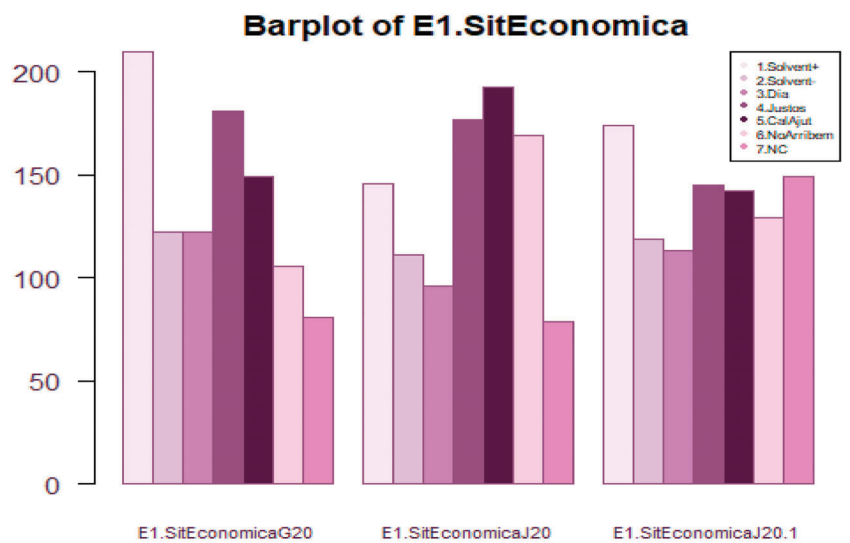
For temporal basic variables: Apparently looks like a multivalued frequency table. The main difference is that it has been built from a set of several qualitative variables (one per timestamp), each of them are simple choice and represented in a different column in the dataset. It quantifies the information shown in the Trajectories map. See in Table 5 the trajectory frequency table corresponding to the R1. RelUConv variable presented later in the Results section.

### 2.13.7. Multiple Bar Plot

As usual, it represents the joint probability distribution of 2 qualitative variables. In this case, one is time. The other is a nominal, ordinal or Likert variable. For temporal basic variables. See an example in Figure 10.

**Table 5.** Example of Trajectory Frequency Table.

R1.RelUConv	Frequencies
01.Satisf + 01.Satisf + 01.Satisf	596
10.NC + 10.NC + 10.NC	64
02.Preoc + 02.Preoc + 02.Preoc	33
09.Inexistents + 09.Inexistents + 09.Inexistents	25
01.Satisf + 02.Preoc + 02.Preoc	24
01.Satisf + 02.Preoc + 01.Satisf	23
02.Preoc + 01.Satisf + 01.Satisf	14
02.Preoc + 02.Preoc + 01.Satisf	12
01.Satisf + 01.Satisf + 02.Preoc	9
01.Satisf + 01.Satisf + 10.NC	7
04.Tenses + 04.Tenses + 04.Tenses	7
09.Inexistents + 01.Satisf + 01.Satisf	7
01.Satisf + 04.Tenses + 01.Satisf	6
02.Preoc + 01.Satisf + 02.Preoc	6
05.Conflic + 05.Conflic + 05.Conflic	6
03.Igno + 01.Satisf + 01.Satisf	5
04.Tenses + 04.Tenses + 01.Satisf	5
01.Satisf + 03.Igno + 01.Satisf	4
01.Satisf + 09.Inexistents + 10.NC	4
02.Preoc + 04.Tenses + 04.Tenses	4
04.Tenses + 05.Conflic + 01.Satisf	4
05.Conflic + 01.Satisf + 01.Satisf	4
09.Inexistents + 03.Igno + 10.NC	4
10.NC + 01.Satisf + 01.Satisf	4
01.Satisf + 03.Igno + 03.Igno	3
01.Satisf + 04.Tenses + 04.Tenses	3
03.Igno + 03.Igno + 03.Igno	3
04.Tenses + 01.Satisf + 01.Satisf	3
04.Tenses + 04.Tenses + 03.Igno	3
05.Conflic + 04.Tenses + 01.Satisf	3



**Figure 10.** Multiple bar plot of Economic situation.

2.13.8. Grid of Pies

For temporal basic variables the T columns representing time can be analysed independently as if they were ordinary qualitative variables. A pie for each timestamp can be done and they are presented in a grid See an example in Figure 11 for economic situation.

**E1.SitEconomicaG20      E1.SitEconomicaJ20      E1.SitEconomicaG21**

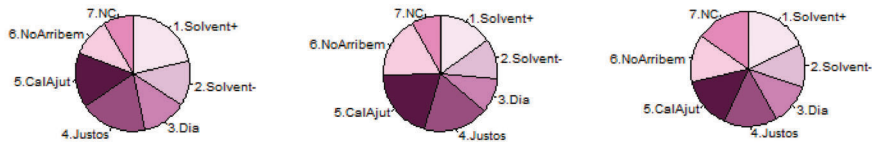


Figure 11. Grid of Pie Charts of question E1. Economic Situation.

2.13.9. Transition Tables

Tables quantifying the transitions between two consecutive timestamps, in counts or proportions. Given a temporal basic variable (X,T), it is the cross table between  $X_t$  and  $X_{t+1}$ ,  $t = \{1:T - 1\}$ . See an example in Figure 12 for the changes in the quality of convivial Unit Relationships between January 2020 and July 2020.

	01.Sat- isf	02.Preoc	03.Igno	04.Tenses	05.Con- flic	06.Vio -Verb	07.Vio -Psi	08.Vio -Fis	09.Inexis- tents	10.NC
01.Satisf	616	50	9	12	5	0	1	0	6	4
2.Preoc	21	48	0	8	2	1	0	0	1	0
03.Igno	5	0	3	0	1	0	0	0	0	0
04.Tenses	3	1	0	16	4	1	1	0	1	0
05.Conflic	4	1	3	4	7	0	1	0	0	1
06.VioVerb	1	3	2	0	0	0	0	0	0	0
07.VioPsi	1	1	0	0	0	0	2	1	0	0
08.VioFis	0	0	0	0	0	0	1	1	1	0
09.Inexistents	9	0	4	0	0	0	0	0	28	3
10.NC	4	0	0	1	0	0	0	0	1	66

Figure 12. Changes between January 2020 and July 2020 (variable gener 2020–Juliol 2020).

2.13.10. Changing Tables

Cross table of the categorization of successive transition tables that quantifies how many state changes are observed in both the first and second transition. Figures 13–16 are examples on changes of the quality of the relationships in the Convivial Unit.

	Millor	Igual	Pitjor	NC	Sum		Freq	Prop
Millor (better)	0.009	0.042	0.010	0.004	0.066	Improve	76	0.078
Igual (Same)	0.027	0.690	0.016	0.009	0.743	V pattern	4	0.004
Pitjor (Worse)	0.056	0.040	0.004	0.009	0.109	Balance	670	0.690
NC	0.001	0.005	0.000	0.076	0.082	Λ pattern	10	0.010
Sum	0.093	0.778	0.031	0.099	1.000	Enworse	59	0.061

(a)

(b)

Figure 13. (a) Changes reported along time (relative); (b) Change patterns (convivial unit).

	Freq	Prop		Freq	Prop
Improve	76	0.078	Improve	65	0.067
V pattern	5	0.005	V pattern	7	0.007
Balance	642	0.661	Balance	662	0.682
Λ pattern	13	0.013	Λ pattern	10	0.010
Enworse	83	0.085	Enworse	80	0.082

(a) (b)

Figure 14. (a) Change patterns in familiar relationships with person living out of home; (b) Change of patterns in relationships with neighbours.

2.13.11. Multiple Stacked Bar Plot

This is a graphical representation proposal to provide a compact view of a TQQ type variable with a Q, X and T. In this case, the three stacked bar plots represent participation in society through time. For each timestamp T = (G20, J20, G21), a stacked bivariate bar plot represents the relationship between the Likert Q (1Molt (high participation), 2.Una mica (moderate participation), 3Gens (no participation), 4NC (unknown)) (in bars) and the modalities of X, here indicating if the participation in different social activities (like neighborhood networks (Xarxes), associations (Associacions) voluntary movements (Voluntari) or Others (Altres)). Changes along time can be analyzed as well. See Figure 15.

Soc1.ParticipGener2020 Soc2.ParticipJuliol2020 Soc3.ParticipGener2021

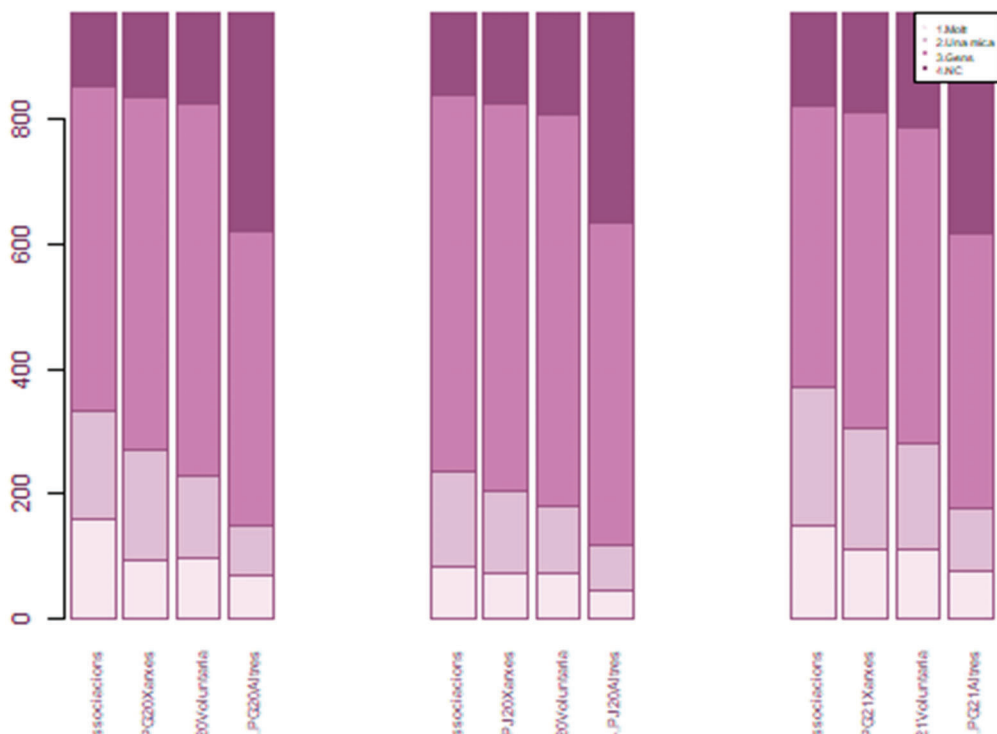
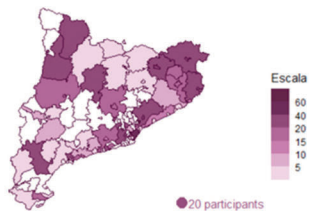


Figure 15. Multiple stacked plot of Question Soc1-2-3. From left to right the vertical labels are: Soc1.1.PG20Associacions (participation in associations in January 2020); Soc1.2.PG20 Xarxes; Soc1.3.PG20Voluntaria; Soc1.4.PG20Altres and so on.



## 2 Distribució de les respostes

Nro de respostes per Area Basica de Serveis Socials



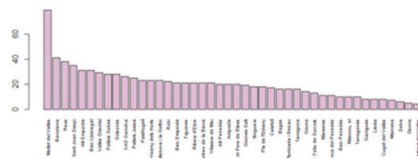
Number of modalities: 67

### Frequency table

Àrees bàsiques de Serveis Socials	Freq.	Prop.
Mollet del Vallès	79	0.086
Barcelona	41	0.044
Reus	38	0.041
Sant Joan Despí	35	0.038
Alt Empordà	31	0.034
Baix Llobregat	31	0.034
Vallès Oriental	29	0.031
Pallars Sobirà	28	0.030
Solsonès	28	0.030
CAS Garrotxa	26	0.028
Pallars Jussà	25	0.027
Palafurgell	23	0.025
Sant Vicenç dels Horts	23	0.025
Vilanova i la Geltrú	23	0.025
Rubi	22	0.024
Baix Empordà	21	0.023
Figueras	21	0.023
Ribera d'Ebre	21	0.023

Sant Andreu de la Barca	21	0.023
Vilassar de Mar	21	0.023
Alt Penedès	20	0.022
Ampostà	20	0.022
Sant Pere de Ribes	20	0.022
Gironès-Salt	19	0.021
Noya	18	0.020
Fla de l'Estany	18	0.020
Calafell	17	0.018
Bages	16	0.017
Montcada i Reixac	16	0.017
Tarragona	16	0.017
Girona	14	0.015
Sant Feliu de Guíxols	13	0.014
Maresme	11	0.012
Vilafranca del Penedès	11	0.012
Baix Penedès	10	0.011
Masnou, el	10	0.011
Tarragonès	10	0.011
Garrigues	8	0.009
Lleida	8	0.009
Sant Cugat del Vallès	8	0.009

Diagrama de Pareto (n. respostes)



## 3 Detall de la participació de les ABSS comarcals per municipis

Nota, la participació de les ABSS municipals es pot consultar al capítol anterior. En aquesta secció donem detalls de com les respostes s'han distribuït entre els diferents municipis per aquelles ABSS que són Consells Comarcals.

### 3.1 Alt Camp

Nro de respostes de l'ABSS: 2

Number of modalities: 2

#### Frequency table

Àrees bàsiques de Serveis Socials	Freq.	Prop.
Brufill	1	0.5
Montferrit	1	0.5

### 3.2 Alt Empordà

Nro de respostes de l'ABSS: 31

Number of modalities: 18

#### Frequency table

Àrees bàsiques de Serveis Socials	Freq.	Prop.
Escala, l'	7	0.226
Llança	5	0.161
Cabanes	2	0.065

Port de la Selva, el	2	0.065
Roses	2	0.065
Agullana	1	0.032
Cadaqués	1	0.032
Castelló d'Empúries	1	0.032
Darnius	1	0.032
Palaus-saverdera	1	0.032
Peraleu	1	0.032
Port de Molins	1	0.032
Sant Pere Pescador	1	0.032
Torreella de Fluvià	1	0.032
Ventalò	1	0.032
Vilabertran	1	0.032
Vilajuïga	1	0.032
Vilameurola	1	0.032

### 3.3 Alt Penedès

Nro de respostes de l'ABSS: 20

Number of modalities: 16

#### Frequency table

Figure 16. Sample pages of automatic report.

### 2.13.12. Error Estimation

The results of all estimates build over questionnaires data have associated sampling errors. The main statistical offices in our context have been consulted and two different methods are used to compute them.

### 2.14. Statistical Institute of Catalonia (IDESCAT)

IDESCAT is the statistical office from Catalonia and uses the Variance Coefficient (CV) of the estimate  $\hat{\theta}$  as the estimation of the relative sampling error for the estimate  $\hat{\theta}$ . CV is published in the sampling error tables. The estimated CV allows obtaining a confidence interval at 95% of the estimated characteristic ( $\theta$ ):

$$\left[ \hat{\theta} \pm 1.96 \widehat{CV} \times \hat{\theta} \right] \tag{1}$$

In turn, computing  $\widehat{CV}$  follows the recommendations of Eurostat and the Net-SILC2 working group [18], so that the error clustering and the ultimate cluster approach are used. According to this methodology, for the calculation of the variance of the sampling error, only the variation between the totals of the primary sampling units (the census tracts) is taken into account. This might parallel the BASS role in our case.

### 2.14.1. Statistical National Institute (INE, Instituto Nacional de Estadística)

[The sampling errors of the estimates of some of the main investigated characteristics are calculated quarterly. A resampling method is used to obtain the sampling errors. The INE uses the reiterated semi samples method [19,20] in most of their important panels, among them the APS (the Active Population Survey, EPA in Spanish) [21].

This procedure consists of obtaining  $r$  semi samples from data (being a semi sample a subsample of size  $n/2$ , with  $n$  the original sample size). From each semi sample  $s$ , the estimate  $\hat{\theta}_s$  of the target parameter  $\theta$  is calculated. Once all the estimates have been calculated, as well as the estimate of the full sample  $\hat{\theta}$ , the variance estimator is given by:

$$\widehat{V}(\hat{\theta}) = \frac{1}{r} \sum_{s=1}^r (\hat{\theta}_s - \hat{\theta})^2 \tag{2}$$

where  $r$  is the number of subsamples considered,  $\hat{\theta}_s$  is the estimate of  $\theta$  obtained with the semisample  $s$  (a reweighting technique is applied using the CALMAR software) and  $\hat{\theta}$  is the global estimation of the target parameter, based on complete sample.

In the case of the APS, the number of reiterations used is 40, formed by making pairs with the sections of each strata, ensuring that the two sections of each pair belong to the same APS rotation shift; the first section of each was randomly assigned for 20 reiterations and the other section for another 20. In this way, each reiteration is constituted by a number of sections equivalent to 50% of the sample (semi sample) and each section appears in the half of the iterations. The survey publishes the relative sampling error as a percentage (coefficient of variation):

$$\widehat{CV}(\hat{\theta}) = \sqrt{\widehat{V}(\hat{\theta})} \times 100/\hat{\theta} \tag{3}$$

#### 2.14.2. Calculation of Sampling Error in INSESS-COVID19

In our case, we provide the CV of each item of the questionnaire based on the same expression used by IDESCAT

$$\left[ \hat{\theta} \pm 1.96 \widehat{CV} \times \hat{\theta} \right] \tag{4}$$

For the numerical variables  $\hat{\theta}$  is the observed mean and for the qualitative ones is the observed proportion. The

$$\widehat{CV}(\hat{\theta}) = \sqrt{\widehat{V}(\hat{\theta})}/\hat{\theta} \tag{5}$$

as usual. So, the most important part in our case is to estimate  $V(\hat{\theta})$ . For numerical variables, it is estimated as the square of the sample quasi-standard deviation. For the qualitative variables, each modality is considered as following a Bernoulli distribution, so that  $\theta$  represents the proportion of that modality, whereas

$$V(\hat{\theta}) = \frac{\hat{\theta}(1 - \hat{\theta})}{n} \tag{6}$$

In addition, a confidence of the qualitative question as a whole is provided by means of the pooled standard deviation of all modalities.

#### 2.15. Privacy

Many of the questions contained in the INSESS-COVID19 questionnaire are sensitive (being the object of violence, being in irregular situation in the country, suffering from mental disorder, etc.). Guaranteeing the privacy and anonymity of the respondents is crucial to make them sure that they can answer all the questions without being scared.

This is the reason why the questionnaire is self-contained and anonymous, such that the respondent cannot be identified and their answers cannot be crossed with any other database at individual level. In particular, they cannot be crossed with the Social Services information systems. So that we cannot expect to get any extra information about the person out of the questionnaire. Some questions require information that Social Services already have about people, but we preferred to ask again and avoid mistrust feelings that could limit the answers provided by the respondents.

To guarantee this security, the BASS professionals identify the people to participate into the workshops, but they do not share with INSESS-COVID19 team their identities. They

communicate to the participants the links and passwords to enter the project website and the questionnaire but using a common password the system cannot trace the identities of the respondents, so that the responses keep anonymous and secure. The server hosting the questionnaire database is RGPD compliant as well, and INSESS-COVID19 team preserves the microdata without sharing with any other institution other than aggregated data.

However, all these good practices are not sufficient to guarantee the statistical secrecy of the respondents.

#### 2.16. Risk of Revelation of Statistics Secrecy and Preservation

The citizen's profiles targeted by INSESS-COVID19 project focus on some subpopulations that represent minorities presumed to be impacted by the COVID. The data collection process has been distributed along the territory in order to minimize the efforts required to BASS professionals, already collapsed by the management of the cases impacted by pandemics. Some of the BASS were providing more than the required 20 citizens, but a number of them provided around 20 or sometimes less. This means that for some profiles, a BASS can provide one or two single people. This raises serious limitations for publishing classical descriptive statistics at the BASS level, as it would be easy for the BASS professionals to disclose the statistical secrecy by identifying the person. This phenomenon happens not only when data is presented at BASS level, but even when minority profiles are studied at Catalan level, by crossed with other variables that can reveal sufficient information to identify the people.

The classical practice of not publishing results about too small subpopulations is not a solution in the context of this project, as vulnerable minorities (even is not statistically significant) require attention and cannot disappear from the picture (let us think about women victim of domestic violence, they are never too much, but this is not a reason to hide in the analysis what happens with this segment of population, right?)

INSESS-COVID19 is proposing and applying some good practices that preserve statistical secrecy even in front of very small subpopulations.

All data has been taken into account for the computation of global statistics.

All modalities of qualitative variables with too small number of responses have been hidden from the public report (only those with a minimum of 10 responses have been published). The modalities with some responses but not enough to be public are listed in the report. Therefore, one can know that less than 10 people have been accounted in the study for those modalities, but exact number is not available.

Target profiles with less than three participants are only listed as present profiles in the sample, but without the exact number of respondents. This is particular important when the results are reported at BASS level.

The target profiles are not mutually exclusive. Thus, many of the citizens participating in the study simultaneously meet several profiles: for example, single-parent women who also work in the field of essential services, or men with very low wages and in a situation of under-housing, etc. This makes possible to decrease the publishable threshold until three, since one cannot know if the people in this "hidden profile" have only this characteristic or some others and identification of the person keeps preserved.

#### 2.17. Territorial Information

As usual when data is collected over a territory, a map visualizing the statistical information is very relevant. In INSESS-COVID19, four territorial levels were apparently suitable: Cities and villages, BASS, Vegueries, and Provinces. The 947 Catalan municipalities are grouped at a first administrative level in 42 Comarcas. Each comarca is a BASS managing all municipalities inside the comarca with less than 20,000 inhabitants. The municipalities with more than 20,000 inhabitants are a BASS themselves as well. Therefore, Catalonia has 107 BASS in the territory. Vegueries is an intermediate grouping of comarcas. Catalonia has eight Veguerias and four provinces. The province is too big to be considered in the INSESS-COVID19 study as the heterogeneity inside a single province is too high

from the social vulnerability side. Thus, BASS and Veguerias are the two territorial levels considered for geographical representation.

It is worth to mention that qualitative variables cannot be represented in maps as a whole, but some specific modalities have to be selected and their territorial proportions represented one by one.

### 2.18. Metainformation Model

Once the different types of variables have been defined, and the statistical tools to analyze each type of variable is clear, a mechanism to provide intelligence to the scripts performing the descriptive analysis is required. This is based on variable declaration and the implementation is designed on the basis of a metainformation file that provides all required conceptual information to the R system to run proper descriptive analysis, able to use predefined descriptive procedures for each type of variable. The metainformation file has to contain all contextual information from data. Our proposal is to use a metainformation file in form of a table (implementable as a csv file for example) with the following structure: The rows are associated to variables. Some variables provide metainformation through several rows.

- Col: Number of column where the variable is in the dataset
- Etilcol: For rows containing modalities of a qualitative variable. It indicates the column of the variable if it is represented in a single column. For TQQ variables it contains the number of the columns containing the modalities.
- Block: As the questionnaire was designed by thematic blocks (economy, health, etc.) the number of the block of the question is specified
- Block name: Specifies the name of the block
- Block label: Short label for the block to be used in the report
- Question: Complete text of the question as it is appearing in the digital questionnaire
- Answers: The rows below the question contain all modalities in D
- Columns: Values of Q for TQQ variables
- Rephrasing: A short expression for both questions and possible answers to be used in the statistical tables and graphs, as the long texts will overlap and make reading difficult
- Colcut: Short labels for Q values in TQQ variables
- Object: Indicates the kind of information in the row (question, modality, name of Block, separator (between variables))
- Type of variable, according to the typology defined above
- Descriptive type: Descriptive procedure associated (some types of variable follow similar descriptive tools). Each descriptive procedure uses a specific combination of descriptive tools (classic or including the innovative ones proposed in this paper)
- Comments: There is a space to write context information if it is needed.
- Reference: In case that question is inspired in a reference survey (see Table 1). Useful for validation
- View: For pattern interpretation, we divided the variables in different groups, depending on the principal topic.

### 3. Automatic Analysis and Reporting

The key for a getting a quick feedback and, as a consequence, a quick support for the decision-making is to have the technological infrastructure ready to collect data as well as to analyze the data as soon as the collection period is closed.

Data arrives to the on-line questionnaire automatically as soon as participants provide their responses without additional intervention of the research team, other than ensuring the permanent availability of the server.

At any moment, data can be downloaded from the on-line questionnaire in form of a csv file, so several waves can be treated as well to form a continuous panel if required.

The contents of the csv file represents the several questions from the questionnaire following the formats described in Figure 7 according to the type of the variables representing the different questions.

Given a certain questionnaire, a metadata file can be linked with it, by indicating which type correspond to each variable, and which columns contains the information relative to that variable in the csv.

Each questionnaire requires its own metadata file. Changing questionnaire is relatively simple, so that modifications in the corresponding digital questionnaire can be easily done, and the corresponding metadata file must be modified accordingly.

The analysis of the data collected in the questionnaire is automatically processed through some R and Rmarkdown scripts, which inputs both the dataset in csv format and the corresponding metadata file.

A knowledge component is also implemented, so the procedures know in each moment which kind of analysis is appropriated for each variable, according to its type. This gives the intelligence to the system and is able to manage exceptions. In addition, it can be modified to add new data types including other analysis tools when required. This component is the one including all the guidelines that guarantee the preservation of the statistical secrecy in front of small samples mentioned in previous sections.

In addition, a very important part of the procedure is that Rmarkdown has been designed for automatic reporting in such a way that it produces a formatted Word document with the results. So, the result of the analysis is an editable Word file ready to be read, commented, and post-processed in a very easy way by the decision-maker itself, just requiring specific domain expertise to select the relevant results, to add complementary explanations for the analytical findings, to synthesis the findings in a short overview or to reorder them in a rational that makes sense for the communication of results.

When the analysis must be repeated periodically (every six months for example), the system is also prepared to add those reordering and selection criteria into the automatic reporting part, thus producing a results document much closer to what the expert need to communicate results.

As said before, the INSESS-COVID19 questionnaire is generating a csv file with 195 columns representing 25 blocks of information. Some of the variables split in many columns by internal representation, as explained before. The total elapsed time between downloading the csv file from the questionnaire (located in the server) and getting the Word file containing the results of the analysis by using the scripts designed in the project is about 15 min on average. And the aspect of the obtained document is very close to a final report, as it can be seen in the Figure 16.

This means that the methodology developed by INSESS-COVID19 project provides a technological infrastructure that permits to get direct and fresh information from the citizens, specific professional collectives or relevant actors involved in a certain decision by direct participation tools, where:

- The decision-maker can decide what to ask, even if its information system is not collecting that information (modification of the questionnaire require less than 2 h)
- The decision-maker can decide who must receive the questionnaire and when (sample design and representability of respondents being crucial)
- The decision-maker can decide if answering the questionnaire is voluntary or mandatory and the response deadlines

Depending on the case, call the respondents may be immediate if personal mails are available, or might require more time, if intermediate institutions must find them and call. However, this is out of the technological part of the proposed methodology.

Once the participants have been called and new questionnaire activated, 20 min would be enough for responding a questionnaire of similar extension as the one build for INSESS-COVID19, and 15 min would provide the working document with the results of the analysis for diagnoses and interpretation, thus constituting a very powerful tool for

quick diagnoses of relevant situations for further decision-making, and for implementing direct participatory strategies in a new way of policy-making.

Of course, the proposed tools are not restrictive for policy making, but its use can be extended to monitor any kind of industrial or business process through data monitoring, just modifying the questionnaire, or the input data of the corresponding scripts.

In the following, we synthesize the results of the analysis of the INSESS-COVID19 questionnaire.

#### *Sample Validation*

After the data collection, a further validation of the representability of the sample should be required. In addition, this can be pursued by making proportion comparison statistical tests and homogeneity tests to check whereas the distribution of the sample is homogeneous to the distribution of the population. However, this is the first time in Catalonia (and probably in Spain) that a study is conducted targeting 20 vulnerable profiles, independently if they are current users of Social Services System or not. In addition, there are not population data available to make this validation. In fact, all reference official statistics or reports consulted as State of the Art have some similarities with our study, but target populations are not directly comparable, so precluding the possibility to test this part. Being the first time that such a population is analyzed, this work will become the reference to test other studies in the future.

In spite of this limitation, we tried to go further and inspected some of the referent official statistics and reports to see if we could get some clues and indications that our sample is indeed well representing the reference population.

Official statistics from INE or IDESCAT like census or the *padró* provide data about the proportion of disabled population in Catalonia, for example, and since all disabled people gets a certification from Social Services, it happens that if the INSESS-COVID19 sample is valid, the sample proportion of disabled persons should be equal to the real proportion reported in the IDESCAT. Same happens with the assigned housing; all families that gained the right to have a gratuit house have been linked to the Social Services system to manage it and the IDESCAT in the Anuari Estadístic de Catalunya 2019 reports the proportion of the Catalan population in this situation that is as well comparable with the one appearing in INSESS-COVID19 sample. Same situation occurs with Widow people, which is officially reported in the census from INE and all of them process their pension through the Social Services system as well. (Table 6) However, the proportion of married people would not be comparable. Indeed, since census is done for the entire population and being vulnerable or not directly impacts in the capacity of marrying (which is an indicator of stability), official census statistics on married people cannot be directly compared with those from our sample, where only vulnerable population is targeted.

The official report from Social Services in Catalonia (the Rudel report) cannot be used for the comparison, since is only reporting about Basic Social Services, and we are also including in our study other segment of populations like mental health patients which are users of Specialized Social Services and same happens with other profiles included in the sample. In addition, the Third Sector Barometer provides interesting information, but only regarding Third Sector users, as expected, and in our sample, we are including people that never before has been linked to the Social Services System neither to other Third Sector entities. For example, entrepreneurs that had bankrupt are included in the INSESS-COVID19 sample cause they are a vulnerable group that merits attention and might become users of the Social Services system in the near future, but these people have never been part of any of the statistics provided by Third Sector Barometer or Rudel report. In addition, workers from essential services occurred in the INSESS-COVID19 sample come from healthcare system, social services system and hostelry sector. None of them structurally linked with Social Services before. In addition, official statistics about the size of those professional sectors are unusual as well, since they include non-vulnerable people, which are not targeted in the INSESS-COVID19 sample design.

In synthesis, for those indicators where an external official statistics is available and comparable with the configuration of INSESS-COVID19 sample, the sample looks representative, but the global validation is non-suitable, being INSESS-COVID19 a pioneer study in its category.

Finally, the global statistic error of the sample is 3%, which is small enough to provide significant results.

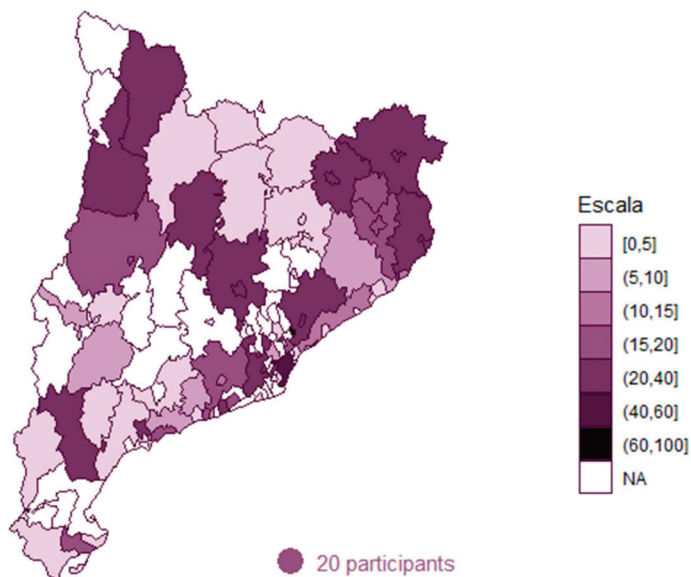
**Table 6.** Validation. Sample proportion against Population Proportion.

Population Segment	Sample Prop	Population Proportion	p-Val	Significant Difference	Source
Disabled people	15.8	14.8	0.82	No	IDESCAT economic Survey
Assigned housing	0.033	0.036	0.64	No	IDESCAT Anuari Estadistic de Catalunya
Widows	0.084	0.075	0.27	No	INE census

#### 4. Results

In the following the main results of the questionnaire, presented to the Catalan government last 15th December 2020 are synthesized in such a way that the different tools used in the analysis are illustrated and global results discussed. The territorial coverage of the respondents is reasonable (971 responses), although some areas in Tarragona province did not engage the INSESS-COVID19 project as a consequence of the overflow in Social Services already referred before. Here number of responses are presented in aggregated way. Later, those BASS with less than five respondents are preserved from public results, and only used for internal analysis and for building the final global results.

Figure 17 visualizes the participation of the BASS providing some response to the questionnaire. White corresponds to BASS that did not participate into the project. Figures 14 and 18 shows the Pareto diagram. It can be seen that some specific BASS provided more than the required 20 participants. Figure 19 provides participation at the level of Vegueria.



**Figure 17.** Number of responses per BASS.



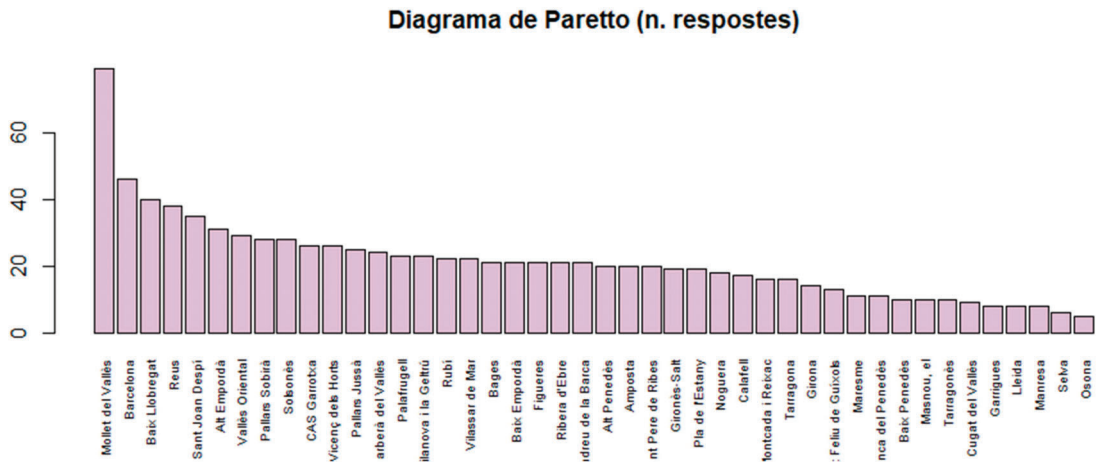


Figure 18. Number of answers per BASS.

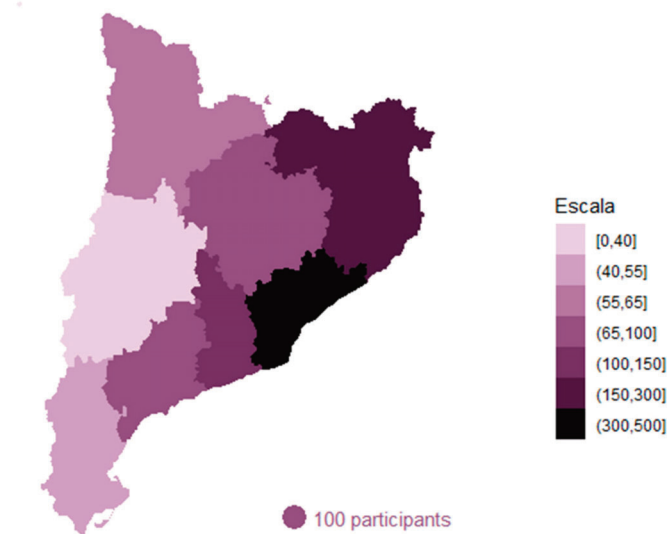


Figure 19. Number of answers per Vegueria.

In the following (Figures 20 and 21) the Age (Figure 20a and Table 7) and gender (Figures 20b and 21) distribution of the sample

Table 7. Extended 5-Number Summary of Age.

Min	Q1	Median	Mean	Q3	Max	StDev	CV
−10	33	43	45.39	56	95	18.316	0.404

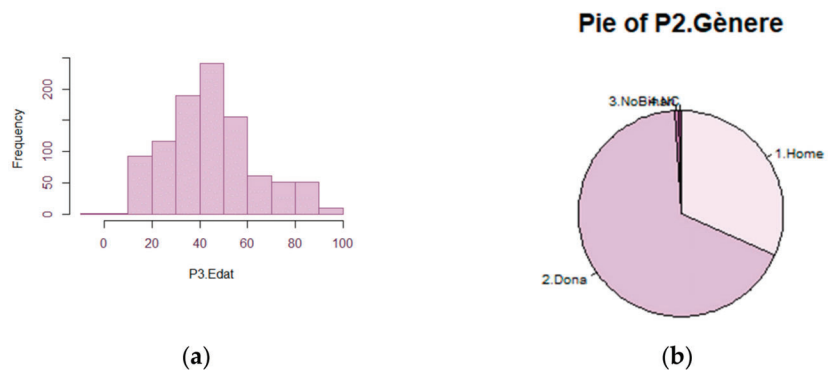


Figure 20. (a) Histogram of Edat; (b) Pie chart of Gender.

### Number of modalities: 4

#### Frequency table

P2.Gènere	Freq.	Prop.	Std. Err
2.Dona (Female)	655	0.675	0.0152
1.Home (Male)	307	0.316	0.0148
3.NoBinari (No binary)	5	0.005	0.0032
4.NC (Unknown)	4	0.004	0.0000

95% CI error:  $\pm 5 \times 10^{-4}$

Std. Error of the question: 0.0107

Figure 21. Frequency table of Gender.

#### 4.1. Economic and Working Impact

**Question L3.1.:** Indicates your personal working category in January 2020, July 2020 and your forecasting for January 2021 (“Indica la teva categoria laboral a gener i juliol de 2020 i quina creus que serà la teva categoria laboral al gener de 2021”)

Responded by the entire sample. Some conclusions are visible in Figure 22, the working Category Frequencies.

- The number of people who do not work and receive no benefits raises a 50%
- The n. of people who do not work and receive some benefit raises by 17.6%
- The number of people who have no job or occupation increases by 11%

**Question L3.2. and L3.3.:** Indicates your personal working situation in January 2020, July 2020 and your forecasting for January 2021 (“Indica la teva situació laboral a gener i juliol de 2020 i quina creus que serà la teva situació laboral al gener de 2021”) See Working situation frequencies in Figure 23.

These two questions provide different modalities for the working situation:

- The number of people who have been licensed (Acomiadat) or folded (Plegat) increases by 78.04%.
- The number of people who reduced their working hours (Reducció) increases by 36.36%.

	L3.1.CategLaboralG20	L3.1.CategLaboralJ20	L3.1.CategLaboralG21
1.Estudiant	84	53	64
2.1aFeina	35	46	83
3.NoTreballaNprest	86	130	73
4.NotreballaSprest	125	147	86
5.Mcasa	39	42	37
6.Jubilat	123	126	129
7.Cap	312	275	287
8.NC	167	152	212

Figure 22. L3.1 Working Category Frequencies.

	L3.3.SituacioLaboralG20	L3.3.SituacioLaboralJ20	L3.3.SituacioLaboralG21
1.Ampliatio	61	44	112
2.Reduccio	55	75	59
3.Acomiadat	36	57	16
4.Plegat	5	16	6
5.Autonom	35	27	27
6.Empressari	<5	<5	<5
7.Cap	575	548	493
8.NC	203	203	257

Figure 23. L3.3. Working Situation Frequencies.

From similar tables made of question L3.2. (1. C indefinit (Permanent contract), 2. CtempActiu (fixed term contract), 3. TreballPerCTemp (intermitent temporal contracts), 4. TeballNregul (irregular working activity), 5. ERT0 (temporal regulation process), 6. RecentCtemp (recent temporal contract initiated), 7. TrobaFeinaFixa (Fix work found) it was found that:

- The number of people who had a business and stopped working during the lockdown or went into bankrupt increased by 110%
- The number of people with non-precarious working conditions (permanent or fixed-term contracts) decreased by 41.62%.
- The 51.25% of people without a job are afraid of not working by January 2021 (the most mentioned reasons are that many companies closed because of COVID19, after a certain age, possibilities to be contracted again decrease, for certain sectors, the people is afraid to be infected by the employee and prefer not to contract new workers).

Regarding Economic Situation

**Question E1:** Economic situation at January 2020 and July, and forecast for January 2021 (“Situació econòmica a gener i juliol de 2020 i previsió per gener de 2021”) See multiple barplot in Figure 10, grid of Pie charts in Figure 11 and proportions in Figure 24 which show question.

**Question E2.:** Did you need to submit for some of the special supports to receive funds to mitigate the problematic created by COVID-19? (“Has necessitat acollir-te a algun dels ajust especials que s’han posat en marxa per mitigar la problemàtica per la COVID-19?”).

- The number of people with economic problems increases a 23.34% (this accounting for those with difficulties to resist the entire month, those with new debts by the end of the month and those that require external economic help to go ahead)

- A 42.8% of them is convinced that they will have economic scarcity by January 2021
- A 62.20% of respondents had some Social Services need
- A 46.1% needed support for food (from them a 64.51% searched it into the BASS)
- A 25.00% needed support to pay the rent of the house (and 51.85% of them searched for it into the BASS)
- A 11.8% asked for Renda mínima garantida (minimum vital rent) and 51.3% of them searched for it in Catalan government
- A 15.7% needed psychologic support, and 51.3% of them searched it into the BASS
- A 51.8% needed a support that implied some economic benefit. Unfortunately, a 70.37% of them did not received the payment by 1st July 2020. Some of them couldn't complete the electronic submission by lack of digital skills, some of them (14.41%) were out of the restrictive eligibility criteria. See in Figure 25 Cross table of E2.AjutsCOVID19 per levels.

	E1.SitEco- nomi- caG20	E1.SitEco- nomi- caJ20	E1.SitEco- nomi- caJ20.1
1.Solvent+	0.216	0.150	0.179
2.Solvent-	0.126	0.114	0.123
3.Dia	0.126	0.099	0.116
4.Justos	0.186	0.182	0.149
5.CalAjut	0.153	0.199	0.146
6.NoArribem	0.109	0.174	0.133
7.NC	0.083	0.081	0.153

Figure 24. Temporal proportions table of variable E1.SitEconomica (economic situation) per levels. Each column represents the observed distribution of the variable E1 in one timestamp.

	E2.1. Ali- mentacio	E2.2. CuraIn- fants	E2.3. Mat- Infor- matic	E2.4. Reparti- ment- Domicili	E2.5. Lloguer Vi- venda	E2.6. Sub- mini- strame nts	E2.7. Tax- esTri- buts	E2.8. Ren- daMi- nim	E2.9. IMV	E2.10. Acollid a	E2.11. Psi- coleg	E2.12. Tele- assis- tencia	E2.13. ERTO iniciat	E2.14. Altres
	Food	Child- ren- Care	ICT mate- rials	Home Delivery	House Rent	Energy Ser- vices Bills	Tax pay- ment s	Mini- mum Vital Rent	IMV	Resi- dence	Psy- cholo- gist	Teles- sis- tance	ERTO initi- ated	Other
1.NoNecessita	423	568	591	570	491	487	572	519	517	569	521	543	557	463
2.Ajuntament	254	65	30	70	109	162	43	18	23	31	68	50	14	15
3.Gencat	21	20	14	8	40	10	5	51	23	3	18	7	1	4
4.Gobierno	6	1	1	1	5	4	3	8	32	3	4	1	12	6
5.Entitats	73	14	7	29	18	26	9	4	3	9	14	7	3	5
6.Altres	17	16	22	21	29	26	23	14	15	15	27	21	12	25
7.NC	132	253	288	254	253	236	303	345	343	333	303	331	365	431

Figure 25. Cross table of E2.AjutsCOVID19 (Subsidies per COVID19) per levels.

Special attention requires the difficulties on life conditions, smoothed by the alarm state, as all eviction processes were interrupted. Nevertheless, they will emerge again in the next months:

- A 27.18% lives in social houses or shares a room in a flat (Question F4. Way of living (Habitatge) from the questionnaire)
- A 27.1% needed support to pay the electricity or gas bills (and 67.30% of them searched help into the BASS)
- A 10.8% needed help to pay taxes

4.2. Social Impact

- A 15.24% are dependent people (Question D1: Do you have some dependency degree (tens algun grau de dependència) of the questionnaire?) See barplot in Figure 26a
- From them a 55.40% refer a worsening dependency process from January 2020 (and a 53.65% of them attributes worsening directly to COVID19) (Question D2: Do you think your dependency level would be different if you would be reevaluated now? Creus que si et valoressin ara tindries una variació en el grau de dependència?) See Cross Table of T1.1 per levels in Figure 26b.

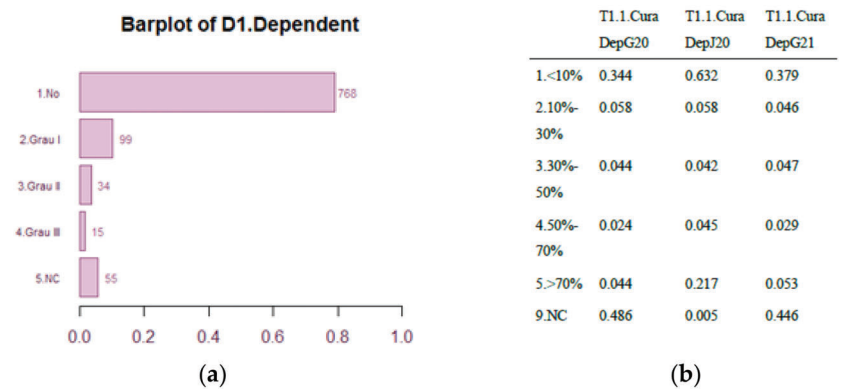


Figure 26. (a) Barplot of D1.Dependent. (b) Cross table of T1.1 per levels.

The questionnaire gets information also from the other side of dependency. The side of the informal caregivers:

- A 16.99% of respondents had dependent people in charge in January 2020
- The number of people with dependents in charge increased by 40.43%

**Question PC2.1:** How many dependent people have you in charge, according to the age? (Quantes persones en Grau I de dependència tens a càrrec en les diferents franges d’edat? (0–11) anys)

This variable has one more complexity level, because dependency is classified in three groups of increasing severity by introducing a fourth variable into the analysis. So, to analyze this item the three variables considered are:

1. Severity of dependency: Ordinal variable with three modalities: Degree I (lower impairment), Degree II and Degree III (higher impairment)
2. Age group: Ordinal variable with 4 modalities determined by experts: Children: (1–11) years, Teenagers: (12–17) years, Adults: (18–69) years, Elderly: more than 70 years.
3. Number of dependent people in charge: discrete variable: (0,1,2 . . . )

Also, the questionnaire includes an entire block dedicated to the use of time. See Multiple stacked barplot of questions PC2, PC3 and PC4 in Figure 27.

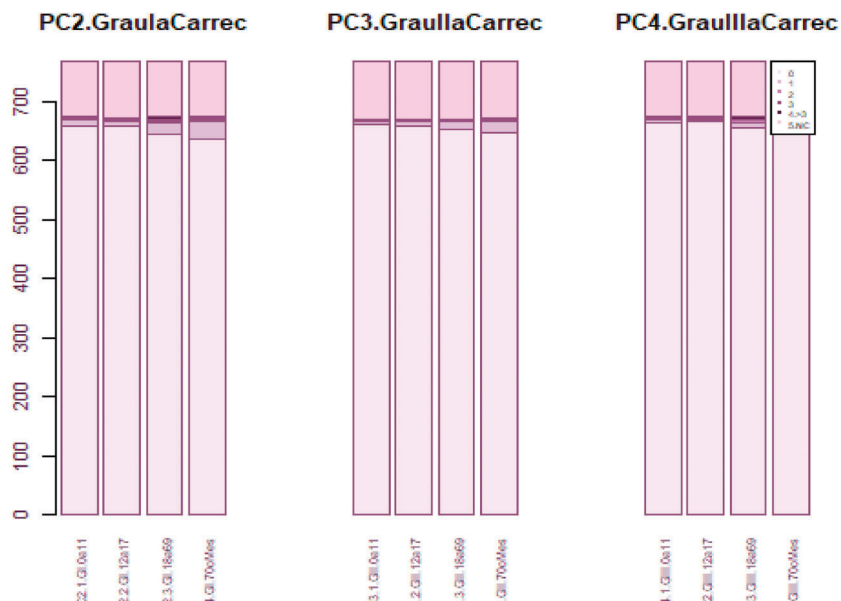


Figure 27. Multiple stacked bar plot of Question PC2-3-4. Dependent people in charge.

For each degree of severity, the inner analysis replicates the structure of the previous question. See multiple barplot of question PC2 with the number of degree I dependent people in charge per age group Figure 28.

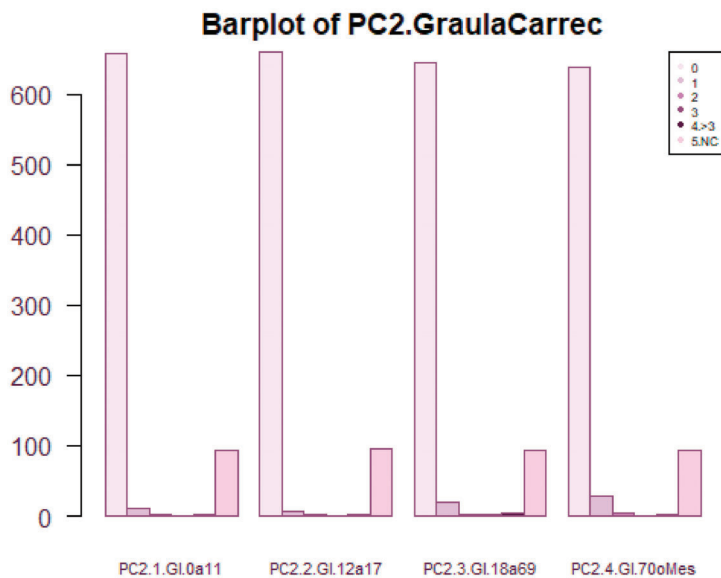


Figure 28. Multiple bar plot or degree I dependent people in charge per age group.

See Cross table of question PC2 with the number of degree I dependent people in charge per age group Figure 29, temporal proportions table in Figure 30 grid of pie charts in Figure 31.

	PC2.1.GI.0a11	PC2.2.GI.12a17	PC2.3.GI.18a69	PC2.4.GI.70oMes
0	658	660	645	638
1	12	6	20	29
2	3	2	2	4
3	0	1	2	1
4.>3	2	3	5	2
5.NC	93	96	94	94

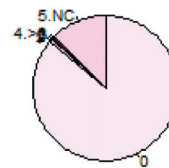
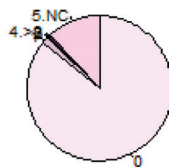
Figure 29. Cross table of PC2.GraulaCarrec per levels.

	PC2.1.GI.0a11	PC2.2.GI.12a17	PC2.3.GI.18a69	PC2.4.GI.70oMes
0	0.857	0.859	0.840	0.831
1	0.016	0.008	0.026	0.038
2	0.004	0.003	0.003	0.005
3	0.000	0.001	0.003	0.001
4.>3	0.003	0.004	0.007	0.003
5.NC	0.121	0.125	0.122	0.122

Figure 30. Temporal proportions of PC2.GraulaCarrec per levels.

**Pie of PC2.1.GI.0a11**

**Pie of PC2.2.GI.12a17**



**Pie of PC2.3.GI.18a69**

**Pie of PC2.4.GI.70oMes**

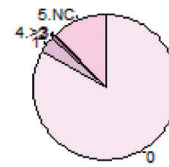
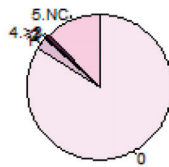


Figure 31. Grid of pie charts of question PC2.1-PC2.4. The variable shows modalities 0 (means 0 dependent persons in charge),1,2,3,4. > 3 (more than 3) and 5.NC (unknown). Since some modalities are so infrequent labels overlap. The exact figures are shown in Figures 29 and 30.



In addition, the questionnaire includes an entire block dedicated to the use of time, from which we can see that:

- A 29.69% of caregivers require now more time to take care of their dependents in charge
- Some caregivers increased required dedication until 5 times more than before pandemics.

**Question R1.** RelUConv: How were on average the relationships in the following environments (“Com eren majoritàriament les relacions que mantenies amb les persones en els diferents àmbits?”).

This is a pack of questions asking for Convivencial unit (Unitat convivencial), Family, Neighbours, Friends, WorkingMates and other. In all of them, the pattern “V^” is observed more or less intensively.

A total of 93 patterns are observed from which 30 can be listed as the others have a too small frequency to be published under guaranty of preserving statistical secrecy. See trajectory map in Figure 9.

See Trajectory frequency table in for question R1: RelConv in Figure 32.

R1.RelUConv	Frequencies		
01.Satisf+01.Satisf+01.Satisf	596	04.Tenses+04.Tenses+01.Satisf	5
10.NC+10.NC+10.NC	64	01.Satisf+03.Igno+01.Satisf	4
02.Preoc+02.Preoc+02.Preoc	33	01.Satisf+09.Inexistents+10.NC	4
09.Inexistents+09.Inexistents+09.Inexistents	25	02.Preoc+04.Tenses+04.Tenses	4
01.Satisf+02.Preoc+02.Preoc	24	04.Tenses+05.Conflic+01.Satisf	4
01.Satisf+02.Preoc+01.Satisf	23	05.Conflic+01.Satisf+01.Satisf	4
02.Preoc+01.Satisf+01.Satisf	14	09.Inexistents+03.Igno+10.NC	4
02.Preoc+02.Preoc+01.Satisf	12	10.NC+01.Satisf+01.Satisf	4
01.Satisf+01.Satisf+02.Preoc	9	01.Satisf+03.Igno+03.Igno	3
01.Satisf+01.Satisf+10.NC	7	01.Satisf+04.Tenses+04.Tenses	3
04.Tenses+04.Tenses+04.Tenses	7	03.Igno+03.Igno+03.Igno	3
09.Inexistents+01.Satisf+01.Satisf	7	04.Tenses+01.Satisf+01.Satisf	3
01.Satisf+04.Tenses+01.Satisf	6	04.Tenses+04.Tenses+03.Igno	3
02.Preoc+01.Satisf+02.Preoc	6	05.Conflic+04.Tenses+01.Satisf	3
05.Conflic+05.Conflic+05.Conflic	6		
03.Igno+01.Satisf+01.Satisf	5	Preserved modalities (frequency less than 3): 63	
		[1] “Number of observations: 971”	

Figure 32. Trajectory frequency table for question R1: Rel UConv.

The automatic report provides the bivariate multiple plot and the frequencies table and the grid of pie charts as well. Here the proportions table is shown. See Figure 33.

	R1.RelUConvG20	R2.RelUConvJ20	R3.RelUConvG21
01.Satisf	0.724	0.684	0.734
02.Preoc	0.083	0.107	0.087
03.Igno	0.009	0.022	0.020
04.Tenses	0.028	0.042	0.020
05.Conflic	0.022	0.020	0.009
06.VioVerb	0.006	0.002	0.001
07.VioPsi	0.005	0.006	0.004
08.VioFis	0.003	0.002	0.000
09.Inexistents	0.045	0.039	0.033
10.NC	0.074	0.076	0.093

Figure 33. Proportions of R1.RelUConv per time.

Figure 12 shows the transition table between January and July 2020 and Figure 34 between July 2020 and January 2021 and one can see which changes in the quality of the relationships are more frequent. During the lockdown a 7.92% of the participants moved from satisfactory relationships with people living in the same home to worse situations

(most of them to worrying relationships or tense), whereas a 4.53% improved their initial relationships to satisfactory.

	01.Satisf	02.Preoc	03.Igno	04.Tenses	05.Conflc	06.VioVerb	07.VioPsi	09.Inexistents	10.NC
01.Satisf	635	15	2	2	0	0	0	3	7
02.Preoc	37	60	3	2	0	0	1	0	1
03.Igno	7	1	8	0	1	0	0	0	4
04.Tenses	16	5	4	14	0	0	0	0	2
05.Conflc	8	2	0	0	7	0	1	0	1
06.VioVerb	0	0	0	0	0	1	0	0	1
07.VioPsi	2	0	0	1	1	0	2	0	0
08.VioFis	2	0	0	0	0	0	0	0	0
09.Inexistents	2	1	1	0	0	0	0	28	6
10.NC	4	0	1	0	0	0	0	1	68

Figure 34. Expected changes July 2020–January 2021.

See changes reported along time in question R1,RelUConv per time in Figure 13a.

See Change patterns in convivial unit in Figure 13b, change patterns in familiar relationships with people living out of home in Figure 14a and change patterns in relationships with neighbours in Figure 14b

See trajectories map in Figure 35 and change patterns in Figure 36a about relationship with friends. See change patterns in labour relations in Figure 36b.

### Evolució de R1.RelAmics al llarg del temps

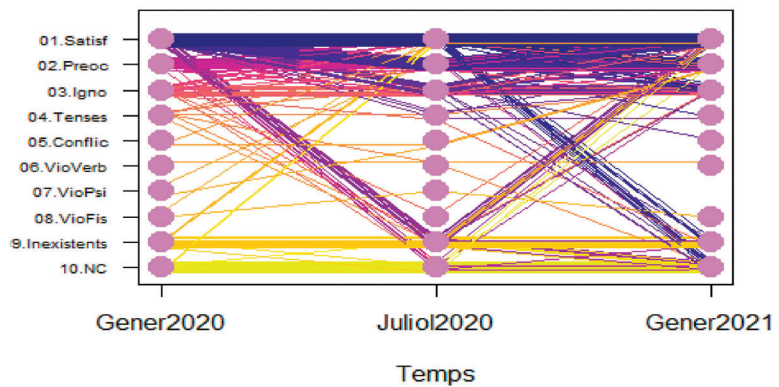


Figure 35. Trajectories map of relationships with friends.

	Freq	Prop		Freq	Prop
Improve	44	0.045	Improve	60	0.062
V pattern	7	0.007	V pattern	4	0.004
Balance	677	0.697	Balance	523	0.539
Λ pattern	5	0.005	Λ pattern	12	0.012
Enworse	84	0.087	Enworse	64	0.066

(a)

(b)

Figure 36. (a) Changing patterns in relationships with friends. (b) Changing patterns in labour relationships.

The “V^” pattern appears again here, with certain proportion of people that behaves more links with other people during the pandemics, and those that feel more isolated

**Question Soc4.:** The pandemics created links with other people (family, friends, neighbours, etc.)? La pandèmia: T’ha creat vincles d’unió amb altres persones (família, amistsats, veïnatge, . . . ). Barplot of isolation feelings during pandemics in Figure 37a and intensification of links in Figure 37b. Figure 38a shows frequency table of isolation feelings and Figure 38b intensification on links.

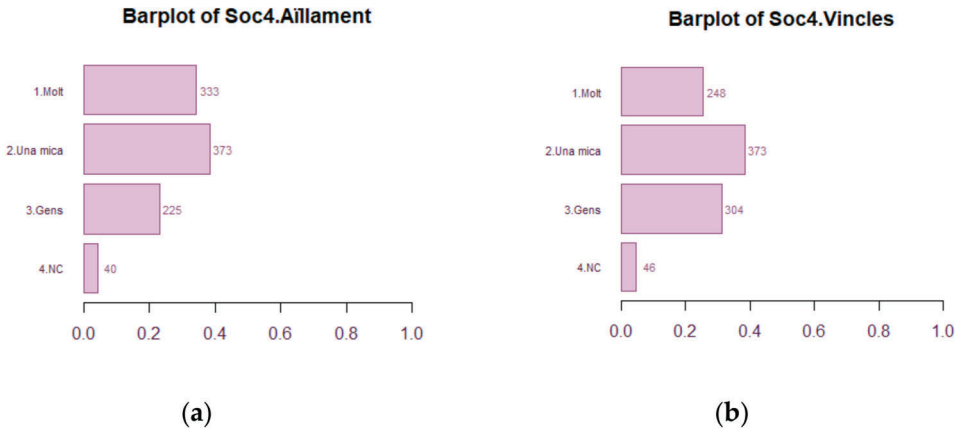


Figure 37. (a) Barplot of isolation feelings during pandemics (b) Barplot of intensification of links due to pandemics.

Soc4.Aïllament	Freq.	Prop.	Std. Err
2.Una mica	373	0.384	0.0155
1.Molt	333	0.343	0.0152
3.Gens	225	0.232	0.0134
4.NC	40	0.041	0.0063

95% CI error:  $\pm 5 \times 10^{-4}$

**Std. Error of the question: 0.0131**

(a)

Soc4.Vincles	Freq.	Prop.	Std. Err
2.Una mica	373	0.384	0.0155
3.Gens	304	0.313	0.0148
1.Molt	248	0.255	0.0141
4.NC	46	0.047	0.0071

95% CI error:  $\pm 5 \times 10^{-4}$

**Std. Error of the question: 0.0133**

(b)

Figure 38. (a). Frequency table of isolation feelings during pandemics (b) Frequency table of intensification of links due to pandemics.

**Question:** Do you feel or have you felt alone? (T’has sentit o et sents sol?).

Results are shown in different figures. See trajectory map of loneliness feelings in Figure 39, Trajectories frequency table in Figure 40, multiple barplot in Figure 41a and proportions per level in Figure 41b. See Grid of pie charts in Figure 42. Changes January 2020–July 2020 are shown in Figure 43a and see planned changes in July 2020–January 2021 in Figure 43b.

### Evolució de Soc5.SolG20 al llarg del temps

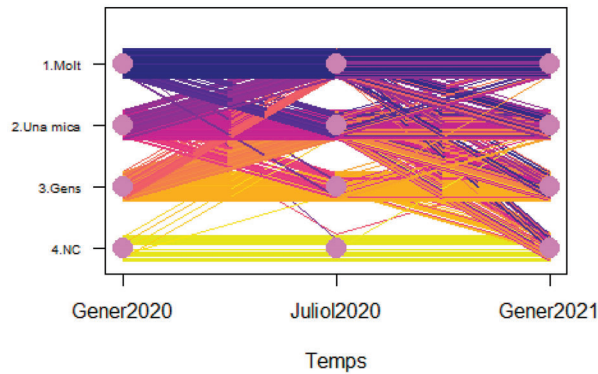


Figure 39. Trajectory map of loneliness feelings.

Soc5.SolG20	Frecuencies		
3.Gens+3.Gens+3.Gens	277	3.Gens+1.Molt+1.Molt	13
2.Una mica+2.Una mica+2.Una mica	142	3.Gens+1.Molt+2.Una mica	12
1.Molt+1.Molt+1.Molt	80	2.Una mica+2.Una mica+1.Molt	9
3.Gens+2.Una mica+2.Una mica	59	1.Molt+1.Molt+4.NC	8
3.Gens+2.Una mica+3.Gens	45	3.Gens+3.Gens+2.Una mica	8
2.Una mica+1.Molt+1.Molt	37	1.Molt+1.Molt+3.Gens	7
2.Una mica+1.Molt+2.Una mica	32	3.Gens+1.Molt+3.Gens	7
4.NC+4.NC+4.NC	31	3.Gens+2.Una mica+1.Molt	7
2.Una mica+2.Una mica+3.Gens	26	1.Molt+2.Una mica+1.Molt	6
2.Una mica+3.Gens+3.Gens	24	1.Molt+2.Una mica+3.Gens	6
3.Gens+2.Una mica+4.NC	21	1.Molt+2.Una mica+4.NC	5
2.Una mica+2.Una mica+4.NC	19	2.Una mica+1.Molt+4.NC	5
3.Gens+3.Gens+4.NC	17	3.Gens+1.Molt+4.NC	5
1.Molt+1.Molt+2.Una mica	16	1.Molt+3.Gens+3.Gens	4
1.Molt+2.Una mica+2.Una mica	15	2.Una mica+3.Gens+2.Una mica	3
2.Una mica+1.Molt+3.Gens	13	2.Una mica+3.Gens+4.NC	3

Figure 40. Trajectories Frequency table of trajectories with frequency greater than 3.

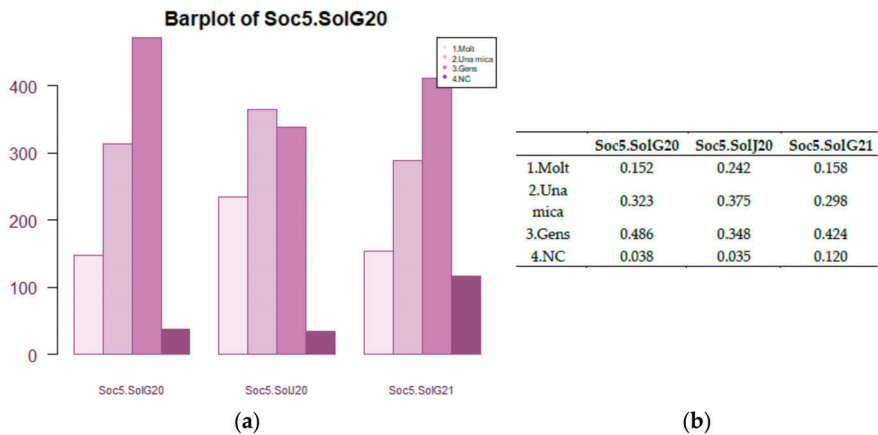
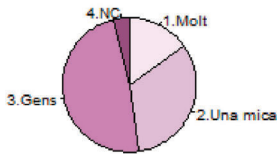
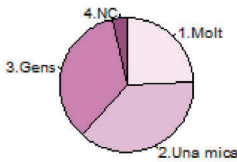


Figure 41. (a) Multiple bar plot of loneliness feelings (b) Proportions of Soc5.SolG20 per levels.

Soc 5.SolG20



Soc5.SolJ20



Soc5.SolG21

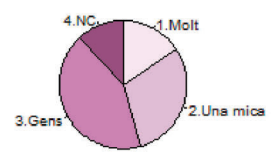


Figure 42. Grid of pie charts for loneliness feelings.

	1.Molt	2.Una mica	3.Gens	4.NC
1.Molt	111	32	4	1
2.Una mica	87	196	30	1
3.Gens	37	132	303	0
4.NC	0	4	1	32

(a)

	1.Molt	2.Una mica	3.Gens	4.NC
1.Molt	130	60	27	18
2.Una mica	22	217	78	47
3.Gens	1	11	306	20
4.NC	0	1	1	32

(b)

Figure 43. (a) Changes January 2020–July 2020 (b) Planned changes July 2020–January 2021.

- A 72.7% of respondents felt more isolated
- The loneliness feeling increases by 29.3%. This is a pattern mainly followed by women (70%) more than 60 years old (on average), living alone, with some lack of digital skills and a 52.18% of them requiring emotional support during pandemics.
- A 41.45% of participants required psychological support due to COVID19 and from them a 30.95% required emotional support.
- A 23.58% of participants referred some mental disorder in January 2020. From them: a 96.94% received pharmacological treatment. Among people with mental disorders, 46.28% suffer from depression and 58.51% suffer from anxiety disorder.
- The 73.78% of people with mental disorders declared to feel worse in July 2020
- A 68.86% of people with depression declares to feel worse in July 2020
- A 72.3% of people with anxiety declare to feel worse by July 2020
- A 5.12% of people without mental disorders in January 2020 declare feeling worse in July 2020
- A 57.93% of people with disabilities feel worse and 58.33% of them attributes worsening to the COVID19

Moreover, a 49.64% of participants have teleworked or have followed online training during the pandemics, while only a 5.19% of them already did tele-activities in January 2020. In July 2020, a 21.84% of people involved in tele-activities (working or education) suffered the impact on care activities (relatives, elderly, children . . . ). A 54.4% of them required emotional support. See frequency table of Mental Disorders in Figure 44a and Marginal barplot of Mental Disorders in Figure 44b.

Number of modalities: 6

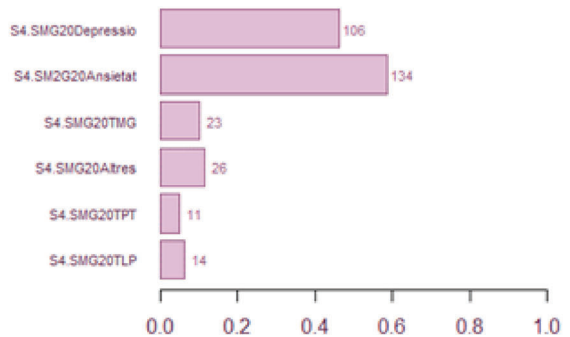
Freq. table Problemes de Salut Mental

Salut Mental	Freq.	Prop.	95% CI error
S4.SM2G20Ansietat	134	0.585	0.0326
S4.SMG20Depressio	106	0.463	0.0330
S4.SMG20Altres	26	0.114	0.0210
S4.SMG20TMG	23	0.100	0.0197
S4.SMG20TLP	14	0.061	0.0158
S4.SMG20TPT	11	0.048	0.0141

95% CI error: ± 0.0021; SE: 0.0239

(a)

Barplot of Salut Mental



(b)

Figure 44. (a) Frequency table of Mental disorders. (b) Marginal bar plot of Mental Disorders.

### 4.3. Violence

Among the options to choose for the quality of the relationships at different environments (questions R1 to R9 of the questionnaire), particular options asked if the person is being object of violence, either physic emotional or psychic. In total, a 6.38% of the respondents declare to be victims of some form of violence. From them, a 72.58% are women. Civil status, profession and academic level are transversal among these group (17.4% have university studies). A common characteristic of these people is that 90.33% of people have working precariat (no stable or temporal contract, but other irregular ways of work or unemployment). The questionnaire poses two additional questions to get more details about the pattern of the aggressor and the forces balance with the victim.

**Question R4.** Aggressor: If you have been object of violence, who performs it? (“Si has indicat ser objecte de violència, qui exerceix aquesta violència?”) See frequency table of kind of aggressor in Figure 45a and bar plot showing who is the aggressor in Figure 45b. See the frequency table of R4 .Agessor in Figure 46.

R4.Agressor	Freq.	Prop.	Std. Err
1.NoViolencia	775	0.798	0.0130
5.NC	125	0.129	0.0110
3.Igual	58	0.060	0.0077
2.Superior	27	0.028	0.0055
4.Subaltern	8	0.008	0.0032

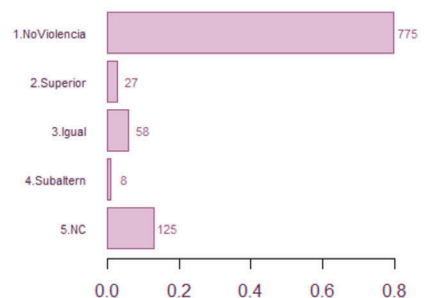
Table 29: Frequency table of kind of aggressor

95% CI error: ± 5 × 10<sup>-4</sup>

Std. Error: 0.0088

(a)

Barplot of R4.Agressor



(b)

Figure 45. (a) Frequency table of kind of aggressor (b) Who is the aggressor?

R4.Agressor	Freq.	Prop.	Std. Err
1.NoViolencia	764	0.787	0.0130
5.NC	117	0.120	0.0105
3. Un igual (germà, company, amic, veí...)	48	0.049	0.0071
2. Un superior o ascendent (pare, tiet, responsable de feina, professor...)	20	0.021	0.0045
1.NoViolencia;5.NC	7	0.007	0.0032
4. Un subaltern o descendent (fills, empleats, ...)	5	0.005	0.0032
2. Un superior o ascendent (pare, tiet, responsable de feina, professor...);3. Un igual (germà, company, amic, veí...)	4	0.004	0.0000

95% CI error:  $\pm 5 \times 10^{-4}$  ; Std. Error of the question: 0.0056

#### Preserved modalities (frequency less than 3): 5

Figure 46. Frequency table of Who is the aggressor.

#### 4.4. Synthesis of Remaining Results

In the following, we synthesize the results of applying the automatic intelligent scripts to the entire dataset.

##### 4.4.1. Economic and Working Impact

- The number of people who do not work and receive no benefits raises a 50%
- The number of people who do not work and receive some benefit raises by 17.6%.
- The number of people who have no job or occupation increases by 11%
- The number of people who have been licensed or folded raises by 78.04%
- The number of people who reduced their working hours increases by 36.36%.
- The number of people who had their own business and stopped working during the lockdown or went into bankrupt increased by 110%
- The number of people with non-precarious working conditions decreased by 41.62%.
- The 51.25% of people without a job are afraid of not working yet by January 2021 (the most mentioned reasons are that many companies closed because of COVID19, after a certain age possibilities to be contracted again decrease, for certain sectors, the people is afraid to be infected by the employee and prefer not to contract new workers).
- The 51.25% of people without a job are afraid of not working yet by January 2021
- The number of people with economic problems increases a 23.34%.
- A 42.8% of them is convinced that they will have economic scarcity by January 2021
- A 62.20% of respondents had some Social Services need
- A 46.1% needed support for food (from them a 64.51% searched it into the BASS)
- A 25.00% needed support to pay the rent of the house (and 51.85% of them searched for it into the BASS)
- A 11.8% asked for Renda mínima garantida (minimum vital rent) and 51.3% of them searched for it in Catalan government
- A 15.7% needed psychologic support, and 51,3% of them searched it into the BASS
- A 51.8% needed a support that implied some economic benefit. Unfortunately, a 70.37% of them did not received the payment by July 1st, 2020. Some of them couldn't complete the electronic submission by lack of digital skills, some of them (14.41%) were out of the restrictive eligibility criteria



- A 27.18% live in social houses or share a room in a flat
- A 27.1% needed support to pay the electricity or gas bills (and 67.30% of them searched help into the BASS)
- A 10.8% needed institutional help to pay taxes and tributes
- A 51.8% of participants submitted applications involving economic support to the administrations during the first wave
- The 70.37% of applicants for economic support did not receive a penny by July 1st 2020

(They mention a variety of reasons among which we can highlight the delay on resolutions, the difficulties to make the submission, the impact of digital gap of making the digital application, the restrictive eligibility criteria that left excluded a 14.41% of the people that declare to need the support).

#### 4.4.2. Social Impact

- A 67.5% of the participants to INSESS-COVID19 project are women.
- A 15.24% of the participants are dependent people
- From them a 55.40% refer a worsening dependency process from January 2020 (and a 53.65% of them attributes worsening directly to COVID19)
- A 16.99% of respondents had dependent people in charge in January 2020
- The number of people with dependents in charge increased by 40.43% by July
- An 18.02% of the respondents didn't declare dedication to dependent people in January and declared to dedicate more than 70% of their daily time to this matter by July
- A 72.7% of respondents felt more isolated
- The loneliness feeling increases a 29.3% and this is a pattern mainly followed by women (70%) more than 60 years old (on average), living alone, with some lack of digital skills and a 52.18% of them requiring emotional support during pandemics.
- A 41.45% of participants required psychological support due to COVID19 and from them a 30.95% required emotional support.
- The 73.78% of people with mental disorders declared to feel worse in July 2020
- A 68.86% of people with depression declares to feel worse in July 2020
- A 72.3% of people with anxiety declare to feel worse by July 2020
- A 5.12% of people without mental disorders in January 2020 declare feeling worse in July 2020
- A 57.93% of people with disabilities feel worse and 58.33% of them attributes worsening to the COVID19
- A 54.4% of the people making teleworking or tele-education during the pandemics required emotional support

#### 4.4.3. Violence

- In total, a 6.38% of the respondents declare to be victims of some form of violence.
- From them, a 72.58% are women.
- Civil status, profession and academic level are transversal among these group (17.4% have university studies).
- A 90.33% of people are under a working scarcity.
- Often the person is victim of more than one aggressor simultaneously: 14.51% receive violence from two profiles of aggressors simultaneously; 27.42% from three
- In a 93.54% of the cases, the aggressor shares an equality relationship with the victim, being a neighbor, a friend, a brother . . .
- In a 45.55% of the cases the aggressor has a power relationship with the victim, being a father, a boss, a professor, etc.

### 5. Discussion

In 2020 the crisis of the Covid-19 is impacting segments of population that were already affected by the previous crisis, being the job area, the working class, and the young population the segments more punished again. Nevertheless, not only that, as we shall see.

INSESS-COVID19 shows as some of the most relevant impacts between January and July 2020 the economic indicators listed in Section 4.4.1. Economic and Working Impact Among the working class, there is a pessimism that merits attention. People is worried about their future and an important part of them think this situation will not improve neither in the short or midterm.

Uninterrupted decrease of incomes of many people and families is acceleration the increment of poverty risk in Catalan society, for both moderate and extreme poverty. This has raised the demands of need to public social services and third sector entities, as well as help and economic support submission applications.

Special attention requires the difficulties about life conditions, smoothed by the alarm state. The sudden impoverishment of wide segments of population will have a delayed effect on the lack of capacities to pay taxes, bills of domestic services like gas or electricity, the house rent, or the bank quotes for loans and mortgages. The alarm state declared by the government has interrupted all eviction processes. Nevertheless, they will emerge again in the next months, as soon as alarm state is abolished, and these processes will be reactivated in a much worse context than when they were interrupted.

In fact, the ERTA (temporal regulation of occupation procedures), institutional helps to self-employed people and other economic funds provided by the governments contributed somehow to smooth the economic effects of the pandemics, inefficiencies and delays on the management and resolution of applications sensibly diminished the positive impact they could have had.

Until here, the main conclusion is that the COVID19 has raised a crisis that impacts on population segments already punished by the previous crisis on 2008, which were not recovered yet, thus creating an amplified impact towards poverty and social vulnerability in many critical needs, like housing or work.

However, as said before, there is something new in the COVID19 crisis, that was not observed in previous crises and that worsen even more the social vulnerabilities of the people. As it has been seen in Section 4.4.2. Social Impact, where indicators are showed. The COVID19 crisis is also a social crisis and a crisis of relationships, and it is impacting as well to other population segments different from those affected by the 2008 crisis, as a consequence of the new social vulnerabilities emerged from social distance measures that pandemics management required: Restricted mobility, home lockdown, social isolation, teleworking, accelerated digital transformation, interruption and delay of court and administrative processes, etc. These measures caused serious impacts to women and elderly.

In addition, this is in seriously different from official common figures. Indeed, according to CCI2018, a 58.5% of the users of Social Services are women. The proportion of women in the INSESS-COVID is significantly higher. In addition, this points out to a bigger impact of the pandemics on the vulnerability of women, provided that gender was not a criterion used in any of the 20 target profiles defined to participate in the project. So, when BASS were finding people following those 20 profiles, it happened that most of them were satisfied by women. Indeed, women assumed a heavy load in the worse periods of the pandemics, informal caregivers tend to be women, single-parent families, women in charge of dependent people, children or people with disability or mental disorders. Nevertheless, not only, most of the health and social services professional profiles strongly stressed during the pandemics tend to be female works as well. Finally, many widows are also women, so old women leaving alone are also seriously impacted by isolation, loneliness and dependency issues during the pandemics.

On the other hand, mobility restrictions and the high vulnerability to COVID-19 of elderly people caused serious confinement-related impacts to this segment of population: loneliness, isolation, depression, digital gap, etc.

The questionnaire gets information also from the other side of dependency. The side of the informal caregivers

Regarding social relationships and participation, in all areas, working, familiar, friendships... the pattern “V^” is observed for the confinement period involving the double dynamics of:

- Strengthening the links, increasing solidarity and intensifying participation, even in voluntary activities
- Disconnecting and isolating from relatives, friends, neighbors, colleagues

Thus, many segments of population required psychological and emotional support. The loneliness feelings, isolation and mental impairment raised in many people, especially in elderly.

Finally, the violence has also been present during the pandemics as show violence indicators in Section 4.4.3 Violence

The questionnaire also includes information about the digital gap and the interruption/delay or court and administrative processes (divorces, regularization, evictions, etc.). The most impacted groups are women in different forms, and one of the more impressive patterns is that of women victims of violence, who had to pass the confinement at home together with the aggressor while divorces or restraining orders were interrupted in court.

As a main advantage of the proposed methodology, we are providing a tool for direct participation that can provide access to query citizens (or professionals whenever needed) in quick times, and process collected data very quickly. The questions can be adapted to each application experience, and the analysis will keep in automatic, provided that the Metainformation csv file is provided together with dataset. In addition, a special design of the questionnaire can solve the small number of responses by substitution of respondents that can delay along time without losing validity of dataset a long time. Finally, the results are offered in form or working document in Word that is radically unshortening the capacity of using this results in strategic meetings immediately after data download from digital questionnaire.

As all studies based on citizens data, results depend on the truth of the answers provided by participants.

Time required to complete the study depends on the celerity of BASS searching participants, and response time of the participants. Face-to-face workshops proposed in INSESS-COVID19 were designed to mitigate the time required for data collection step and pilots proved their effectiveness, even if the pandemics constrained to work under the “free” modality. Limitations to celebrate workshops as originally designed were solved by developing new modalities for the workshops. Counterpart, free open workshop increases coverage but loses control of time to provide the response. In any case, the technology developed and the statistical new descriptive tools proposed perform well and quickly and they will provide much valuable results when answering the questionnaire becomes mandatory (this depends on the topic of the consultation).

The proposed methodology constitutes a powerful tool to disclose the underlying patterns of social vulnerability in Catalan territory, but is still not providing predictions on Social needs of population in the current months. Once the patterns have been discovered, the predictive model for the specific patterns will become reachable with the next step of the analysis and classifier techniques will be used.

Data comes from the entire Catalan Territory, but selected participants identified by the BASS are not obliged to answer, so that, some of them might skip the commitment and generate poor data from some BASS. In addition, sample heterogeneity between territories might appear. However, this heterogeneity is associated with intrinsic characteristic of the territory itself, so, it is not necessarily wrong. However, this can be compensated by calling new substitute participants with same profiles and their answers will be valid thanks to the introduction of temporal variables included in the questionnaire, even if he/she was answering the questions with important delay.

## 6. Conclusions

The impact of the COVID19 crisis on the Social Services along the territory involves two related dimensions: the impact on people in need of social care, and the impact on the praxis of social services professional teams. The COVID19 crisis appeared in a moment when, as a society, we had not fully recovered yet from the economic crisis started at 2008, and has added an additional burden to social services from all over the country that are already far overflowed.

This pandemic came when our technological maturity was less advanced than we, as a society, would have liked. We still cannot use the data as an immediate asset in crisis management.

The INSESS-COVID19 project has developed a technological tool suitable for collecting fresh and direct information from citizens almost immediately, and ready to analyze it with generic and automated processes that can be very helpful to deal with new and unexpected situations, like the management of emergencies and disruptions (as COVID19 was). Data Science and Artificial Intelligence are the disciplines that enable it. We stopped collecting data just one week before public presentation of results in front of the government. Everything was processed in a very short time. INSESS-COVID19 solves the extraction of added value from data in a very short time. The single current limitation is now the availability of data, the time taken by the citizen to get involved, to participate, to answer the questionnaire, the time required by the administration to think about which questions need to be addressed, to whom and when. The analysis of the 971 answers collected from citizens belonging to the 20 target profiles and distributed all over entire Catalonia, and including as well some individuals that had never been users of Social Services, but they start to be after the pandemics first round.

An innovative methodology to collect, analyze and report this data is proposed. It helps to see what is happening, to understand the main trends in different parts of the territory and to identify the relevant indicators for future studies where predictive models of the identified relevant socioeconomic parameters can be analyzed to build predictive models helping the decision-makers to anticipate. In addition, our results are producing the key inputs to be used in simulators for Decision Support.

The methodology is flexible to work upon other questionnaires and other subpopulation of respondents with minimum modifications required. It provides a new tool to get relevant decisional knowledge for decision-making support at many different decision-making levels, from most operative, to most strategic, including policy-making support.

The INSESS-COVID19 project shows how the COVID19 crisis is impacting, on the one hand, on same populations segments that were already damaged of the last economic crisis (2008) and, on the other, it impacts on new groups, rising emerging social needs that also demand the attention of Social Services: women and elderly.

The economic slowdown that began in 2008 affected the labour market in a particularly negative way and generated a double process of impoverishment due, on the one hand, to falling incomes and rising inequality in their job distribution and, on the other, the collapse of lower incomes. This situation limited the opportunities that individuals and families had to resolve their economic difficulties and increased existing social differences. Rising unemployment, prolonged unemployment, precarious wages, job discontinuity or the low purchasing power of retirement pensions weakened family economies, and this increased the problems and complexity of social situations, and it accentuated the processes of social exclusion of individuals and families. Among the most affected profiles were those who had lost their jobs, unemployed young people looking for their first job, young families with dependent children, single women with family responsibilities, single men without a home, elderly women with pensions. non-contributory and irregular immigrants.

In this work, a specific methodology to guarantee that the analysis of the territorial data is preserving the statistical secrecy for minority subpopulations is proposed, so that the information can be used for the decisions without risks of revealing the identity of participants.

In the future work, in-depth analysis is in progress to find BASS with similar profiles that admit a common local report that preserves statistical secrecy while providing specific information to the BASS. Clustering of BASS through multiview clustering techniques and semi-automatic knowledge-based dimensionality reduction techniques are being used to characterize BASS with a synthesis indicator of each block of information so that a better perspective of territorial similarities is elicited, and the real policy strategies can be descended at territorial level.

**Author Contributions:** K.G. was in charge of the conceptualization and methodology; both K.G. and X.A. were developing the software; validation was led by K.G. with BASS from different places in Catalonia; K.G. in charge of formal analysis; K.G. and X.A. developed the investigation; X.A. was in charge of resources; data curation, was led by K.G. and done by X.A.; writing—original draft preparation, K.G.; writing—review and editing, K.G. and X.A.; visualization, K.G. and X.A.; supervision, K.G.; project administration, K.G.; funding acquisition, K.G. All authors have read and agreed to the published version of the manuscript.

**Funding:** UPC funded this project with a special call that Development Cooperation Centre (CCD) of the UPC opened on the occasion of the emergency generated by the COVID-19 crisis.

**Institutional Review Board Statement:** The study was conducted according to the guidelines of the Declaration of Helsinki, and approved by the Institutional Review Board (or Ethics Committee) of IDEAL.

**Informed Consent Statement:** Informed consent was obtained from all subjects involved in the study.

**Data Availability Statement:** As the research is using personnel data from vulnerable persons, data is hosted in a UPC server compliant with current RGPD and individual data is not available out of the project, as it was agreed with the respondents. Only aggregated data is provided.

**Acknowledgments:** Authors want to express special thanks to INSESS-COVID19 team, specially to Toni Codina from iSocial for his unselfish dedication to the project. In addition, to all participant citizens that shared with the project their situation and problematics. To the Social Services professionals who gave an initial impetus to the project by organizing the pilots (in Platja d’Aro and La Noguera) or creativity to find new formulas for citizen involvement that enabled participation in the project despite the BASS overflow. To the follow-up committee (Meritxell Benedí, Mireia Mata, Montserrat Dolz, Miquel Angel Manzano, Albert Cònsola, Sònia Oriola, Rafael Cuenca i Sílvia Madrid) and to the institutions of the advisory board for their support to the project: General Directorate of Social Services and the General Directorate of Equity (GenCat, Catalan Government); the Diputació de Barcelona; the Barcelona Metropolitan Area; the Catalan Association of Municipalities and Counties; and the Federation of Municipalities of Catalonia. To Miquel Sastre, Yaroslav Hernandez, Paula Pedrós, Carles Alsinet, Montse Torredelot, Massimiliano Giacalone, Sergi Ramirez, Cervemakers, Institut de Cervelló, Social Services Teams and Culture councillor from Mollet del Vallès.

**Conflicts of Interest:** The authors declare no conflict of interest.

## References

1. La Gatta, V.; Moscato, V.; Postiglione, M.; Sperli, G. An Epidemiological Neural network exploiting Dynamic Graph Structured Data applied to the COVID-19 outbreak. *IEEE Trans. Big Data* **2020**, *7*, 45–55. [CrossRef]
2. Gibert, K.; Horsburgh, J.S.; Athanasiadis, I.N.; Holmes, G. Environmental data science. *Environ. Model. Softw.* **2018**, *106*, 4–12. [CrossRef]
3. Gibert, K.D. Conti aTLP: A color-based model of uncertainty to evaluate the risk of decisions based on prototypes. *Artif. Intell. Commun.* **2015**, *28*, 113–126.
4. EUR-LEX. Available online: <https://eur-lex.europa.eu/eli/reg/2016/679/2016-05-04> (accessed on 15 February 2021).
5. Fayyad, U.; Gregory, P.-S.; Padhraic, S. The KDD process for extracting useful knowledge from volumes of data. *Commun. ACM* **1996**, *39*, 27–34. [CrossRef]
6. Royal Society. Explainable AI. Available online: <https://royalsociety.org/topics-policy/projects/explainable-ai/> (accessed on 14 March 2021).
7. DIXIT Centre de Documentació de Serveis Socials. Available online: [https://dixit.gencat.cat/ca/detalls/Noticies/tsf\\_presenta\\_eina\\_cribratge\\_ajudar\\_identificar\\_gestionar\\_casos\\_socials\\_complexos.html](https://dixit.gencat.cat/ca/detalls/Noticies/tsf_presenta_eina_cribratge_ajudar_identificar_gestionar_casos_socials_complexos.html) (accessed on 15 February 2021).

8. Pla Estratègic de Serveis Socials. Available online: [https://treballiafersocials.gencat.cat/web/.content/03ambits\\_tematicas/15\\_serveissocials/pla\\_estrategic\\_serveis\\_socials/Pla\\_estrategic\\_serveis\\_socials\\_catalunya\\_NOU/01\\_Plana\\_principal/1.-2020-12-29-Pla-estrategic-de-serveis-socials-2021-2024.pdf](https://treballiafersocials.gencat.cat/web/.content/03ambits_tematicas/15_serveissocials/pla_estrategic_serveis_socials/Pla_estrategic_serveis_socials_catalunya_NOU/01_Plana_principal/1.-2020-12-29-Pla-estrategic-de-serveis-socials-2021-2024.pdf) (accessed on 15 February 2021).
9. Lauriks, S.; Buster, M.C.A.; de Wit, M.A.S.; van de Weerd, S.; Tigchelaar, G.; Fassaert, T. The Dutch Version of the Self-Sufficiency Matrix (SSM-D). 2012. Available online: <http://www.self-sufficiencymatrix.org> (accessed on 15 February 2021).
10. The Self-Sufficiency Standard. Available online: <https://depts.washington.edu/selfsuff/standard.html> (accessed on 15 February 2021).
11. Brooks, J.; Pearce, D. Meeting needs, measuring outcomes: The self-sufficiency standard as a tool for policy-making, evaluation, and client counseling. *Clgh. Rev.* **2000**, *34*, 34.
12. Gibert, K.; Nonell, R.; Velarde, J.M.; Colillas, M.M. Knowledge Discovery with clustering: Impact of metrics and reporting phase by using KCLASS. *Neural Network World* **2015**, *15*, 319–326.
13. Barometre del Tercer Sector Social del 2017. Available online: <http://www.tercersector.cat/el-sector-catalunya> (accessed on 15 February 2021).
14. Krejcie, R.V.; Morgan, D.W. Determining sample size for research activities. *Educ. Psychol. Meas.* **1970**, *30*, 607–610. [CrossRef]
15. Moore, D.; McGabe, G.P.; Craig, B.A. *Introduction to the Practice of Statistics*; H. Freeman: New York, NY, USA, 1993.
16. Gibert, K.; García Rudolph, A.; Curcoll, L.; Soler, D.; Pla, L.; Tormos, J.M. Knowledge discovery about quality of life changes of spinal cord injury patients: Clustering based on rules by states. *Stud. Health Technol. Inform.* **2009**, *150*, 579–583. [PubMed]
17. Gibert, K.; Rodriguez Silva, G.; Rodriguez Roda, I. Knowledge discovery with clustering based on rules by states: A water treatment application. *Environ. Model. Softw.* **2010**, *25*, 712–723. [CrossRef]
18. Di Meglio, E.; Osier, G.; Berger, Y.G.; DiFalco, E. Standard Error Estimation for EU-SILC Target Indicators—First Results of the Net-SILC2 Project. 2013. Available online: <https://eprints.soton.ac.uk/354892/> (accessed on 15 February 2021).
19. Encuesta de Población Activa. Diseño de la Encuesta y Evaluación de la Calidad de los Datos. Informe Técnico. Available online: [https://www.ine.es/inebaseDYN/epa30308/docs/epa05\\_disenc.pdf](https://www.ine.es/inebaseDYN/epa30308/docs/epa05_disenc.pdf) (accessed on 15 February 2021).
20. Encuesta de Población Activa, Metodología 2005. Descripción general de la Encuesta. Available online: <https://www.ine.es/inebaseDYN/epa30308/docs/resumetepa.pdf> (accessed on 24 March 2021).
21. EPA. Available online: [https://www.ine.es/dyngs/INEbase/es/operacion.htm?c=Estadistica\\_C&cid=1254736176918&menu=ultiDatos&idp=1254735976595](https://www.ine.es/dyngs/INEbase/es/operacion.htm?c=Estadistica_C&cid=1254736176918&menu=ultiDatos&idp=1254735976595) (accessed on 15 February 2021).





Article

# Comparison of Dengue Predictive Models Developed Using Artificial Neural Network and Discriminant Analysis with Small Dataset

Permatasari Silitonga <sup>1</sup>, Alhadi Bustamam <sup>1,\*</sup>, Hengki Muradi <sup>2</sup>, Wibowo Mangunwardoyo <sup>3</sup> and Beti E. Dewi <sup>4</sup>

<sup>1</sup> Department of Mathematics, Faculty of Mathematics and Natural Sciences, Universitas Indonesia, Kampus Baru UI, Depok 16424, Indonesia; permatasari.silitonga@sci.ui.ac.id

<sup>2</sup> Department of Mathematics, Faculty of Science and Information Technology, Institut Sains dan Teknologi Nasional, Jl. Moh Kahfi II Srengseng Sawah Jagakarsa, Jakarta Selatan 12640, Indonesia; hengki.muradi@sci.ui.ac.id

<sup>3</sup> Department of Biology, Faculty of Mathematics and Natural Sciences, Universitas Indonesia, Kampus Baru UI, Depok 16424, Indonesia; wibowo.mangun@ui.ac.id

<sup>4</sup> Department of Microbiology, Faculty of Medicine, Universitas Indonesia, Jl. Salemba Raya no. 5, Kota Jakarta Pusat, Daerah Khusus Ibu Kota Jakarta 10430, Indonesia; beti.ernawati@ui.ac.id

\* Correspondence: alhadi@sci.ui.ac.id

**Abstract:** In Indonesia, dengue has become one of the hyperendemic diseases. Dengue consists of three clinical phases—febrile phase, critical phase, and recovery phase. Many patients have died in the critical phase due to the lack of proper and timely treatment. Therefore, we developed models that can predict the severity level of dengue based on the laboratory test results of the corresponding patients using Artificial Neural Network (ANN) and Discriminant Analysis (DA). In developing the models, we used a very small dataset. It is shown that ANN models developed using logistic and hyperbolic tangent activation function with 70% training data yielded the highest accuracy (90.91%), sensitivity (91.11%), and specificity (95.51%). This is the proposed model in this research. The proposed model will be able to help physicians in predicting the severity level of dengue patients before entering the critical phase. Furthermore, it will ease physicians in treating dengue patients early, so fatal cases or deaths can be avoided.

**Keywords:** Artificial Neural Network; Discriminant Analysis; dengue



**Citation:** Silitonga, P.; Bustamam, A.; Muradi, H.; Mangunwardoyo, W.; Dewi, B.E. Comparison of Dengue Predictive Models Developed Using Artificial Neural Network and Discriminant Analysis with Small Dataset. *Appl. Sci.* **2021**, *11*, 943. <https://doi.org/10.3390/app11030943>

Received: 8 December 2020

Accepted: 30 December 2020

Published: 21 January 2021

**Publisher's Note:** MDPI stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.



**Copyright:** © 2021 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

## 1. Introduction

Dengue is an acute febrile disease caused by dengue virus (DENV). Commonly, dengue incidences happen in tropical and subtropical countries, such as South America, Southeast Asia, etc. [1]. Dengue incidences usually occur in the rainy season, and happen in urban and suburban areas. Based on a study that analyzed 130 countries, there were around 9221 dengue deaths per year from 1990 to 2013, with the lowest of 8277 in 1992, and the highest of 11,302 in 2010 [2]. Dengue had made Indonesia suffered the most significant economic loss in Southeast Asia. The average annual economic burden of dengue in Indonesia was approximately USD 381.5 million. Indonesia has the highest dengue infection rate in Southeast Asia and the second-highest dengue infection rate in the world after Brazil [3]. In 2017, there were 59,047 Dengue Hemorrhagic Fever (DHF) cases and 444 DHF-associated deaths in Indonesia, with 0.75% case fatality rate and 22.55 incidence rate per 100,000 person-years [4]. In 2018, the national incidence of dengue was 24.75 cases per 100,000 population, resulting in 467 deaths [5].

DENV is a member of the family Flaviviridae and genus Flavivirus [6]. DENV is transferred by *Aedes aegypti* and *Aedes albopictus* female mosquitoes [7]. Those female mosquitoes consume blood as their regular meal to mature their eggs [8]. They fulfill their need for blood by biting humans. The incubation period of DENV is around 4 to 10 days.

After the incubation period ends, an infected mosquito can transfer DENV in its lifetime [9]. There are four serotypes of DENV, which are DENV-1, 2, 3, and 4. A DENV-infected person can be infected by one serotype of DENV or more [10].

Dengue consists of three clinical phases, which are febrile phase (occurs on the first until the third day of fever), critical phase (occurs on the fourth until sixth day of fever), and recovery phase (occurs on the seventh day of fever or afterwards). The common clinical symptoms of dengue are as follows: during the febrile phase, the symptoms are high fever, headache, nausea, myalgia, arthralgia, malaise, retro-orbital pain, and vomiting [11]. During the critical phase, the symptoms are thrombocytopenia, leukocytopenia, and plasma leakage, which is clinically manifested by hemoconcentration, pleural effusion, and/or ascites. Patients may also experience severe bleeding and shock [11]. During the recovery phase, the extravasated fluid is re-absorbed into the intravascular compartment. Some patients may experience an erythematous rash. The severity level of dengue consists of Dengue Fever (DF), Dengue Hemorrhagic Fever (DHF), and Dengue Shock Syndrome (DSS) [12]. DF is commonly classified as dengue, while DHF and DSS are commonly classified as severe dengue. The main difference between dengue and severe dengue is, patients who suffer severe dengue experience plasma leakage, while patients who suffer dengue do not. In this research, we only considered level DF and DHF. DHF itself was divided into two different levels, which were DHF grade 1 and DHF grade 2.

The gold standard to confirm a DENV infection are Reverse Transcription-Polymerase Chain Reaction (RT-PCR) test, culture of DENV, hemagglutination inhibition test, laboratory test, and tourniquet test. A suspected person can go through a laboratory test that measures hematocrit, thrombocyte (platelet count), and white blood cell (WBC). DENV infection will decrease in platelet count below 100,000 per  $\mu\text{L}$  between the third to eighth day onset of fever, and an increase in hematocrit of 20% or more. A positive tourniquet test also indicates that the corresponding patient might suffer dengue. Laboratory test parameters that we analyzed consist of hematocrit, hemoglobin, platelet count, WBC, monocyte, lymphocyte, and neutrophil.

In this research, we developed models that can predict the severity level of a dengue patient, based on the values of the seven laboratory test parameters. In other words, the laboratory test parameters were used as predictors to predict the severity level of a dengue patient. We used two methods, which were Artificial Neural Network (ANN) and Discriminant Analysis (DA). We decided to use ANN because ANN is one of machine learning methods that imitates the neuron structure of a human brain that has been used in many fields, including medical diagnosis prediction. Meanwhile, we decided to use DA because DA is one of dependency statistical analysis techniques in Multivariate Analysis that is used to classify an object into one of the statistically independent groups or categories. This is aligned with our purpose of this research, which was to classify data of dengue patients into one of the statistically independent severity levels of dengue.

We only used patients' data on the third day onset of fever. Because on the third day, dengue patients are going to enter the critical phase. We expected that this model can help physicians to treat dengue patients early before the patients enter the critical phase, so they can have timely treatment and fatal cases or deaths can be avoided. We also used the data on the third day because there haven't been any previous researches that used data on the third day.

The objectives of this research were to develop models that can predict the severity level of dengue using Artificial Neural Network (ANN) and Discriminant Analysis (DA), to evaluate the performances of the models, and to conclude which model has the best performance. A model with the best performance would be the proposed predictive model in this research. The predictive model will assist physicians in predicting the severity level of dengue.

## 2. Research Significance

Many patients have died in the critical phase due to the lack of proper and timely treatment. Therefore, we developed models that can predict the severity level of dengue based on the laboratory test results of the corresponding patients using ANN and DA. Our proposed predictive model—the one with the highest accuracy—will be able to help physicians in predicting the severity level of dengue patients before entering the critical phase. Furthermore, it will ease physicians in treating dengue patients early. So, dengue patients can receive proper and timely treatment, and fatal cases or deaths can be avoided.

## 3. Related Works

Abdiel E. Laureano-Rosario et al. [13] utilized ANN, which was trained with genetic algorithm to predict dengue fever outbreak in Puerto Rico and some areas in the coast of Mexico. They concluded that the model they developed using ANN had a good predictive ability.

Jorge D. Mello-Román et al. [14] compared two machine learning methods, which were Artificial Neural Networks multilayer perceptron (ANN-MLP) and Support Vector Machine (SVM) as the tools to assist medical diagnosis. ANN-MLP produced a better result with an average of 96% accuracy, 96% sensitivity, and 97% specificity. In conclusion, ANN-MLP could be used as a classifier to diagnose dengue infection with high accuracy, sensitivity, and specificity.

Oswaldo Santos Baquero et al. [15] compared Seasonal Autoregressive Integrated Moving Average (SARIMA), Generalized Additive Models (GAM), Artificial Neural Networks (ANN), naïve model, and ensemble model to predict dengue cases in São Paulo for one month ahead.

Tanujit Chakraborty et al. [16] developed a novel hybrid model, which was a combination of Autoregressive Integrated Moving Average (ARIMA) and Neural Network Autoregressive (NNAR). The model was used to analyze time series dengue data of three dengue endemic regions, which were San Juan, Iquitos, and Filipina. They concluded that the proposed hybrid model was easy to interpret, had an excellent performance in forecasting dengue epidemic for three dengue time series data from different regions, and had better forecasting accuracy compared to the other methods used in previous researches, such as traditional methods or other hybrid methods.

Siriyasatien et al. [17] developed models that enable forecasting of outbreaks of dengue, giving medical professionals the opportunity to develop plans for handling the outbreak, well in advance. They utilized several methods, one of them is ANN. Based on their results, the ANN model had two advantages: easy to be used in incremental learning and can learn to ignore irrelevant attributes.

Yulia Resti et al. [18] applied Quadratic Discriminant Analysis (QDA) in mapping the incidence of dengue into five areas in Palembang based on significant factors, such as age, gender, blood group, etc. The overall correct percentage of the mapping results was 66.7%.

Abdul Halim Poh et al. [19] utilized Principal Component Analysis (PCA) and DA to predict the patients' clinical dengue positivity. The DA method yielded accuracy between 93–98%, sensitivity between 75–89%, and specificity between 94–100%.

ANN and DA have been widely used to predict dengue incidences. That is why we decided to use both methods separately in our research to develop models that can predict the severity level of dengue based on the laboratory test results.

## 4. Materials and Methods

### 4.1. Dataset

We used dengue patients' data from the year 2009 and 2010 to develop the models (<https://drive.google.com/drive/folders/1C1ZciLa2Cwsb1lrDpBpUlp-2IZrckBQ?usp=sharing>). Many information was written in that data, such as age, gender, patients' length of stay, clinical symptoms, laboratory test results, and patient diagnosis. The diag-

noses were classified into three distinct severity levels, which were DF as the mild level, DHF grade 1 as the intermediate level, and DHF grade 2 as the severe level.

The data was obtained from Department of Microbiology, Universitas Indonesia. The data only consists of 77 dengue patients' data. It is very small because it is difficult to collect laboratory test results of dengue patients needed in this research. The data was split into training and testing data. We developed the models with three different data splits. That is, with the ratio of training: testing = 70%:30%, 80%:20%, and 90%:10%. Training data was used for learning, which was to fit the parameters (i.e., weights). While testing data was used to assess the (generalization) performance of the neural network [20]. Table 1 is the summary of our data:

**Table 1.** Dengue patients' data summary.

	Hemoglobin	Hematocrit	WBC	Platelet Count	Neutrophil	Lymphocyte	Monocyte	Severity Level
Minimum value	8.5	30.3	1000	6000	22	12.1	0	0
1st quartile	12.6	38.4	2330	93,000	55	20.7	3	0
Median	13.9	40.6	3100	119,000	64	25	10	1
Mean	13.9	41.3	3178	126,188	61.7	28	9.5	0.6
3rd quartile	15.4	45	4000	161,000	73	34	14	1
Maximum value	17.8	52.3	6900	278,000	84	68	23	2

#### 4.2. Discriminant Analysis (DA)

Discriminant Analysis (DA) is one of dependency statistical analysis techniques in Multivariate Analysis. It means that in DA, there are dependent and independent variables being analyzed, where the value of the dependent variable depends on the values of the independent variables. In DA, there is only one dependent variable, but there are more than one independent variables. DA can be used if the dependent variable is categorical (in nominal or ordinal scale), and the independent variables are metric (in interval or ratio scale). A categorical dependent variable means the corresponding variable consists of certain categories.

DA is used to classify an object into one of the statistically independent groups or categories based on the values of the independent variables. Classification process in DA is mutually exclusive, that is, if an object is classified into one particular group, it will not be classified into another group.

In DA, discriminant function(s) will be formulated. A discriminant function is used to classify an object into a group or category based on the value of the discriminant function itself. The values of the independent variables of an object will be inputted in the function. Then, the obtained value of the function, which is called the discriminant value, will determine in which group an object belongs. For  $k$  categories,  $k-1$  discriminant functions have to be formulated.

Based on the amount of the categories of the dependent variable, there are two types of DA, which are two-group discriminant analysis and Multiple Discriminant Analysis (MDA). Two-group discriminant analysis is a type of DA where the dependent variable consists of two categories. While MDA is another type of DA where the dependent variable consists of more than two categories.

Based on the type of the discriminant function, there are also two types of DA, which are Linear Discriminant Analysis (LDA) and Quadratic Discriminant Analysis (QDA). LDA is a classification method used to find the optimum linear combination of features to separate two or more groups of objects [21]. QDA can be considered as a direct extension of LDA. There are some differences between LDA and QDA. First, the discriminant function in LDA is a linear combination of the independent variables. While the discriminant function in QDA is in the form of quadratic function. Second, in LDA, every covariance matrix of

every independent variable has to be the identical. While in QDA, every covariance matrix of every independent variable can be different. The third difference, the decision curve in LDA is in the form of a straight line, while the decision curve in QDA is in the form of a quadratic curve [22].

Assume that observations  $X_j^{(k)}$  of a group  $k$  are random vectors of size  $p$  sampled from a Gaussian distribution  $\mathcal{N}(\mu^{(k)}, \Sigma^{(k)})$ , for all  $j \in \{1, \dots, N\}$ . With a Bayesian approach, an object  $x$  can be classified into group  $k^*$  that reach the maximum value of the posterior density function

$$k^* = \underset{k}{\operatorname{argmax}}(f_k(x)\pi_k) = \underset{k}{\operatorname{argmax}}(\log(f_k(x)\pi_k)) \tag{1}$$

where  $f_k(x) = \frac{1}{(2\pi)^{\frac{p}{2}} |\Sigma^{(k)}|^{\frac{1}{2}}} \exp\left\{-\frac{1}{2}(x - \mu^{(k)}) (\Sigma^{(k)})^{-1} (x - \mu^{(k)})^T\right\}$  is the density function of a Gaussian vector and  $\pi_k$  is the probability of an object belongs to group  $k$ . Formula (1) is the Quadratic Discriminant Analysis (QDA) formula. Linear Discriminant Analysis (LDA) has a similar formula. The only difference is in LDA, the covariance matrices  $\Sigma^{(k)}$  are assumed to be identical for all classes [23].

#### 4.3. Artificial Neural Network (ANN)

Artificial Neural Network (ANN) is a simple imitation of neuron structure of a human brain [24]. Similar to human brain, ANN is capable to analyze incomplete or unclear information, and furthermore, evaluate them. ANN imitates human brain in processing input signals and translating it into output signals. ANN is also capable to learn from data without any assumptions of certain functions.

ANN is a part of Artificial Intelligence, along with Support Vector Machines, Expert Systems, and Fuzzy Logic. ANN consists of processing units which are called artificial neuron. Artificial neurons try to imitate the structure and behavior of biological neurons. A neuron can consist of more than one input (dendrite), but commonly consists of only one output (synapsis through axon).

A neuron has a function which determines the activation of the neuron itself. That function is called an activation function. An activation function processes input signals that have been combined together, then transforms them into an output signal. Mathematically, the procedure of signal processing can be expressed as follows:

$$y(x) = \Phi\left(\sum_{i=1}^n w_i \cdot x_i\right)$$

where  $y$  is the output signal,  $\Phi(\cdot)$  is the activation function,  $x$  is the input variable, and  $w$  is the weight that is given to each of the input variable [25].

There's a lot of types of activation functions, which are linear, sigmoid, step, ramp, hyperbolic tangent, etc. The most frequently used activation function is sigmoid function. Hyperbolic tangent function has a similar shape to sigmoid function, but its value lies between  $-1$  to  $+1$ , unlike sigmoid which the values lies between  $0$  to  $1$ .

ANN can be widely used in different scopes of problems, such as finding new features of an object, and classifying or predicting an object/event using huge sets of data. Some of the fields where ANN is frequently used are medical diagnosis prediction, character recognition, speech recognition, human face recognition, signature verification application, etc.

ANN has a different way of work with normal computers in many ways. ANN has strengths compared to normal computer programs. Some of the strengths of ANN are as follows: (i) ANN is an adaptive learning method. It imitates human brain on how to do its task while learning, even with different types of inputs; (ii) ANN can organize itself while learning; (iii) ANN works parallelly like human brain; (iv) ANN has a high fault tolerance.

It is able to work even on fuzzy, noisy, and incomplete data; and (v) ANN is applicable to classify data, recognize patterns, and any other tasks that involve obscure data.

4.4. Confusion Matrix and Performance Measurement

Confusion matrix is a matrix used to evaluate the performance of classifier models in general. For binary classification problems, confusion matrix size  $2 \times 2$  is used. However, in real life problems, classification can involve more than two classes or categories. These problems are considered as multi-class classification problems. For a multi-class classification problem, confusion matrix size  $3 \times 3$  is used. Principally,  $3 \times 3$  confusion matrix is similar to  $2 \times 2$  confusion matrix. The general form of  $3 \times 3$  confusion matrix is as follows:

$$\begin{array}{ccc}
 & A & B & C \\
 A & aa & ab & ac \\
 B & ba & bb & bc \\
 C & ca & cb & cc
 \end{array}$$

where the columns denote actual cases, the rows denote predicted cases.

- For category A,
  - True Positive (TP):  $aa$
  - True Negative (TN):  $bb + cb + bc + cc$
  - False Positive (FP):  $ba + ca$
  - False Negative (FN):  $ab + ac$
- Calculations for other categories are similar.

In this research, our problem was a multi-class classification problem, because the severity level consisted of three distinct categories, and we aimed to classify each dengue patient into one of those level. So, we used  $3 \times 3$  confusion matrix and these performance measurements below:

1. *Accuracy*: This metric measures the overall performance of the model. Generally, accuracy is the proportion of true results among the total number of cases examined. *Accuracy* can be expressed as follows:

$$Accuracy = \frac{\text{number of correctly classified examples}}{\text{total number of cases}} \tag{2}$$

*Accuracy* can also be calculated from the elements of the  $3 \times 3$  confusion matrix using the formula below:

$$Accuracy = \frac{\sum_{i=1}^r x_{ii}}{\sum_{i=1}^r \sum_{j=1}^r x_{ij}} \tag{3}$$

where  $x$  denotes an element of the confusion matrix,  $i$  denotes the number of rows of the confusion matrix, and  $j$  denotes the number of columns of the confusion matrix. Note that the numerator of the Formula (3) is the sum of the main diagonal elements of the confusion matrix, and the denominator is the sum of all of the elements of the confusion matrix.

2. *Sensitivity*: this metric measures the proportion of the True Positive (TP) cases among the total number of positive cases. The formula to calculate *sensitivity* is as follows:

$$Sensitivity = \frac{TP}{TP + FN}$$

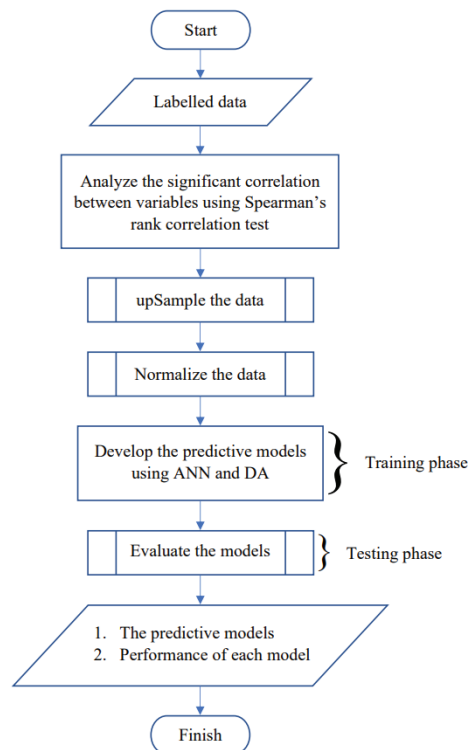
3. *Specificity*: This metric measures the proportion of the True Negative (TN) cases among the total number of negative cases. The formula to calculate *sensitivity* is as follows:

$$Specificity = \frac{TN}{TN + FP}$$

## 5. Results and Discussion

### 5.1. Model Construction

As mentioned previously, we developed models to predict the severity level of dengue with ANN and DA. The models will be able to predict the severity level of dengue based on the laboratory test results of the corresponding patients. The accuracy of both models developed with ANN and DA would be evaluated, and the one with the higher accuracy would be the proposed model. For short, we mentioned the models as “predictive models” in Figure 1.



**Figure 1.** Experiment workflow.

By inputting the values of hematocrit, hemoglobin, platelet count, WBC, monocyte, lymphocyte, and neutrophil of the corresponding patient into the predictive model, the model would be able to process those values and predict the severity level of the dengue patient. In other words, the input variables of the model were the independent variables of this research, which were the seven laboratory test parameters. While the output variable was the dependent variable of this research, which was the severity level of dengue.

Before we developed the model, we conducted Spearman’s rank correlation test. This correlation test was conducted for two purposes:

- To examine whether there is a significant correlation between the independent variables. To be able to use DA, there has to be no multicollinearities between the independent variables. This means all of the independent variables are not correlated. If there’s a couple of correlated independent variables, one of them has to be eliminated.
- To conduct feature selection. This means to select only the necessary independent variables to develop the models.

First, we conducted a two-ways hypothesis test, where the hypotheses were as follows:



$H_0: \rho = 0$ , which means there is no significant correlation between variable  $x_i$  and  $x_j$ , where  $i, j = 1, \dots, 7, i \neq j$ .  $H_1: \rho \neq 0$ , which means there is a significant correlation between variable  $x_i$  and  $x_j$ .

We used a significance level of 0.05. Below were the results of the correlation test.

The elements of Table 2 were the correlation coefficients between the corresponding independent variables. The starred coefficients mean that there is a significant correlation between the corresponding independent variables, or in other words, the corresponding independent variables were correlated. The correlated independent variables were as follows:

- Hemoglobin and hematocrit
- Hemoglobin and platelet count
- Hematocrit and platelet count
- WBC and platelet count
- Neutrophil and lymphocyte
- Neutrophil and monocyte

**Table 2.** The result of Spearman’s rank correlation test between the laboratory test parameters.

	Hemoglobin	Hematocrit	WBC	Platelet count	Neutrophil	Lymphocyte	Monocyte
Hemoglobin	1	0.957 *	0.126	−0.475 *	−0.120	0.137	0.001
Hematocrit	0.957 *	1	0.186	−0.380 *	−0.071	0.144	−0.080
WBC	0.126	0.186	1	0.262 *	0.069	−0.153	0.028
Platelet count	−0.475 *	−0.380 *	0.262 *	1	0.061	−0.133	0.099
Neutrophil	−0.120	−0.071	0.069	0.061	1	−0.833 *	−0.581 *
Lymphocyte	0.137	0.144	−0.153	−0.133	−0.833 *	1	0.143
Monocyte	0.001	−0.080	0.028	0.099	−0.581 *	0.143	1

\* Correlation is significant at the 0.05 level (2-Tailed).

Therefore, we eliminated some of the independent variables which were correlated with more than one other independent variables. The eliminated independent variables were hematocrit, platelet count, and neutrophil. Furthermore, we only used the remaining four independent variables to develop the models – hemoglobin, WBC, lymphocyte, and monocyte.

Then, we applied the upSample technique because the data that had been used in this research was imbalanced. upSampling means doing a sampling with replacement, in a random manner, to the data in the minority category until the sample size equals the size of the majority category. upSample is one of the techniques to handle imbalanced dataset. An imbalanced dataset is usually upSampled before it is used to develop the model, so the developed model will tend to have higher accuracy. If the dataset is not upSampled and is directly used to develop a model, it is unlikely to yield a model with high accuracy. The dataset used in our research was imbalanced, in the sense that the classes are not represented equally. Before the data was upSampled, it consisted of 77 cases. From the 77 cases, 38 cases were from category DF, 28 cases were from category DHF grade 1, and 11 cases were from category DHF grade 2. It is clear that the majority category was DF with a size of 38, and the minority category was DHF grade 2 with size 11. After the dataset was upSampled, each category has the size 38. So, in total, the upSampled data consisted of  $38 \times 3 = 114$  cases.

After we applied the upSample technique, we standardized the data. We only standardized the data of the independent variables, which were the seven laboratory test parameters. We didn’t standardize the data of the dependent variable, which was the severity level of dengue, because we wanted the real values of the dependent variable. Data of the independent variables was standardized because the values varied greatly and had different units. It was also standardized in order to develop better models compared to models developed using unstandardized data. After the data was standardized, we started to develop the models. We developed twelve models in total. Six models were developed

with ANN, and the other six were developed with DA. We developed the models using R programming version 3.6.3. We decided to use R because it has some advantages, such as: (i) it has data science engine, specifically the statistics and machine learning packages, and (ii) it is an open-source language programming, so it is accessible by anyone.

### 5.2. Predictive Models Developed Using DA

We developed two DA models. The first model that we developed was Linear Discriminant Analysis (LDA) model, and the second one was Quadratic Discriminant Analysis (QDA) model. Below were the table of performance measurements of both models.

#### 1. LDA

Based on Table 3, it is shown that the LDA model with 70% training data yielded the highest accuracy (45.45%), sensitivity (45.73%), and specificity (72.89%).

**Table 3.** Performance measurements of the Linear Discriminant Analysis (LDA) model.

	Training Data Percentage		
	70%	80%	90%
<i>Accuracy</i>	45.45%	30.43%	27.27%
<i>Sensitivity</i>	45.73%	31.95%	27.78%
<i>Specificity</i>	72.89%	65.05%	63.10%

#### 2. QDA

Based on Table 4, it is shown that the QDA model with 80% training data yielded the highest accuracy (60.87%) and sensitivity (59.72%).

**Table 4.** Performance measurements of the Quadratic Discriminant Analysis (QDA) model.

	Training Data Percentage		
	70%	80%	90%
<i>Accuracy</i>	60.61%	60.87%	54.55%
<i>Sensitivity</i>	56.07%	59.72%	55.56%
<i>Specificity</i>	82.43%	80.67%	80.56%

Based on the obtained results, QDA model with 80% training data yielded the highest accuracy (60.87%) and sensitivity (59.72%). Therefore, QDA model with 80% training data was the proposed predictive model so far. This model was then compared with the other predictive models developed using ANN.

### 5.3. Predictive Models Developed Using ANN

We also developed the predictive models using ANN—next will be called ANN model for short. We developed the ANN models using R programming with the nnet package. The nnet trains feed-forward neural networks with traditional backpropagation algorithm [26].

Feed-Forward Neural Networks (FFNN) is a type of ANN that doesn't have any feed-back connection from the output to the input [27]. In FFNN, neurons of the previous layer are entirely connected to the consecutive layer, but there are no intra-layer connections [28]. FFNN is a supervised learning method that is utilized for classification problems and consists of input, hidden, and output layers [29].

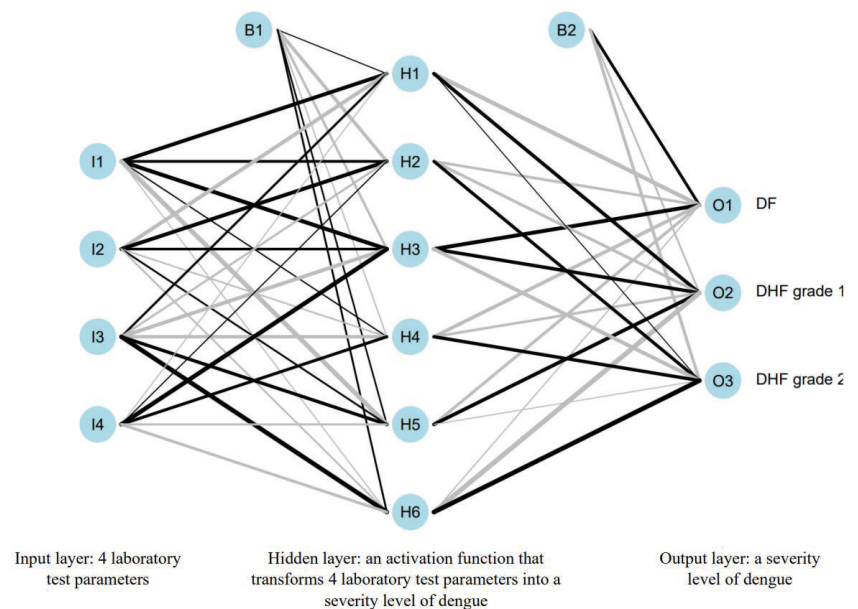
Our ANN architecture consisted of one input layer, one hidden layer, and one output layer. There were four neurons in input layer that represented hemoglobin, WBC, lymphocyte, and monocyte.

There was only one hidden layer in this ANN architecture. The number of hidden layers and neurons in each hidden layer were obtained through the iteration process until the accuracy could not be furtherly improved [30]. The number of hidden layers is arbitrary. However, one hidden layer is commonly used for simple problems. The numbers of input and output neurons are determined according to the problem, while the number of hidden neurons was well optimized [20]. The problem in this research could be considered as a simple problem because it only consisted of four independent variables and the purpose was related to medical diagnosis, one of the common purposes of using ANN. That was why one hidden layer was enough for this research. Meanwhile, to ease the iteration process, we decided to use the rule of thumb in determining the hidden neurons. Based on the research conducted by Panchal and Panchal [31], one of the rules in determining the number of hidden neurons is “the number of hidden neurons should be 2/3 of the input layer size, plus the size of the output layer”.

We followed this rule when developing the ANN model. So, the number of hidden neurons is:

$$\begin{aligned}
 \text{Hiddenneuron} &= \left(\frac{2}{3} \times \text{inputneuron}\right) + \text{outputneuron} \\
 &= \left(\frac{2}{3} \times 4\right) + 3 \\
 &= \frac{17}{3} \\
 &= 5.66\dots \\
 &\approx 6
 \end{aligned}$$

Meanwhile, the number of input neurons was 4 because there were four remaining laboratory test parameters. The number of the output neuron was 3 because it represented three categories of severity level. The output that would come out was the probability values from each neuron that would sum up to 1. The probability value means the probability of a patient to suffer from dengue in a corresponding level. For example, the output value from neuron DF is 0.3, from DHF grade 1 is also 0.3, and from DHF grade 2 is 0.4. So, the patient is suffering dengue level DHF grade 2 because the probability value is the highest compared to the other two. Figure 2 was the architecture of the developed ANN predictive model.



**Figure 2.** Artificial Neural Network (ANN) architecture of the predictive models developed using R.

The activation function is frequently a bounded nondecreasing, differentiable, and nonlinear function such as the hyperbolic tangent or the logistic function [26]. We tried to use logistic and hyperbolic tangent as the activation functions. First, we developed an ANN model with logistic activation function. Then, we developed another ANN model with hyperbolic tangent activation function.

Logistic function may have different forms, but the one used in R programming—the one we used in this research—is the standard logistic function, which is also known as sigmoid function. The formula of the logistic function is as follows

$$f(x_i) = \frac{1}{1 + e^{-x_i}} = \frac{e^{x_i}}{e^{x_i} + 1}$$

where  $x_i$  is the value of the  $i$ th independent variable [32].

Meanwhile, hyperbolic tangent function, which is also known as tanh function, has the formula as follows

$$\tanh(x_i) = \frac{\sinh(x_i)}{\cosh(x_i)} = \frac{e^{x_i} - e^{-x_i}}{e^{x_i} + e^{-x_i}} \tag{4}$$

where  $x_i$  is the value of the  $i$ th independent variable [33].

We provided the biases and weights of our ANN models, so our developed ANN models can be useful for other researchers who want to conduct similar researches. We displayed them in tables so it would be easier to read. Below is the annotation of the notations in the tables below:

Bi–Hj: weight from Bi to Hj, where  $i = 1, 2$  and  $j = 1, \dots, 6$ .

Ik–Hj: weight from Ik to Hj, where  $k = 1, \dots, 4$ .

Bi–Ol: weight from Bi to Ol, where  $l = 1, 2, 3$ .

Hj–Ol: weight from Hj to Ol.

Tables 5 and 6 were the tables of the biases and weights of our ANN models.

1. ANN model developed using logistic activation function

**Table 5.** The biases and weights of the ANN model developed using logistic activation function.

	Data Training Percentage		
	70%	80%	90%
B1–H1	5.4039439	−4.4910300	4.5469032
I1–H1	−15.0766425	11.3254070	−11.0499554
I2–H1	1.5627953	−1.7490682	2.7245021
I3–H1	2.7587930	−0.9163399	0.9683699
I4–H1	−0.1562694	−0.6655746	−1.9540237
B1–H2	−13.793552	−1.4021008	−3.2660749
I1–H2	6.031370	9.1146927	0.7803872
I2–H2	6.958133	−0.3967136	6.6604275
I3–H2	4.942339	3.8428082	−0.7337680
I4–H2	7.859338	−7.9523955	5.1178065
B1–H3	−4.721011	4.965760	−8.241048
I1–H3	4.605214	−6.577878	1.584072
I2–H3	−5.536911	5.634194	6.583553
I3–H3	−13.412774	7.212313	−2.312179
I4–H3	10.487666	−6.157306	6.642880
B1–H4	−2.739654	−5.661526	−3.080241

Table 5. Cont.

	Data Training Percentage		
	70%	80%	90%
I1-H4	9.506876	1.911992	-6.917437
I2-H4	6.985396	10.440174	-1.723676
I3-H4	-4.915390	11.821140	-2.475820
I4-H4	-5.293412	-2.478422	10.804223
B1-H5	10.342958	-6.097498	5.354134
I1-H5	-5.340765	4.321218	-2.957952
I2-H5	-12.073148	6.730848	-7.420634
I3-H5	-1.495051	-1.673032	-1.183222
I4-H5	-9.681449	7.487540	-4.575078
B1-H6	1.1958247	-10.231457	5.565350
I1-H6	-12.1059627	-1.590068	-4.021247
I2-H6	0.0344864	8.865150	5.553954
I3-H6	-9.2652208	-4.616111	5.455334
I4-H6	12.0185146	12.299163	-10.527717
B2-O1	14.765571	-11.104739	-0.62485862
H1-O1	-6.662708	5.525194	0.90165390
H2-O1	-10.577831	-6.403159	5.92878925
H3-O1	-7.284002	1.918726	-9.87260412
H4-O1	-8.744731	8.004559	0.08648525
H5-O1	-12.319315	8.997052	-7.63236367
H6-O1	-4.108802	-12.639489	-0.07664491
B2-O2	-8.337048	-0.07675744	-3.107592
H1-O2	-9.444444	8.49491407	-9.655582
H2-O2	8.975756	-7.48027294	-6.024728
H3-O2	-8.726450	9.93562367	5.738516
H4-O2	4.006490	-7.01414123	9.385994
H5-O2	10.649440	-10.11985671	2.076983
H6-O2	14.583526	11.58227775	8.594578
B2-O3	-4.4530081	-1.13586835	1.294068
H1-O3	9.7557132	0.02180394	8.395259
H2-O3	-3.0576919	8.99669282	-1.187068
H3-O3	13.5887864	-9.10282385	4.191767
H4-O3	-0.8085856	-6.27224615	-8.351199
H5-O3	0.8808544	-1.02923910	5.886176
H6-O3	-12.4993457	4.83479909	-8.863369

2. ANN model developed using hyperbolic tangent (tanh) activation function

**Table 6.** The biases and weights of the ANN model developed using tanh activation function.

	Data Training Percentage		
	70%	80%	90%
B1–H1	5.4039439	−0.1157967	4.5469032
I1–H1	−15.0766425	−5.3827556	−11.0499554
I2–H1	1.5627953	−3.4223409	2.7245021
I3–H1	2.7587930	2.9422321	0.9683699
I4–H1	−0.1562694	4.3960790	−1.9540237
B1–H2	−13.793552	−9.118127	−3.2660749
I1–H2	6.031370	2.836462	0.7803872
I2–H2	6.958133	12.101105	6.6604275
I3–H2	4.942339	1.334247	−0.7337680
I4–H2	7.859338	9.306791	5.1178065
B1–H3	−4.721011	4.775554	−8.241048
I1–H3	4.605214	−5.185507	1.584072
I2–H3	−5.536911	7.562261	6.583553
I3–H3	−13.412774	8.626503	−2.312179
I4–H3	10.487666	−9.557800	6.642880
B1–H4	−2.739654	10.686570	−3.080241
I1–H4	9.506876	−1.389448	−6.917437
I2–H4	6.985396	−7.260299	−1.723676
I3–H4	−4.915390	4.507227	−2.475820
I4–H4	−5.293412	−11.680316	10.804223
B1–H5	10.342958	2.1681860	5.354134
I1–H5	−5.340765	−11.4221402	−2.957952
I2–H5	−12.073148	−0.8285786	−7.420634
I3–H5	−1.495051	−5.1982137	−1.183222
I4–H5	−9.681449	9.6726127	−4.575078
B1–H6	1.1958247	4.9131114	5.565350
I1–H6	−12.1059627	−13.9610034	−4.021247
I2–H6	0.0344864	2.4310158	5.553954
I3–H6	−9.2652208	2.2366086	5.455334
I4–H6	12.0185146	0.2850172	−10.527717
B2–O1	14.765571	−17.24445758	−0.62485862
H1–O1	−6.662708	6.63597309	0.90165390
H2–O1	−10.577831	9.95768136	5.92878925
H3–O1	−7.284002	3.98047085	−9.87260412
H4–O1	−8.744731	9.88497394	0.08648525
H5–O1	−12.319315	−0.06539451	−7.63236367
H6–O1	−4.108802	−7.47772434	−0.07664491
B2–O2	−8.337048	6.135447	−3.107592
H1–O2	−9.444444	−1.290546	−9.655582
H2–O2	8.975756	−10.614987	−6.024728

Table 6. Cont.

	Data Training Percentage		
	70%	80%	90%
H3–O2	−8.726450	8.371778	5.738516
H4–O2	4.006490	−9.933384	9.385994
H5–O2	10.649440	11.691794	2.076983
H6–O2	14.583526	−8.421655	8.594578
B2–O3	−4.4530081	7.2697955	1.294068
H1–O3	9.7557132	−0.7564020	8.395259
H2–O3	−3.0576919	−2.9147165	−1.187068
H3–O3	13.5887864	−10.5569871	4.191767
H4–O3	−0.8085856	−0.6266332	−8.351199
H5–O3	0.8808544	−9.8940081	5.886176
H6–O3	−12.4993457	9.8877980	−8.863369

Below were the tables of performance measurements of both models.

1. ANN model developed using logistic activation function

Based on Table 7, it is shown that the ANN model with 70% training data yielded the highest accuracy (90.91%), sensitivity (91.11%), and specificity (95.51%).

Table 7. Performance measurements of the ANN model developed using logistic activation function.

	Training Data Percentage		
	70%	80%	90%
Accuracy	90.91%	78.26%	72.73%
Sensitivity	91.11%	81.39%	72.22%
Specificity	95.51%	88.19%	86.90%

2. ANN model developed using hyperbolic tangent (tanh) activation function

Based on Table 8, it is shown that the ANN model with 70% training data yielded the highest accuracy (90.91%), sensitivity (91.11%), and specificity (95.51%).

Table 8. Performance measurements of the ANN model developed using tanh activation function.

	Training Data Percentage		
	70%	80%	90%
Accuracy	90.91%	78.26%	72.73%
Sensitivity	91.11%	79.68%	72.22%
Specificity	95.51%	89.10%	86.90%

#### 5.4. The Proposed Predictive Model

Based on the obtained results, both ANN models developed using logistic and hyperbolic tangent activation function with 70% training data yielded the highest accuracy (90.91%), sensitivity (91.11%), and specificity (95.51%). These models also have the highest accuracy, sensitivity, and specificity compared to the previously proposed model, QDA model with 80% training data. Therefore, both ANN models developed using logistic and tanh activation function with 70% training data were the proposed predictive models in this research. We displayed the proposed predictive models' performances in Table 9.



**Table 9.** Performance measurements of the proposed predictive ANN model.

	Activation Function	
	Logistic	Tanh
Accuracy	90.91%	90.91%
Sensitivity	91.11%	91.11%
Specificity	95.51%	95.51%

## 6. Conclusions and Recommendation

As shown in the previous chapter, both ANN models developed using logistic and hyperbolic tangent activation function with 70% training data yielded the highest accuracy (90.91%), sensitivity (91.11%), and specificity (95.51%). Therefore, both ANN models developed using logistic and tanh activation function with 70% training data are the proposed models to be used as the models to predict the severity level of dengue based on the laboratory test results.

We suggest other researchers consider developing another ANN architecture to be applied to the same data, to which the link has been given to obtain a new model with better performance. Researchers may improvise with the number of hidden neurons and/or hidden layer(s).

**Author Contributions:** Writing—original draft, developing the ANN models, and algorithm implementation, P.S.; project administrator, designing methods, formal analysis, and final revision, A.B.; Software—developing the DA models, H.M.; Funding acquisition, W.M.; Writing—review & editing, designing methods, and clinical analysis, B.E.D. All authors have read and agreed to the published version of the manuscript.

**Funding:** This work was fully funded by PDUPT 2020 with the contract no. NKB-2827/UN2.RST/HKP.05.00/2020 from Kementrian Riset dan Teknologi/Badan Riset dan Inovasi Nasional, Indonesia.

**Institutional Review Board Statement:** Not applicable.

**Informed Consent Statement:** Not applicable.

**Data Availability Statement:** The data presented in this study are openly available in the link provided in Section 4.1.

**Conflicts of Interest:** The authors declare no conflict of interest.

## References

- Silitonga, P.; Dewi, B.E.; Bustamam, A.; Siswantining, T. Correlation between laboratory characteristics and clinical degree of dengue as an initial stage in a development of machine learning predictor program. *AIP Conf. Proc.* **2020**, 030008, In Symposium on Biomathematics 2019, SYMOMATH 2019. Apri, M., Akimenko, V., Eds.; American Institute of Physics Inc. 2020. 030008. (AIP Conference Proceedings). [[CrossRef](#)]
- Stanaway, J.; Shepard, D.; Undurraga, E.; Halasa, Y.; Coffeng, L.; Brady, O.; Hay, S.I.; Bedi, N.; Bensenor, I.M.; Castañeda-Orjuela, C.A.; et al. The global burden of dengue: An analysis from the Global Burden of Disease Study 2013. *Lancet Infect Dis.* **2016**, *16*, 712–723. [[CrossRef](#)]
- Widoretno Sjahrurachman, A.; Dewi, B.; Lischer, K.; Pratami, D.; Flamandita, D.; Sahlan, M. Surface plasmon resonance analysis for detecting non-structural protein 1 of dengue virus in Indonesia. *Saudi J. Biol. Sci.* **2020**, *27*, 1931–1937. [[CrossRef](#)] [[PubMed](#)]
- Harapan, H.; Michie, A.; Mudatsir, M.; Sasmono, R.; Imrie, A. Epidemiology of dengue hemorrhagic fever in Indonesia: Analysis of five decades data from the National Disease Surveillance. *BMC Res. Notes* **2019**, *350*, 1–6. [[CrossRef](#)] [[PubMed](#)]
- Santoso, M.; Yohan, B.; Denis, D.; Hayati, R.; Haryanto, S.; Trianty, L.; Noviyanti, R.; Hibberd, M.L.; Sasmono, R.T. Diagnostic accuracy of 5 different brands of dengue virus non-structural protein 1 (NS1) antigen rapid diagnostic tests (RDT) in Indonesia. *Diagn. Microbiol. Infect. Dis.* **2020**, *98*, 1–7. [[CrossRef](#)] [[PubMed](#)]
- Beltrán-Silva, S.; Chacón-Hernández, S.; Moreno-Palacios, E.; Pereyra-Molina, J. Clinical and differential diagnosis: Dengue, chikungunya and Zika. *Rev. Méd. Hosp. Gen. Méx.* **2016**, *81*, 146–153. [[CrossRef](#)]
- Smartt, C.; Shin, D.; Alto, B. Dengue serotype-specific immune response in *Aedes aegypti* and *Aedes albopictus*. *Mem. Inst. Oswaldo Cruz.* **2017**, *112*, 829–837. [[CrossRef](#)] [[PubMed](#)]
- Bustamam, A.; Aldila, D.; Yuwanda, A. Understanding Dengue Control for Short- and Long-Term Intervention with a Mathematical Model Approach. *J. Appl. Math.* **2018**, *2018*, 1–13. [[CrossRef](#)]

9. Götz, T.; Altmeier, N.; Bock, W.; Rockenfeller, R.; Sutimin, S.; Wijaya, K. Modeling dengue data from Semarang, Indonesia. *Ecol. Complex.* **2016**, *30*, 57–62. [[CrossRef](#)]
10. Lardo, S.; Utami, Y.; Yohan, B.; Tarigan, S.M.; Santoso, W.D.; Nainggolan, L.; Sasmono, R.T. Concurrent infections of dengue viruses serotype 2 and 3 in patient with severe dengue from Jakarta, Indonesia. *Asian Pac. J. Trop. Med.* **2016**, *9*, 134–140. [[CrossRef](#)]
11. Kuo, H.; Lee, I.; Liu, J. Analyses of clinical and laboratory characteristics of dengue adults at their hospital presentations based on the World Health Organization clinical-phase framework: Emphasizing risk of severe dengue in the elderly. *J. Microbiol. Immunol. Infect.* **2016**, *51*, 740–748. [[CrossRef](#)] [[PubMed](#)]
12. Dewi, B.E.; Lugito, N.P.; Sutrisna, B.; Pohan, H.T.; Syahrurachman, A.; Widodo, D.; Ronoatmodjo, S.; Sudaryo, M.K.; Windiyansih, C.; Sudiro, T.M. Scoring Model to Predict Dengue Infection in the Early Phase of Illness in Primary Health Care Centre. *Arch. Clin. Microbiol.* **2015**, *6*, 1–8.
13. Laureano-Rosario, A.E.; Duncan, A.P.; Mendez-Lazaro, P.A.; Garcia-Rejon, J.E.; Gomez-Carro, S.; Farfan-Ale, J.; Savic, D.A.; Muller-Karger, F.E. Application of Artificial Neural Networks for Dengue Fever Outbreak Predictions in the Northwest Coast of Yucatan, Mexico and San Juan, Puerto Rico. *Trop. Med. Infect. Dis.* **2018**, *3*, 1–16. [[CrossRef](#)]
14. Mello-Román, J.; Mello-Román, J.; Gómez-Guerrero, S.; García-Torres, M. Predictive Models for the Medical Diagnosis of Dengue: A Case Study in Paraguay. *Comput. Math. Methods Med.* **2019**, *2019*, 1–7. [[CrossRef](#)] [[PubMed](#)]
15. Baquero, O.; Santana, L.; Chiaravalloti-Neto, F. Dengue forecasting in São Paulo city with generalized additive models, artificial neural networks and seasonal autoregressive integrated moving average models. *PLoS ONE* **2018**, *13*, e0195065. [[CrossRef](#)]
16. Chakraborty, T.; Chattopadhyay, S.; Ghosh, I. Forecasting dengue epidemics using a hybrid methodology. *Phys. A: Stat. Mech. Its Appl.* **2019**, *2019*, 1–8. [[CrossRef](#)]
17. Sิริyasatien, P.; Chadsuthi, S.; Jampachaisri, K.; Kesorn, K. Dengue Epidemics Prediction: A Survey of the State-of-the-Art Based on Data Science Processes. *IEEE Access* **2018**, *6*, 53757–53795. [[CrossRef](#)]
18. Resti, Y.; Ismail, N.; Verawaty, M. Quadratic Discriminant Analysis of Dengue Viruses Disease Incidence in Palembang. *Am. J. Appl. Sci.* **2017**, *14*, 578–582. [[CrossRef](#)]
19. Poh, A.H.; Adikan, F.R.M.; Moghavvemi, M.; Omar, S.F.S.; Poh, K.; Mahyuddin, M.B.H.; Yan, G.; Ariffin, M.A.A.; Harun, S.W. Precursors to non-invasive clinical dengue screening: Multivariate signature analysis of in-vivo diffuse skin reflectance spectroscopy on febrile patients in Malaysia. *PLoS ONE* **2020**, *15*, e0228923. [[CrossRef](#)]
20. Jiang, H.; Xi, Z.; Rahman, A.; Zhang, X. Prediction of output power with artificial neural network using extended datasets for Stirling engines. *Appl. Energy* **2020**, *271*, 1–16. [[CrossRef](#)]
21. Othman, N.H.; Lee, K.Y.; Radzol, A.R.M.; Mansor, W.; Ramlan, N.N.M. Linear Discriminant Analysis for Detection of Salivary NS1 from SERS Spectra. In Proceedings of the TENCON 2017—2017 IEEE Region 10 Conference, Penang, Malaysia, 5–8 November 2017; pp. 2876–2879.
22. Zhang, M. Identification of protein coding regions in the human genome by quadratic discriminant analysis. *Proc. Natl. Acad. Sci. USA* **1997**, *94*, 565–568. [[CrossRef](#)]
23. Le, K.; Chau, C.; Richard, F.; Guedj, E. An adapted linear discriminant analysis with variable selection for the classification in high-dimension, and an application to medical data. *Comput. Stat. Data Anal.* **2020**, *152*, 1–12. [[CrossRef](#)]
24. Kukreja, H.; Bharath, N.; Siddesh, C.S.; Kuldeep, S. An introduction to artificial neural network. *Int. J. Adv. Res. Innov. Ideas Educ.* **2016**, *1*, 27–30.
25. Zhang, Z. A gentle introduction to artificial neural networks. *Ann. Transl. Med.* **2016**, *4*, 1–6. [[CrossRef](#)] [[PubMed](#)]
26. Günther, F.; Fritsch, S. neuralnet: Training of Neural Networks. *R J.* **2010**, *2*, 30–38. [[CrossRef](#)]
27. Mansoor, M.; Grimaccia, F.; Leva, S.; Mussetta, M. Comparison of echo state network and feed-forward neural networks in electrical load forecasting for demand response programs. *Math. Comput. Simul.* **2020**, 1–12. [[CrossRef](#)]
28. Zhang, A.; Zhou, H.; Li, X.; Zhu, W. Fast and robust learning in Spiking Feed-forward Neural Networks based on Intrinsic Plasticity mechanism. *Neurocomputing* **2019**, *365*, 102–112. [[CrossRef](#)]
29. Yibre, A.; Koçer, B. Semen quality predictive model using Feed Forwarded Neural Network trained by Learning-Based Artificial Algae Algorithm. *Eng. Sci. Technol. Int. J.* **2020**, 1–9. [[CrossRef](#)]
30. Zhao, N.; Charland, K.; Carabali, M.; Nsoesie, E.; Maher-Giroux, M.; Rees, E.; Yuan, M.; Balaguera, C.G.; Ramirez, G.J.; Zinszer, K. Machine learning and dengue forecasting: Comparing random forests and artificial neural networks for predicting dengue burden at national and sub-national scales in Colombia. *PLoS Negl. Trop. Dis.* **2020**, 1–16.
31. Panchal, F.; Panchal, M. Review on Methods of Selecting Number of Hidden Nodes in Artificial Neural Network. *IJCSMC* **2014**, *3*, 455–464.
32. Kamson, A.; Sharma, L.; Dandapat, S. Enhancement of the heart sound envelope using the logistic function amplitude moderation method. *Comput. Methods Programs Biomed.* **2019**, *187*, 1–9. [[CrossRef](#)] [[PubMed](#)]
33. Liu, T.; Qiu, T.; Luan, S. Hyperbolic-tangent-function-based cyclic correlation: Definition and theory. *Signal Process.* **2019**, *164*, 206–216. [[CrossRef](#)]

Article

# Data-Dependent Feature Extraction Method Based on Non-Negative Matrix Factorization for Weakly Supervised Domestic Sound Event Detection

Seokjin Lee <sup>1,2,\*</sup> , Minhan Kim <sup>1</sup>, Seunghyeon Shin <sup>1</sup>, Sooyoung Park <sup>3</sup> and Youngho Jeong <sup>3</sup>

<sup>1</sup> School of Electronic and Electrical Engineering, Kyungpook National University, Daegu 41566, Korea; kmh7576@knu.ac.kr (M.K.); sineva123@gmail.com (S.S.)

<sup>2</sup> School of Electronics Engineering, Kyungpook National University, Daegu 41566, Korea

<sup>3</sup> Media Research Division, Electronics and Telecommunications Research Institute, Daejeon 34129, Korea; sooyoung@etri.re.kr (S.P.); yhcheong@etri.re.kr (Y.J.)

\* Correspondence: sjlee6@knu.ac.kr; Tel.: +82-53-950-5523

**Abstract:** In this paper, feature extraction methods are developed based on the non-negative matrix factorization (NMF) algorithm to be applied in weakly supervised sound event detection. Recently, the development of various features and systems have been attempted to tackle the problems of acoustic scene classification and sound event detection. However, most of these systems use data-independent spectral features, e.g., Mel-spectrogram, log-Mel-spectrum, and gammatone filterbank. Some data-dependent feature extraction methods, including the NMF-based methods, recently demonstrated the potential to tackle the problems mentioned above for long-term acoustic signals. In this paper, we further develop the recently proposed NMF-based feature extraction method to enable its application in weakly supervised sound event detection. To achieve this goal, we develop a strategy for training the frequency basis matrix using a heterogeneous database consisting of strongly- and weakly-labeled data. Moreover, we develop a non-iterative version of the NMF-based feature extraction method so that the proposed feature extraction method can be applied as a part of the model structure similar to the modern “on-the-fly” transform method for the Mel-spectrogram. To detect the sound events, the temporal basis is calculated using the NMF method and then used as a feature for the mean-teacher-model-based classifier. The results are improved for the event-wise post-processing method. To evaluate the proposed system, simulations of the weakly supervised sound event detection were conducted using the *Detection and Classification of Acoustic Scenes and Events 2020 Task 4* database. The results reveal that the proposed system has F1-score performance comparable with the Mel-spectrogram and gammatonegram and exhibits 3–5% better performance than the log-Mel-spectrum and constant-Q transform.

**Keywords:** feature extraction; sound event detection; non-negative matrix factorization



**Citation:** Lee, S.; Kim, M.; Shin, S.; Park, S.; Jeong, Y. Data-Dependent Feature Extraction Method Based on Non-Negative Matrix Factorization for Weakly Supervised Domestic Sound Event Detection. *Appl. Sci.* **2021**, *11*, 1040. <https://doi.org/10.3390/app11031040>

Received: 18 December 2020

Accepted: 19 January 2021

Published: 24 January 2021

**Publisher’s Note:** MDPI stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.



**Copyright:** © 2021 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

## 1. Introduction

More and more studies have been targeting machine learning and artificial intelligence recently. Machine recognition of environments and sound events using acoustic signals have been of particular interest to researchers [1–4]. There are two main tasks related to the automatic recognition of acoustic signals: acoustic scene classification (ASC) and sound event detection (SED). These tasks are often not clearly distinguished. ASC mainly focuses on the recognition of long clips, e.g., a 10-s clip, to classify the whole acoustic environment [5], whereas SED tends to analyze short sound events, e.g., dog barking or alarm ringing, to determine their types and obtain onset/offset information [6].

The extraction of proper features from the acoustic signals is the first and important step in SED using machine learning algorithms. The most common acoustic features for ASC and SED are Mel-frequency cepstral coefficients (MFCC) [1,7,8] and Mel-frequency

spectrum [9–11]. These are the variations of frequency-domain features used for a characterization of the human hearing system [12]. Mel-frequency-based features have been successfully used in speech signal processing systems employing machine learning techniques. The features exhibit better performance in ASC and SED compared with other existing ones. However, their performance is limited by the less-structured acoustic environmental signals compared with speech signals [5]. Therefore, several alternative features, including a computer-vision-inspired feature [13] and statistical characteristics [14], have been investigated. Recently, several features inspired by the characteristics of the human hearing system, e.g., Mel-frequency discrete wavelet coefficients [15–17], gammatonegram [18], and gammatone-frequency cepstral coefficients [19], have been investigated. The vast majority of the developed features, including the MFCC and psycho-acoustic-based features, are data-independent feature extraction methods because the feature extraction processes are consistent regardless of the data characteristics in the given problems.

Recently, several data-dependent feature extraction methods, including principal component analysis (PCA) [20] and non-negative matrix factorization (NMF) [21], have also been developed. The NMF method has been widely employed to analyze and extract the signal characteristics in the recent acoustic signal processing fields, including music information retrieval [22–24] and speech signal processing [25–27].

As the NMF method can extract the common property of the given signals, data-dependent feature extraction methods based on NMF have been developed [5,28,29]. The method developed in [5] is a supervised task-driven dictionary learning (TDL) with a limited-memory Broyden–Fletcher–Goldfarb–Shanno (L-BFGS) algorithm [30]. However, feature extraction using the TDL method is strongly related to the classification step. Thus, it is difficult to apply the TDL feature extraction method to other techniques, such as the recently suggested convolutional neural network-based classifiers. Unsupervised NMF-based feature extraction methods were also developed [28,29]. The method in [29] capitalized on the convolutional NMF and K-means clustering to deal with the uncategorized dataset. The disadvantages of the unsupervised NMF-based feature extraction method are that the data matrix to be handled is too large and the computational cost is quite high. To overcome these disadvantages, Lee and Pang developed a supervised feature extraction method for the monitoring domestic activity problem [31]. The algorithm in [31] exhibited a performance comparable to the state-of-the-art features. However, this study was limited to monitoring domestic activities, which included the activity class recognition problem (without onset/offset information) only.

In this paper, we develop and analyze a data-dependent feature extraction method for weakly supervised domestic SED. To achieve this goal, we start with the NMF-based feature extraction method [31] for the monitoring domestic activity problem. Unfortunately, this method [31] cannot be directly applied to our problem, because the configuration of the training data is different. Therefore, we develop and analyze several strategies to utilize the data for feature extraction. In addition, we consider the recent trend of making the feature extraction step a part of the neural network or developing “on-the-fly” feature extraction systems [32,33] in the acoustic signal processing. The conventional NMF-based feature extraction methods [5,31] cannot be applied to “on-the-fly” systems because they require dozens or hundreds of iterations. To overcome this problem, we develop a matrix multiplication-based feature extraction method without any iteration.

## 2. Background

### 2.1. Problem Description

As mentioned in the Introduction, we aim to develop a data-dependent feature extraction method for weakly supervised domestic SED system. The target system and problem are designed in accordance with the Detection and Classification of Acoustic Scenes and Events (DCASE) 2020 Task 4b [34]. The target system aims to detect sound events with class, onset time, and offset time labels through a weakly-supervised training. In this training, some of the dataset annotations are omitted. More specifically, the dataset may be

categorized into three types: strongly-labeled, weakly-labeled, and unlabeled data. The strongly-labeled data provide all the required information: sound class, onset time, and offset time annotations. The weakly-labeled data provide only part of the information, for example the sound class label only. The unlabeled data do not provide any information, only waveforms. A schematic diagram of the target system is presented in Figure 1.

Our goal is not to design the system itself but to develop a data-dependent feature extraction method for the system. Thus, we focus on the non-negative matrix decomposition technique, which is a sparse representation tool for acoustic signal data.

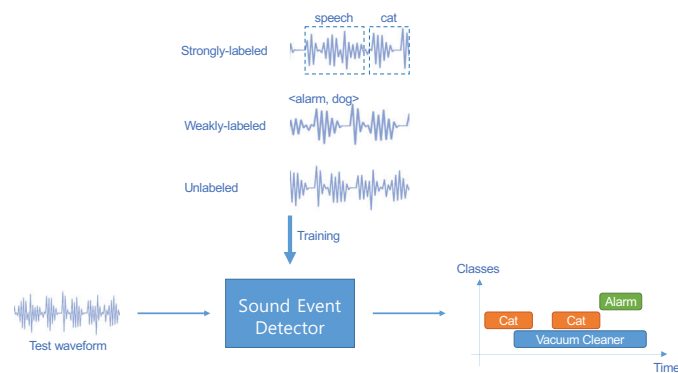


Figure 1. Problem description.

2.2. Non-Negative Matrix Factorization

The NMF method, which was developed by Lee and Seung [21,35], is employed to decompose a non-negative matrix  $V \in \mathbb{R}^+_{K \times N}$  into two matrices,  $W \in \mathbb{R}^+_{K \times R}$  and  $H \in \mathbb{R}^+_{R \times N}$ , that satisfy the following relation [35]:

$$V \approx WH. \tag{1}$$

The matrices  $W$  and  $H$  can be estimated via alternating optimization of the two equations as [35,36]

$$W = \arg \min_W D(V|WH) \quad \text{for fixed } H, \tag{2}$$

$$H = \arg \min_H D(V|WH) \quad \text{for fixed } W, \tag{3}$$

where  $D(V|WH)$  denotes the distance function between  $V$  and  $WH$ . The distance functions can be optimized by the multiplicative update rule as [35]

$$W \leftarrow W \otimes \frac{[V/(WH)]H^T}{\mathbf{1}_{K \times N}H^T}, \tag{4}$$

$$H \leftarrow H \otimes \frac{W^T[V/(WH)]}{W^T\mathbf{1}_{K \times N}}, \tag{5}$$

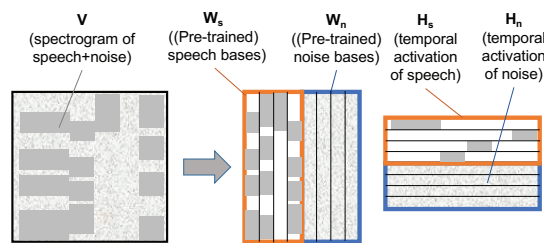
when the Kullback–Leibler divergence (KLD) is used as the distance function, where  $\mathbf{1}_{K \times N}$  denotes a  $K \times N$  matrix of ones, and

$$W \leftarrow W \otimes \frac{VH^T}{WHH^T}, \tag{6}$$

$$H \leftarrow H \otimes \frac{W^TV}{W^TWH}, \tag{7}$$

when the Euclidean distance is used, where  $\otimes$  and the fraction denote the Hadamard (element-wise) product and division, respectively.

The NMF method has been widely employed in the recent acoustic signal processing studies to analyze a music signal into several musical notes or to reduce noise signals from the speech signals. Such interest can be attributed to the fact that the frequency basis matrix,  $\mathbf{W}$ , and the temporal basis matrix,  $\mathbf{H}$ , represent the frequency characteristics and temporal envelop of acoustic signal components, respectively. More specifically, speech denoizing methods divide the bases into two classes, speech and noise, and remove the noise bases after calculating the temporal bases of each class, as shown in Figure 2.



**Figure 2.** Schematic diagram of the applications of the non-negative matrix factorization (NMF) methods to the acoustic signal processing systems.

### 3. Proposed System

#### 3.1. Strategy for the Frequency Basis Learning

Similar to the previous studies [5,31], the temporal basis matrix  $\mathbf{H}$  is used as a feature. To extract the feature matrix, the frequency basis matrix needs to be estimated in advance. Inspired by speech denoizing [27] and reverberation suppression [37] techniques based on NMF [27], the frequency basis matrix of each class is independently estimated from the spectrogram set of the class, as presented in Figure 3a. In this method, the frequency basis matrix is divided into  $C$  groups as [31]

$$\mathbf{W} = [\mathbf{W}_1 \quad \mathbf{W}_2 \quad \cdots \quad \mathbf{W}_C]. \tag{8}$$

Each group denoted by  $\mathbf{W}_c$  is a  $K \times R_c$  matrix, where  $R_c$  denotes the basis number of each class. The frequency and temporal basis matrices can be estimated by optimizing various cost functions. KLD is one of the common choices in the acoustic signal processing field. Therefore, the frequency and temporal basis matrices are iteratively updated by

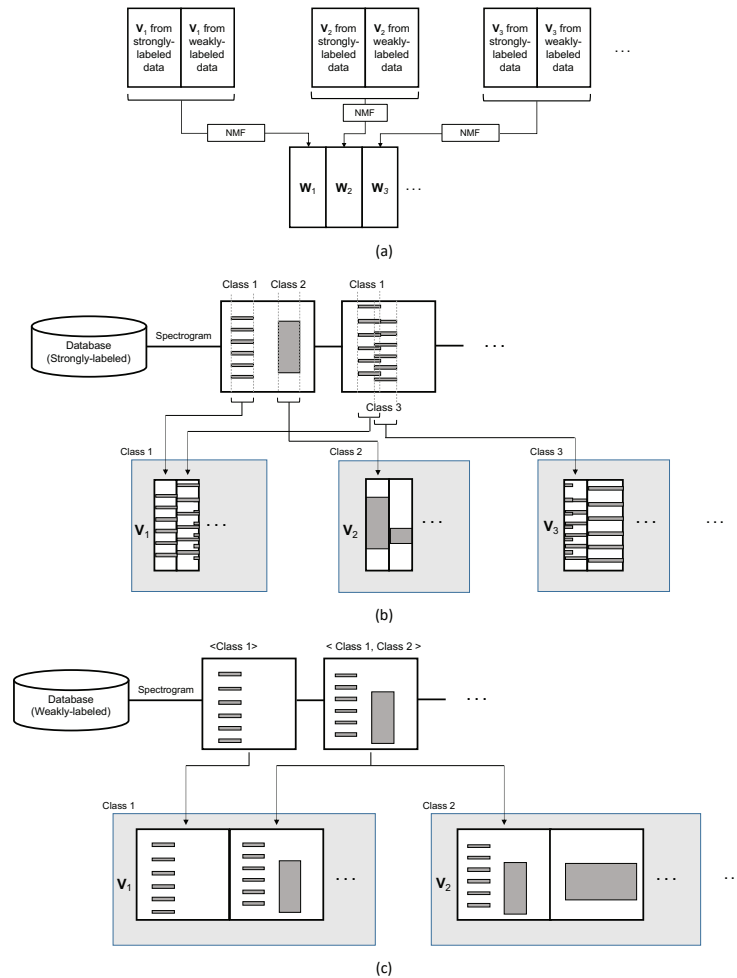
$$\mathbf{W}_c \leftarrow \mathbf{W}_c \otimes \frac{[\mathbf{V}_c / (\mathbf{W}_c \mathbf{H}_c)] \mathbf{H}_c^T}{\mathbf{1}_{K \times N_c} \mathbf{H}_c^T} \tag{9}$$

$$\mathbf{H}_c \leftarrow \mathbf{H}_c \otimes \frac{\mathbf{W}_c^T [\mathbf{V}_c / (\mathbf{W}_c \mathbf{H}_c)]}{\mathbf{W}_c^T \mathbf{1}_{K \times N_c}}, \tag{10}$$

which is similar to Equations (4) and (5), where  $\mathbf{V}_c$  and  $\mathbf{H}_c$  denote  $K \times N_c$  data matrix and  $R_c \times N_c$  temporal basis matrix, respectively, and  $N_c$  stands for the total number of frames in the data matrix of the  $c$ th class.

Since weakly-supervised SED has a heterogeneous database, different learning strategies for each type of data need to be developed. In the case of strongly-labeled data, which contains the class and temporal information, the data matrix is a set of audio clips with a cut-off part, as presented in Figure 3b. As can be seen in the figure, there is a risk of data mixing from different classes. For example, if we assume that we want to generate the data matrix  $\mathbf{V}$  for Class A, there may be various combinations, e.g., A–B, A–C, and A–D. However, even in this case, the data from Class A is expected to be dominant. Thus, the

frequency matrix  $W$  can be expected to represent the characteristics of Class A if we choose a sufficiently small rank for matrix  $W$ .



**Figure 3.** Block diagrams for: (a) the learning of the frequency basis; (b) the learning of the composition of the data matrix from strongly-labeled data; and (c) the learning of the composition of the data matrix from weakly-labeled data.

For the weakly-labeled data that do not contain an onset and offset information, the temporal information cannot be utilized. Therefore, we compose the data matrix  $V$  by just cascading the clips, as shown in Figure 3c. Unlike the case of the strongly-labeled data, where only parts of the event overlap, in the weakly-labeled data, waveforms from different classes are completely mixed when two or more classes are in one clip. Therefore, it is expected that the training of matrix  $W$  is more affected by the unwanted class interference problem compared to strongly-labeled data. Therefore, we evaluate the effect of the unwanted class interference problem on the weakly-labeled data.



### 3.2. Iterative and Non-Iterative Feature Extraction Methods

Once the frequency basis matrix  $\mathbf{W}$  is trained, the feature matrix  $\mathbf{H}_{clip}$  can be extracted from the data matrix  $\mathbf{V}_{clip}$  of an audio clip via the iteration of [31]

$$\mathbf{H}_{clip} \leftarrow \mathbf{H}_{clip} \otimes \frac{\mathbf{W}^T [\mathbf{V}_{clip} / (\mathbf{W}\mathbf{H}_{clip})]}{\mathbf{W}^T \mathbf{1}_{K \times N_{clip}}}. \tag{11}$$

with minimization of the KLD (as Equation (5)).

In this paper, we also develop a non-iterative feature extraction method. To achieve this goal, we develop a closed-form solution for  $\mathbf{H}_{clip}$  which makes the gradient of the divergence function with regard to  $\mathbf{H}_{clip}$  zero (that is,  $\nabla_{\mathbf{H}_{clip}} D(\mathbf{V}_{clip} | \mathbf{W}\mathbf{H}_{clip}) = 0$ ). In addition, a solution with simple operations, e.g. matrix multiplication, is a preferred one because it is easy to implement.

The NMF-based methods for acoustic signal processing mainly use three types of distance functions: KLD, Euclidean distance, and Itakura–Saito divergence (ISD). The three distance functions are in the  $\beta$ -divergence family, which is expressed by [38]:

$$D_{\beta}(x|y) = \begin{cases} \frac{x}{y} - \log \frac{x}{y} - 1 & \beta = 0 \\ x(\log x - \log y) + (y - x) & \beta = 1 \\ \frac{1}{\beta(\beta-1)}(x^{\beta} + (\beta-1)y^{\beta} - \beta xy^{\beta-1}) & \text{otherwise} \end{cases} \tag{12}$$

where  $x$  and  $y$  are arbitrary values and  $\beta$  is a constant to adjust the cost function among the Euclidean ( $\beta = 2$ ), KLD ( $\beta = 1$ ), and ISD ( $\beta = 0$ ). In our problem,  $x$  and  $y$  are elements of the matrices  $\mathbf{V}_{clip}$  and  $\mathbf{W}\mathbf{H}_{clip}$ , respectively, and the gradient of the beta-divergence with regard to  $\mathbf{H}_{clip}$  is [38]

$$\nabla_{\mathbf{H}_{clip}} D_{\beta}(\mathbf{V}_{clip} | \mathbf{W}\mathbf{H}_{clip}) = \mathbf{W}^T \left\{ (\mathbf{W}\mathbf{H}_{clip})^{[\beta-2]} \otimes (\mathbf{W}\mathbf{H}_{clip} - \mathbf{V}_{clip}) \right\} \tag{13}$$

where  $\mathbf{A}^{[n]}$  denotes element-wise  $n$ th power of matrix  $\mathbf{A}$ . Because Equation (13) includes matrix multiplications, element-wise products and power, and the relationship between these operations is also complex. Therefore, it is not easy to find a solution that makes Equation (13) zero, and it is even more difficult to find a solution consisting of simple matrix multiplications. One easy way is to make the term of the left of the Hadamard product,  $(\mathbf{W}\mathbf{H})^{[\beta-2]}$ , equal to one. In this case,  $\beta$  becomes 2.0, and the divergence function becomes the Euclidean distance.

The Euclidean distance function can be defined using the Frobenius norm as

$$D_{EUC}(\mathbf{V}_{clip} | \mathbf{W}\mathbf{H}_{clip}) = \frac{1}{2} \|\mathbf{V}_{clip} - \mathbf{W}\mathbf{H}_{clip}\|_F^2, \tag{14}$$

and the gradient of the cost function with respect to  $\mathbf{H}_{clip}$  is

$$\nabla_{\mathbf{H}_{clip}} D_{EUC}(\mathbf{V}_{clip} | \mathbf{W}\mathbf{H}_{clip}) = -\mathbf{W}^T [\mathbf{V}_{clip} - \mathbf{W}\mathbf{H}_{clip}]. \tag{15}$$

Since the cost function in Equation (14) is convex, the optimal solution of the cost function makes the gradient zero. Therefore, the optimal solution of  $\mathbf{H}_{clip}$  needs to satisfy the relationship:

$$\mathbf{W}^T \mathbf{V}_{clip} - \mathbf{W}^T \mathbf{W}\mathbf{H}_{clip} = 0 \tag{16}$$

and therefore

$$\begin{aligned} {}^{184}\mathbf{H}_{clip} &= [\mathbf{W}^T \mathbf{W}]^{-1} \mathbf{W}^T \mathbf{V}_{clip} \\ &= \mathbf{W}^+ \mathbf{V}_{clip}, \end{aligned} \tag{17}$$

where  $\mathbf{W}^\dagger$  denotes the Moore–Penrose pseudo-inverse of the matrix  $\mathbf{W}$ . Since the frequency basis matrix  $\mathbf{W}$  is pre-trained and fixed,  $\mathbf{W}^\dagger$  is also defined a priori. Thus, feature extraction can be performed via a simple production using Equation (17).

The pseudo-inverse can be implemented via the singular value decomposition [39]. If we assume that the matrix  $\mathbf{W}$  is decomposed as

$$\mathbf{W} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^H, \tag{18}$$

where  $\mathbf{U}$  and  $\mathbf{V}$  denote  $K \times K$  and  $R \times R$  matrices, respectively, and  $\mathbf{\Sigma}$  denotes the  $K \times R$  diagonal matrix whose elements are singular values of  $\mathbf{W}$ , then the pseudo-inverse  $\mathbf{W}^\dagger$  can be calculated as

$$\mathbf{W}^\dagger = \mathbf{V}\mathbf{\Sigma}^\dagger\mathbf{U}^H, \tag{19}$$

where  $\mathbf{\Sigma}^\dagger$  is a matrix formed from  $\mathbf{\Sigma}$  by taking the element-wise inverse of the non-zero elements and shaping the matrix as  $R \times K$ . In practice, the small singular values can be ignored as

$$[\mathbf{\Sigma}^\dagger]_{r,k} = \begin{cases} 1/[\mathbf{\Sigma}]_{k,r} & \text{if } [\mathbf{\Sigma}]_{k,r} > \delta, \\ 0 & \text{otherwise,} \end{cases} \tag{20}$$

where  $[\mathbf{\Sigma}]_{k,r}$  denotes the  $(k, r)$ th element of the matrix  $\mathbf{\Sigma}$  and  $\delta$  is an arbitrary threshold.

### 3.3. Classifier

To develop a feature extraction method and evaluate the performance of our method, a conventional convolutional recurrent neural network (CRNN) with a mean-teacher model is adopted [40–42]. Figure 4 presents the details of the network structure of the CRNN classifier. The convolutional layers and the corresponding max-pooling layers are designed so that the  $n_{feature}$  axis of the output of the CNN layers is equal to unity. The RNN layers and the fully connected layer make the frame-wise probability of the classes, which is denoted by “Time Stamps” in Figure 4. To train the classifier using the weakly-labeled data, then it also generates the “Clip Class” output, which denotes the existing classes in a clip. The element of the clip class output  $\mathbf{C}$  is calculated by weighted average as

$$[\mathbf{C}]_c = \frac{\sum_{n=0}^{N_{out}-1} [\mathbf{P}]_{n,c} [\mathbf{P}_{softmax}]_{n,c}}{\sum_{n=0}^{N_{out}-1} [\mathbf{P}_{softmax}]_{n,c}}, \tag{21}$$

where  $N_{out}$  denotes the frame number of the classifier output of a sound clip.

The mean-teacher model is adopted to train the classifier using both the labeled and unlabeled data, similar to state-of-the-art weakly-supervised SED systems [40,41]. The network parameters of the student model,  $\theta$ , are optimized to minimize the cost function as

$$J_{total}(\theta) = J_{class}(\theta) + \beta J_{consist}(\theta), \tag{22}$$

where  $J_{class}(\theta)$  and  $J_{consist}(\theta)$  denote the classification cost and the consistency cost, respectively, as presented in Figure 5.  $\beta$  is the weight for the consistency cost. The parameters of the teacher model,  $\theta_{teacher}$ , are fixed during the training procedure and updated at the end of the training batch as

$$\theta_{teacher} \leftarrow \alpha \theta_{teacher} + (1 - \alpha) \theta \tag{23}$$

where  $\alpha$  denotes the exponential weighting factor of the moving average.

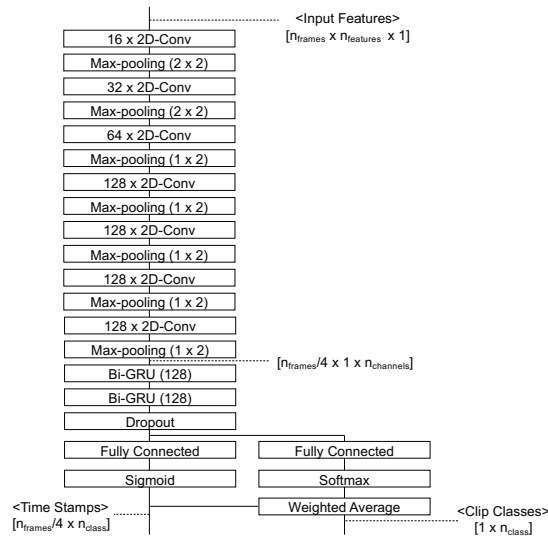


Figure 4. Block diagram of the network structure for the CRNN classifier.

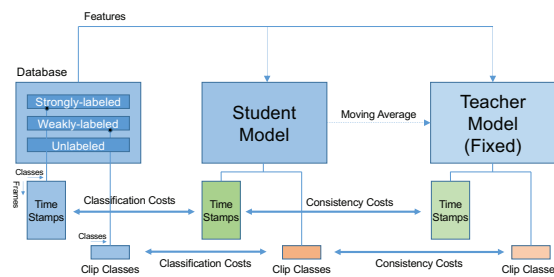


Figure 5. Block diagram of the mean-teacher model.

### 3.4. Post-Processing

The output of the sound event detector is a frame-wise probability of each class. Generally, the classifier output is post-processed with a certain threshold to determine whether the class is activated or not. The magnitude of the threshold value influences the classification performance. For example, a high threshold value can reduce false-positive errors but sometimes underestimates the length of the event, and vice versa. In this paper, the classifier output is first processed with a frame-wise threshold and then with an event-wise threshold to eliminate false-positive errors. The threshold value of the event-wise post-processing is set to be larger than that of the frame-wise post-processing.

When we refer to the classifier output as a matrix  $\mathbf{P}$  of  $N_{out} \times C$ , the frame-wise binary matrix  $\mathbf{B}_{frame}$  is calculated as

$$[\mathbf{B}_{frame}]_{n,c} = \begin{cases} 1 & \text{if } [\mathbf{P}]_{n,c} \geq \tau_{frame}, \\ 0 & \text{if } [\mathbf{P}]_{n,c} < \tau_{frame}, \end{cases} \quad (24)$$

where  $C$  denotes the number of classes and  $\tau_{frame}$  is the threshold for the frame-wise post-processing. Let us define  $\chi(i, c)$  as the  $i$ th event set of class  $c$  as

$$\chi(i, c) = \left\{ (n, c) \mid n_{onset}(i, c) \leq n \leq n_{offset}(i, c) \right\}, \quad (25)$$

where  $n_{onset}(i, c)$  and  $n_{offset}(i, c)$  are the frame numbers of the  $i$ th onset and offset of class  $c$ , respectively. The result of the event-wise thresholding,  $\mathbf{B}_{event}$ , is obtained as

$$[\mathbf{B}_{event}]_{n,c} = \begin{cases} 1 & \text{if } (\max_{(n',c) \in \chi'} [\mathbf{P}]_{n',c}) \geq \tau_{event} \\ 0 & \text{otherwise} \end{cases}, \quad (26)$$

where  $(n', c) \in \chi'$  means that  $(n', c)$  belongs to  $\chi'$ , and  $\chi'$  denotes a certain event set that contains  $(n, c)$  as an element. Figure 6 presents a flow chart and an illustrative example of the proposed event-wise post-processing.

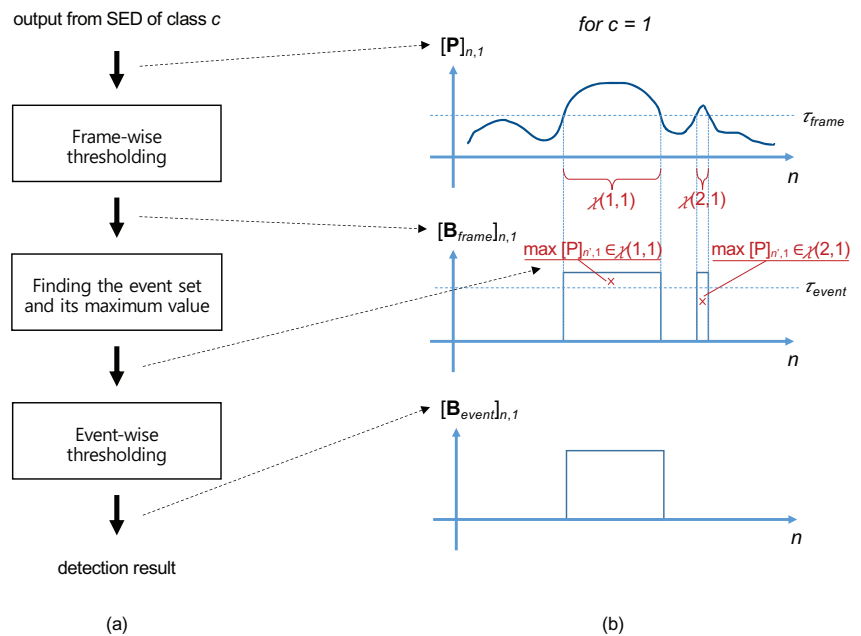


Figure 6. (a) A flow chart; and (b) an illustrative example of the event-wise post-processing.

#### 4. Evaluation

##### 4.1. Evaluation Settings

To evaluate the proposed system for domestic SED, numerical simulations were performed using the DESED dataset of the DCASE 2020 Task 4 database, which is an audio dataset for weakly supervised sound event detection in domestic environments. The database consists of real-recorded subsets of Audioset [43] and Freesound [44]. Some of the data are synthesized using the SINS database [45] as background sounds.

There are 10 event classes: speech, dog, cat, alarm bell, dishes, frying, blender, running water, vacuum cleaner, and electric shaver. Each audio clip is a 10-s wav file with 44.1 (for the weakly-labeled and unlabeled data) or 16 kHz (for strongly-labeled data) sampling frequency. All audio files were resampled to a 16 kHz sampling frequency. The database consists of strongly-labeled data (2045 files, with event classes and time stamp information), weakly-labeled data (1578 files, with event classes information only), and unlabeled data (14412 files, without any annotation). The detailed information of the database configuration can be found in [34].

The audio clips were short-time-Fourier-transformed by a 1024-sample Hanning window with 75% overlap to match the number of input frames in the classifier for the baseline of the DCASE 2020 Task 4 [40]. To train the NMF frequency basis matrix, 200 iterations were performed to calculate both  $\mathbf{W}_c$  and  $\mathbf{H}_c$  matrices (Equations (9) and (10)) for

each class. The number of the basis for each class was set to 13 (11 for the “vacuum cleaner” class) so that the dimension of the extracted feature was 128. In the case of iterative feature extraction, 50 iterations were performed for each clip to calculate  $\mathbf{H}_{clip}$  (Equation (11)). The singular value threshold ( $\delta$  in Equation (20)) was set to be a proportional to the maximum singular value as

$$\delta = \gamma \max_{k,r} [\Sigma]_{k,r} \tag{27}$$

where  $\gamma$  denotes a proportionality constant.

The classifier was trained using Adam optimizer, and the learning rate was exponentially ramped up during 50 epochs to the maximum learning rate of 0.001. The classification and the consistency costs were categorical cross-entropy and mean-squared error, respectively, and the weight for the consistency costs,  $\beta$ , was set to 2.0. The classifier was trained using the training dataset of the DCASE database for 200 epochs and evaluated using the validation dataset. The batch size was set to 24, and the moving-average weighting factor,  $\alpha$ , of the mean-teacher model was 0.999.

The performances of the two types of post-processing systems, with and without the event-wise post-processing, were evaluated. The post-processing system without the event-wise post-processing consisted of only the frame-wise thresholding, and the threshold value was set to 0.5. The system with the event-wise post-processing consisted of the frame-wise and event-wise thresholding. The event threshold value was set to 0.8 as it exhibited good overall performance, while the frame thresholds were set to the optimal values that exhibited the best performance among  $\{0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7\}$  for each class.

In this evaluation, the event-based F1-score was utilized as the performance measure, which is defined by a geometric mean of precision  $P$  and recall  $R$  as

$$F_1 = \frac{2PR}{P + R} \tag{28}$$

where precision and recall are defined as

$$P = \frac{n_{TP}}{n_{TP} + n_{FP}} \tag{29}$$

and

$$R = \frac{n_{TP}}{n_{TP} + n_{FN}} \tag{30}$$

where  $n_{FP}$ ,  $n_{FN}$ , and  $n_{TP}$  denote the numbers of false answers (false positives), missing answers (false negatives), and correct answers (true positives). The answers were considered as “correct answer” if the onset error was smaller than 0.2 s, and the offset error was smaller than 0.2 s (or 20 % of the event duration). The criteria to determine the correct answers were set to be consistent with the baseline of the DCASE 2020 Task 4 [34]. Moreover, the performance of each class was micro-averaged, which aggregated the contributions of all the test samples regardless of the class, and macro-averaged, which averages the individual performance of each class.

#### 4.2. Comparison of Various Features

Table 1 compares the proposed and conventional features that are commonly used in ASC and SED applications. The abbreviations MelSpec, Log-Mel, GAM, and CQT denote the Mel-spectrogram, log-Mel spectrum, gammatonegram, and constant-Q transform (CQT), respectively. The frequency basis matrix used to extract the NMF features in Table 1 was trained using strongly-labeled data only as it exhibited good overall performance. The effect of the data on the frequency basis training will be considered in the next subsection. The dimensions of the extracted features were all set to 128, similar to that in the Delphin-Poulet system [40], to compare the performances of the features without changing the structure of the classifier. The CQT was performed under 16 bins per octave and 32.7 Hz

lower bound frequency. In total, 128 CQT bins (=8 octaves) corresponded to a frequency range of less than approximately 8 kHz. The proposed system was also compared with the Cornell system [46] including the log-Mel features with data augmentation, per-channel energy normalization (PCEN), mean-teacher CRNN, and post-processing with hidden Markov model (HMM), which was submitted to DCASE 2020 Task 4. The details of the sub-systems can be found in [46]. The training parameters (e.g., the weight for the consistency costs, number of epochs, batch size, and the moving-average weighting factor of the mean-teacher model) were set to the same values of the proposed system. Because the Cornell system has its own post-processing consisting of HMM and random forest optimization, the proposed event-wise post-processing was not applied to the Cornell system.

**Table 1.** Averaged results of various features. The performance of the Cornell system was provided as a reference of comparison. The boldface means the best performance of each measure.

	w/o Event-Wise Post-Processing		w/ Event-Wise Post-Processing	
	F1-Score [%] (Micro)	F1-Score [%] (Macro)	F1-Score [%] (Micro)	F1-Score [%] (Macro)
NMF(iterative)	<b>35.06</b>	31.58	40.12	39.23
NMF(non-iterative)	34.87	30.16	40.02	38.45
MelSpec	34.41	32.31	<b>40.41</b>	39.72
Log-Mel	30.27	29.88	35.11	36.60
GAM	32.15	<b>33.09</b>	37.23	<b>39.81</b>
CQT	32.25	28.76	37.28	35.36
Cornell et al. [46]	-	-	(with own post-processing) 42.48	39.56

The performances of the NMF methods are comparable to those of the Mel-spectrogram and gammatonegram, the best among the conventional features, and are better than those of the log-Mel spectrum and CQT in all the performance measures. The comparison between the micro-averaged performance of the NMF and that of the Mel-spectrogram revealed that the NMF methods exhibited slightly better performances without the event-wise post-processing and slightly worse performances with the event-wise post-processing. The gammatonegram had the best performance in the macro-averaged F1-score among the features but demonstrated low performance in the micro-averaged F1-score. Moreover, the features extracted using the iterative and non-iterative NMF methods exhibited similar performances. Thus, the non-iterative NMF method can be considered as an alternative to the iterative NMF method. The micro-averaged and macro-averaged F1-scores of the proposed system are similar to and slightly less than that of the Cornell system, respectively. We think that the difference of the performance was caused by additional sub-systems in the Cornell system, such as PCEN, data augmentation, and random forest optimization of HMM post-processing.

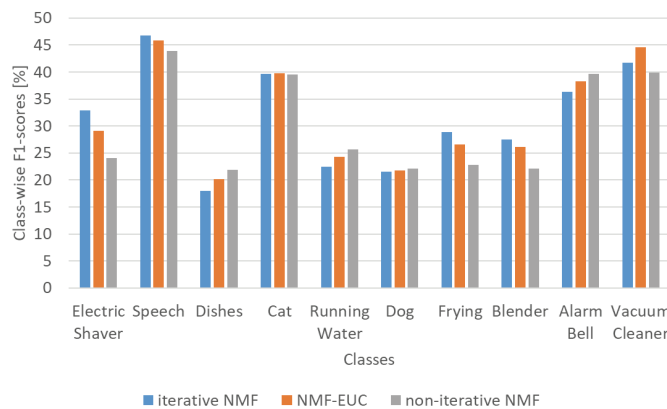
Table 2 presents the class-wise F1-score performances. The gray background denotes the class where the NMF method exhibits better performance than that in the Mel-spectrogram, and the bold-face number stands for the best performance in each class. As can be seen in the table, the NMF methods are advantageous for classes with harmonic structures (speech, dog, cat, and alarm bell) and disadvantageous for noise-like classes (electric shaver, blender, running water, and vacuum cleaner). The class-wise performances of the iterative and non-iterative NMFs are different. The *iterative* NMF has better performances for five classes (electric shaver, speech, frying, blender, and vacuum cleaner), similar performances for two classes (cat and dog), and worse performances for three classes (dishes, running water, and alarm bell). The last three classes (dishes, running water, and alarm bell) contained several impulsive sounds. Thus, the *non-iterative* method

appeared to be useful for the impulsive sounds. However, we cannot easily come to this conclusion, as the frying sounds also contain impulsive ones.

**Table 2.** Class-wise F1-scores [%] of various features. The boldface means the best performance of each class.

	Electric Shaver	Speech	Dishes	Cat	Running Water	Dog	Frying	Blender	Alarm Bell	Vacuum Cleaner
NMF (iterative)	32.9	<b>46.8</b>	18.0	39.7	22.4	21.5	28.9	27.5	36.3	41.7
NMF (non-iterative)	24.1	43.9	21.9	39.5	25.7	<b>22.1</b>	22.8	22.1	<b>39.6</b>	39.9
MelSpec	35.7	45.0	18.0	39.1	<b>30.9</b>	17.4	24.0	32.0	32.0	49.0
Log-Mel	<b>38.3</b>	37.3	14.2	36.7	27.4	13.7	23.5	23.0	35.8	49.0
GAM	29.5	34.2	<b>24.6</b>	<b>41.7</b>	28.3	19.7	<b>31.4</b>	<b>33.9</b>	39.4	48.2
CQT	34.6	46.7	18.4	35.9	21.7	16.5	9.1	24.8	24.1	<b>55.6</b>

There are two main differences between the iterative and non-iterative NMFs: divergence function and optimization method. The *iterative NMF* uses KLD and multiplicative update, and the *non-iterative NMF* employs Euclidean and closed-form solution. To analyze the effect on the class-wise performance, we compare the performances of the NMFs with an *iterative NMF with Euclidean* (for convenience, we call it *NMF-EUC* here). The *iterative NMF* and *NMF-EUC* have the same parameters and settings except for the divergence function. Figure 7 shows the comparison results. The *iterative NMF* is superior to the *NMF-EUC* in the classes of electric shaver, speech, frying, and blender, where the *iterative NMF* is superior to the *non-iterative NMF*. In other words, the tendencies in performances are similar for the *NMF-EUC* and *non-iterative NMF*. However, the difference in the performance of *non-iterative NMF* compared to *iterative NMF* is greater than that of *NMF-EUC* and *iterative NMF*. Therefore, it seems that the difference in the optimization method had a greater effect on the performance than that in the divergence function in this experiment.



**Figure 7.** Performance comparison between the *iterative NMF*, *NMF-EUC*, and *non-iterative NMF*.

#### 4.3. Effect of the Training Data on the Frequency Basis Learning

To analyze the effect of the training data on the frequency basis matrices in the NMF methods, the performances of the NMF methods were compared with the different frequency basis matrices from various parts of the database. Table 3 presents the F1-score results of different frequency basis matrices with the event-wise post-processing. STR, WEAK, and WEAK(U) denote the frequency basis matrices trained by the strongly-labeled



data, weakly-labeled data, and weakly-labeled unitary label data, which consist of single-class clips, respectively.

As presented in Table 3, the strongly-labeled data exhibited relatively good overall performance. The weakly-labeled data demonstrated good micro-averaged performance for iterative NMF but degraded performances for the other criteria, including non-iterative NMF. Using both the strongly- and weakly-labeled unitary label data (STR + WEAK(U)) also had good macro-averaged performance for non-iterative NMF, but it also showed degraded performances for other criteria. Conversely, the performance of the weakly-labeled unitary label data (WEAK(U)) did not exceed that of the weakly-labeled data (WEAK) that has interference classes. For instance, Class 2 data interferes with the training of Class 1 frequency basis matrix (Figure 3c). Therefore, we suggest that the class interference problem does not significantly affect the classification performance.

**Table 3.** Comparison results with different frequency basis matrices from various parts of the database.

	NMF (Iterative)		NMF (Non-Iterative)	
	F1-Score [%] (Micro)	F1-Score [%] (Macro)	F1-Score [%] (Micro)	F1-Score [%] (Macro)
STR	40.12	39.23	40.02	38.45
WEAK(U)	40.02	38.89	38.98	37.65
WEAK	41.53	38.28	39.01	37.76
STR + WEAK(U)	38.91	37.50	38.39	38.79
STR + WEAK	38.51	37.39	39.97	38.40

#### 4.4. Thresholding Singular Values for Calculating the Pseudo-Inverse Matrix

To extract the proposed feature instantaneously, the Moore–Penrose pseudo-inverse of the frequency basis matrix needs to be calculated in advance using Equation (17). As follows from Equations (19) and (20), the pseudo-inverse matrix can be calculated via singular value decomposition with thresholding of the small singular values. The threshold, which is one of the design parameters, is related to stability and sparsity of the pseudo-inverse matrix. Thus, it may affect the performance of the extracted features. Therefore, we evaluated the effect of the threshold on the classification performance.

Table 4 presents the classification performances with various threshold values. As follows from Equation (27),  $\delta$  denotes the ratio of the threshold value to the maximum singular value. Among the test values,  $\gamma = 0.01$  exhibits the best performance, while  $\gamma = 0.005$  and  $\gamma = 0.05$  exhibit comparable performances. However,  $\gamma = 0.001$  and  $\gamma = 0.1$  result in significantly degraded performances. Therefore,  $\gamma$  values between 0.005 and 0.05 are good choices for our system.

**Table 4.** Comparison results with various thresholds of singular values for calculating the pseudo-inverse.

	w/o Event-Wise Post-Processing		w/ Event-Wise Post-Processing	
	F1-Score [%] (Micro)	F1-Score [%] (Macro)	F1-Score [%] (Micro)	F1-Score [%] (Macro)
$\gamma = 0.001$	27.26	24.56	35.16	34.24
$\gamma = 0.005$	33.59	29.95	39.59	37.94
$\gamma = 0.01$	34.87	30.16	40.02	38.45
$\gamma = 0.05$	32.52	28.35	38.51	36.23
$\gamma = 0.1$	21.43 <sup>†</sup>	10.88	26.45	17.39

## 5. Conclusions

In this paper, two NMF-based feature extraction methods are proposed for weakly supervised SED. Inspired by the NMF applications in the acoustic signal processing systems capable of analyzing the frequency characteristics of the acoustic signals, the proposed methods were designed to extract features from heterogeneous database for weakly supervised SED. To generate the frequency basis matrix, the class-wise data matrices were composed from strongly- and weakly-annotated data. The class-wise frequency basis matrices were estimated using the NMF algorithm with the KLD, and then cascaded to compose the whole frequency basis matrix. In the iterative feature extraction method, the temporal basis matrix was calculated via iterations of the NMF equations using the whole frequency basis matrix. Moreover, we developed a non-iterative feature extraction method using a least-squares solution of the NMF problem. The classifier was constructed based on the mean-teacher model for the proposed features and enhanced by the proposed event-wise post-processing method.

To evaluate the proposed data-dependent feature extraction methods, simulations of weakly supervised SED were performed using the DCASE 2020 Task 4 database. The proposed methods were compared with the conventional features, e.g., Mel-spectrogram, log-Mel spectrum, gammatonegram, and CQT. Although the proposed features did not outperform other features, they yielded results comparable to those of the Mel-spectrogram and gammatonegram, which are state-of-the-art features. Moreover, they demonstrated 3–5% better F1-score performance than the log-Mel-spectrum and CQT.

**Author Contributions:** Conceptualization, S.L., S.P., and Y.J.; methodology, S.L., validation, S.L., M.K., and S.S.; and data curation, M.K. and S.S. All authors have read and agreed to the published version of the manuscript.

**Funding:** This work was supported by Institute for Information & Communications Technology Promotion (IITP) grant funded by the Korea government (MSIT) (N0. 2017-0-00050, Development of Human Enhancement Technology for auditory and muscle support).

**Institutional Review Board Statement:** Not applicable.

**Informed Consent Statement:** Not applicable.

**Data Availability Statement:** Publicly available datasets were analyzed in this study. This data can be found here: <https://project.inria.fr/desed/dcase-challenge/dcase-2020-task-4/>.

**Conflicts of Interest:** The authors declare no conflict of interest.

## References

1. Chu, S.; Narayanan, S.; Kuo, C.C.J.; Mataric, M.J. Where am I? Scene recognition for mobile robots using audio features. In Proceedings of the 2006 IEEE International Conference on Multimedia and Expo, Toronto, Canada, 9–12 July 2006; pp. 885–888.
2. Ellis, D.P.; Lee, K. Minimal-impact audio-based personal archives. In Proceedings of the the 1st ACM Workshop on Continuous Archival and Retrieval of Personal Experiences, New York, USA, 15 October 2004; pp. 39–47.
3. Barchiesi, D.; Giannoulis, D.; Stowell, D.; Plumbley, M.D. Acoustic scene classification: Classifying environments from the sounds they produce. *IEEE Signal Process. Mag.* **2015**, *32*, 16–34.
4. Mesaros, A.; Heittola, T.; Virtanen, T. A multi-device dataset for urban acoustic scene classification. In Proceedings of the Detection and Classification of Acoustic Scenes and Events 2018 Workshop (DCASE2018), Surrey, UK, 19–20 November 2018; pp. 9–13.
5. Bisot, V.; Serizel, R.; Essid, S.; Richard, G. Feature learning with matrix factorization applied to acoustic scene classification. *IEEE/ACM Trans. Audio Speech Lang. Process.* **2017**, *25*, 1216–1229.
6. Cakir, E.; Parascandolo, G.; Heittola, T.; Huttunen, H.; Virtanen, T. Convolutional recurrent neural networks for polyphonic sound event detection. *IEEE/ACM Trans. Audio Speech Lang. Process.* **2017**, *25*, 1291–1303.
7. Huang, Z.; Jiang, D. *Acoustic Scene Classification Based on Deep Convolutional Neuralnetwork with Spatial-Temporal Attention Pooling*; Technical Report; DCASE2019 Challenge; DCASE community, 2019.
8. Liu, H.; Wang, F.; Liu, X.; Guo, D. *An Ensemble System for Domestic Activity Recognition*; Technical Report; DCASE2018 Challenge; DCASE community, 2018.
9. Chen, H.; Liu, Z.; Liu, Z.; Zhang, P.; Yan, Y. *Integrating the Data Augmentation Scheme with Various Classifiers for Acoustic Scene Modeling*; Technical Report; DCASE2019 Challenge; DCASE community, 2019.

10. Inoue, T.; Vinayavekhin, P.; Wang, S.; Wood, D.; Greco, N.; Tachibana, R. *Domestic Activities Classification Based on CNN Using Shuffling and Mixing Data Augmentation*; Technical Report; DCASE2018 Challenge; DCASE community, 2018.
11. Valenti, M.; Squartini, S.; Diment, A.; Parascandolo, G.; Virtanen, T. A convolutional neural network approach for acoustic scene classification. In Proceedings of the 2017 International Joint Conference on Neural Networks (IJCNN), Anchorage, Alaska, USA, 14–19 May 2017; pp. 1547–1554.
12. Moore, B.C. *An Introduction to the Psychology of Hearing*; Brill Academy Press: Leiden, Netherlands, 2012.
13. Bisot, V.; Essid, S.; Richard, G. HOG and subband power distribution image features for acoustic scene classification. In Proceedings of the 2015 23rd European Signal Processing Conference (EUSIPCO), Nice, France, 31 August–4 September 2015; pp. 719–723.
14. Roma, G.; Nogueira, W.; Herrera, P.; de Boronat, R. Recurrence quantification analysis features for auditory scene classification. *IEEE Aasp Chall. Detect. Classif. Acoust. Scenes Events* **2013**, doi:10.1109/WASPAA.2013.6701890.
15. Gowdy, J.N.; Tufekci, Z. Mel-scaled discrete wavelet coefficients for speech recognition. In Proceedings of the 2000 IEEE International Conference on Acoustics, Speech, and Signal Processing, (Cat. No. 00CH37100), Istanbul, Turkey, 5–9 June. 2000; Volume 3, pp. 1351–1354.
16. Waldekar, S.; Saha, G. *Wavelet Transform Based Mel-scaled Features for Acoustic Scene Classification*; Interspeech 2018, Hyderabad, India, 2–6 September. 2018; pp. 3323–3327.
17. Waldekar, S.; Kumar, A. K.; Saha, G. *Mel-Scaled Wavelet-Based Features for Sub-Task A and Texture Features for Sub-Task B of DCASE 2020 Task 1*; Technical Report; DCASE2020 Challenge; DCASE community, 2020.
18. Phan, H.; Koch, P.; Katzberg, F.; Maass, M.; Mazur, R.; Mertins, A. Audio scene classification with deep recurrent neural networks. *arXiv* **2017**, arXiv:1703.04770.
19. Valero, X.; Alias, F. Gammatone cepstral coefficients: Biologically inspired features for non-speech audio classification. *IEEE Trans. Multimed.* **2012**, *14*, 1684–1689.
20. Abdi, H.; Williams, L.J. Principal component analysis. *Wiley Interdiscip. Rev. Comput. Stat.* **2010**, *2*, 433–459.
21. Lee, D.D.; Seung, H.S. Learning the parts of objects by non-negative matrix factorization. *Nature* **1999**, *401*, 788–791.
22. Smaragdis, P.; Brown, J.C. Non-negative matrix factorization for polyphonic music transcription. In Proceedings of the 2003 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics, New Paltz, New York, USA, 19–22 October 2003; pp. 177–180.
23. Bertin, N.; Badeau, R.; Vincent, E. Enforcing harmonicity and smoothness in Bayesian non-negative matrix factorization applied to polyphonic music transcription. *IEEE Trans. Audio Speech Lang. Process.* **2010**, *18*, 538–549.
24. Benetos, E.; Dixon, S.; Giannoulis, D.; Kirchhoff, H.; Klapuri, A. Automatic music transcription: challenges and future directions. *J. Intell. Inf. Syst.* **2013**, *41*, 407–434.
25. Wilson, K.W.; Raj, B.; Smaragdis, P.; Divakaran, A. Speech denoising using nonnegative matrix factorization with priors. In Proceedings of the 2008 IEEE International Conference on Acoustics, Speech and Signal Processing, Las Vegas, Nevada, USA, 30 March–4 April 2008; pp. 4029–4032.
26. Kwon, K.; Shin, J.W.; Kim, N.S. NMF-based speech enhancement using bases update. *IEEE Signal Process. Lett.* **2014**, *22*, 450–454.
27. Fan, H.T.; Hung, J.w.; Lu, X.; Wang, S.S.; Tsao, Y. Speech enhancement using segmental nonnegative matrix factorization. In Proceedings of the 2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Florence, Italy, 4–9 May 2014; pp. 4483–4487.
28. Cauchi, B. *Non-Negative Matrix Factorisation Applied to Auditory Scenes Classification*. Master’s Thesis, Master ATIAM, Université Pierre et Marie Curie: Paris, France, 2011.
29. Bisot, V.; Serizel, R.; Essid, S.; Richard, G. Acoustic scene classification with matrix factorization for unsupervised feature learning. In Proceedings of the 2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Shanghai, China, 20–25 March 2016; pp. 6445–6449.
30. Liu, D.C.; Nocedal, J. On the limited memory BFGS method for large scale optimization. *Math. Program.* **1989**, *45*, 503–528.
31. Lee, S.; Pang, H.S. Feature Extraction Based on the Non-Negative Matrix Factorization of Convolutional Neural Networks for Monitoring Domestic Activity With Acoustic Signals. *IEEE Access* **2020**, *8*, 122384–122395.
32. Choi, K.; Joo, D.; Kim, J. Kapre: On-gpu audio preprocessing layers for a quick implementation of deep neural network models with keras. *arXiv* **2017**, arXiv:1706.05781.
33. Cheuk, K.W.; Anderson, H.; Agres, K.; Herremans, D. nnAudio: An on-the-Fly GPU Audio to Spectrogram Conversion Toolbox Using 1D Convolutional Neural Networks. *IEEE Access* **2020**, *8*, 161981–162003.
34. Turpault, N.; Serizel, R.; Parag Shah, A.; Salamon, J. Sound event detection in domestic environments with weakly labeled data and soundscape synthesis. In Proceedings of the Detection and Classification of Acoustic Scenes and Events 2019 Workshop, New York, USA, 25–26 October 2019.
35. Lee, D.D.; Seung, H.S. Algorithms for non-negative matrix factorization. In *Advances in Neural Information Processing Systems*, Vancouver, Canada, 3–8 December 2001; pp. 556–562.
36. Cichocki, A.; Zdunek, R.; Phan, A.H.; Amari, S.i. *Nonnegative Matrix and Tensor Factorizations: Applications to Exploratory Multi-Way Data Analysis and Blind Source Separation*; John Wiley & Sons: Hoboken, NJ, USA, 2009.
37. Lee, S.; Lim, J.s. Reverberation suppression using non-negative matrix factorization to detect low-Doppler target with continuous wave active sonar. *EURASIP J. Adv. Signal Process.* **2019**, *2019*, 11.

38. Févotte, C.; Bertin, N.; Durrieu, J.L. Nonnegative matrix factorization with the Itakura-Saito divergence: With application to music analysis. *Neural Comput.* **2009**, *21*, 793–830.
39. Ben-Israel, A.; Greville, T.N. *Generalized Inverses: Theory and Applications*; Springer Science & Business Media: Berlin/Heidelberg, Germany, 2003; Volume 15.
40. Delphin-Poulat, L.; Plapous, C. *Mean Teacher with Data Augmentation for Dcase 2019 Task 4*; Technical Report; Orange Labs: Lannion, France, 2019.
41. Jiakai, L. *Mean Teacher Convolution System for Dcase 2018 Task 4*; Technical Report; DCASE2018 Challenge; DCASE community, 2018.
42. Tarvainen, A.; Valpola, H. Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. In *Advances in Neural Information Processing Systems*, Long Beach, California, USA, 4–9 December 2017; pp. 1195–1204.
43. Gemmeke, J.F.; Ellis, D.P.W.; Freedman, D.; Jansen, A.; Lawrence, W.; Moore, R.C.; Plakal, M.; Ritter, M. Audio Set: An ontology and human-labeled dataset for audio events. *Proc. IEEE ICASSP* **2017**, doi:10.1109/ICASSP.2017.7952261.
44. Fonseca, E.; Pons, J.; Favory, X.; Font, F.; Bogdanov, D.; Ferraro, A.; Oramas, S.; Porter, A.; Serra, X. Freesound Datasets: A platform for the creation of open audio datasets. In *Proceedings of the 18th International Society for Music Information Retrieval Conference (ISMIR 2017)*, Suzhou, China, 23–27 October 2017; pp. 486–493.
45. Dekkers, G.; Lauwereins, S.; Thoen, B.; Adhana, M.W.; Brouckxon, H.; van Waterschoot, T.; Vanrumste, B.; Verhelst, M.; Karsmakers, P. The SINS Database for Detection of Daily Activities in a Home Environment Using an Acoustic Sensor Network. In *Proceedings of the Detection and Classification of Acoustic Scenes and Events 2017 Workshop (DCASE2017)*, Munich, Germany, 16–17 November 2017; pp. 32–36.
46. Cornell, S.C.; Pepe, G.; Principi, E.; Pariente, M.; Olvera, M.; Gabrielli, L.; Squartini, S. *The UNIVPM-INRIA Systems for The DCASE 2020 Task 4*; Technical Report; DCASE2020 Challenge; DCASE community, 2020.

Article

# Learning Optimal Time Series Combination and Pre-Processing by Smart Joins

Amaia Gil <sup>1,\*</sup>, Marco Quartulli <sup>1</sup>, Igor G. Olaizola <sup>1</sup> and Basilio Sierra <sup>2</sup>

<sup>1</sup> Vicomtech Foundation, Basque Research and Technology Alliance (BRTA), Mikeletegi 57,

20009 Donostia-San Sebastián, Spain; mquartulli@vicomtech.org (M.Q.); iolaizola@vicomtech.org (I.G.O.)

<sup>2</sup> Department of Computer Sciences and Artificial Intelligence, University of the Basque Country (UPV/EHU),  
20018 Donostia-San Sebastián, Spain; b.sierra@ehu.es

\* Correspondence: agil@vicomtech.org

Received: 28 July 2020; Accepted: 8 September 2020; Published: 11 September 2020



**Abstract:** In industrial applications of data science and machine learning, most of the steps of a typical pipeline focus on optimizing measures of model fitness to the available data. Data preprocessing, instead, is often ad-hoc, and not based on the optimization of quantitative measures. This paper proposes the use of optimization in the preprocessing step, specifically studying a time series joining methodology, and introduces an error function to measure the adequateness of the joining. Experiments show how the method allows monitoring preprocessing errors for different time slices, indicating when a retraining of the preprocessing may be needed. Thus, this contribution helps quantifying the implications of data preprocessing on the result of data analysis and machine learning methods. The methodology is applied to two case studies: synthetic simulation data with controlled distortions, and a real scenario of an industrial process.

**Keywords:** optimization; machine learning; preprocessing

## 1. Introduction and Description of the Problem

In machine learning there are several steps to follow in order to perform model construction. Many of them, such as feature selection, feature extraction and model training, are based on mathematical optimization. However, the initial preprocessing is often not explicitly and quantitatively optimized.

In preprocessing, one of the main steps consists of obtaining all the features that will be used in model generation. The features can come from different origins and joining all the data adequately can be hard. The specific case of working with time series has the advantage of the use of a temporal reference system, a timeline that enables merging the observations. Nevertheless, each feature has its own sampling, and all of them should be resampled to construct a single multi-variate time series in a synchronized way.

This resampling is done by feature, and, depending on the application objectives, different characteristics of data joining methods should be taken into account. Examples of objective measures of preprocessing quality might be based on measures of distortion and on the information lost in the process. Further considerations might be related to the causal nature of the resulting system, or to the amount of delay (or anticipation, in case of being shifted to a prior time instance) applied to the different original series to synchronize them.

Note that these properties can have different degrees of practical importance depending on the application domain. On the one hand, in the case of real-time prediction, data anticipation can imply the need of waiting for a new data entrance, resulting a big delay in the prediction; obviously a data prediction approach based on time series analysis could be used to avoid this problem, using a correction method in case a significant difference between predicted and real values is detected. On the

other hand, information loss and data distortion can have a significant impact on the predictive power of the model.

### 1.1. Background

In the context of SQL database engines [1], a time series is a sequence of data values measured at successive, though not necessarily regular, points in time ([https://cloud.ibm.com/docs/sql-query?topic=sql-query-ts\\_intro](https://cloud.ibm.com/docs/sql-query?topic=sql-query-ts_intro)—IBM Cloud SQL Query documentation). Each entry in a time series is called an observation. Each observation comprises a timetick that indicates when the observation was made, and the data recorded for that observation. The set of timeticks defines a temporal base or temporal reference system for the series.

A temporal join is a join operation that operates on time series data. It produces a single array time series based on the original input data and the new temporal reference system.

This section introduces a range of common SQL joining methodologies. Specifically, the following methods will be introduced: left join, nearest join, forward join and backward join. Notably, the outer join is not contemplated, as the obtained results are the same as those from other selected methods (backward join or forward join) depending on the selection of a function for filling in Non-Available (NA) values (*ffill* or *bfill*).

In order to simplify the explanation of these methods, a specific example will be used, together with terminology from the documentation of the widely adopted pandas (<https://pandas.org>—Python data analysis and manipulation tool) data analysis library. Suppose that sensor data  $y(t_O)$  is acquired with the temporal reference system  $t_O$  as shown in Table 1a. For model learning, suppose the temporal reference system  $t_D$  shown in Table 1b is required.

**Table 1.** Problem definition. (a) Captured data. (b) Desired temporal reference system.

(a)	
$t_O$	$y$
10:00	$y_1$
10:02	$y_2$
10:16	$y_3$
10:27	$y_4$
(b)	
$t_D$	$\hat{y}$
10:00	
10:15	
10:30	

Finally, suppose the function *ffill* is selected for filling NA values, and that this function operates by forward-filling such NA values with the nearest prior known data value.

#### 1.1.1. Left Join

The left join method takes only samples from  $y$  that are synchronized with  $t_D$ , in other words, only data that originally had the desired time is used. Table 2a shows the application of a left join to the example. After filling NA values, the results shown in Table 2b are obtained.

In this particular example, three samples from  $y$  are not taken into account in the joined dataset. In this sense, part of the information in the original data is lost.

**Table 2.** Result of left join. (a) After applying the join. (b) After filling Non-Available (NA) values.

(a)	
$t_D$	$\hat{y}$
10:00	$y_1$
10:15	NA
10:30	NA
(b)	
$t_D$	$\hat{y}$
10:00	$y_1$
10:15	$y_1$
10:30	$y_1$

### 1.1.2. Nearest Join

A nearest join takes into account the nearest known available data from  $y$ . Results from the join are shown in Table 3.

**Table 3.** Result nearest join.

$t_D$	$\hat{y}$
10:00	$y_1$
10:15	$y_3$
10:30	$y_4$

Depending on the distribution of  $y$ , future knowledge of future data can be added to the past in a non-causal manner. In the example, the joined series at 10:15 uses data from 10:16.

### 1.1.3. Forward Join

In a forward join, samples of  $t_D$  that are not available in  $t_O$  are selected using subsequent matches from  $y$ . Results from the join are shown in Table 4a, and after filling NA values in Table 4b.

**Table 4.** Result of forward join. (a) After applying the join. (b) After filling NA values.

(a)	
$t_D$	$\hat{y}$
10:00	$y_1$
10:15	$y_3$
10:30	NA
(b)	
$t_D$	$\hat{y}$
10:00	$y_1$
10:15	$y_3$
10:30	$y_3$

### 1.1.4. Backward Join

In a backward join, samples of  $t_D$  that are not available in  $t_O$  are selected using the nearest prior match. Results from the join are shown in Table 5.

Given the above existing methods, the remainder of this paper considers the problem of locally selecting an optimal method by the optimization of a quantitative measure of the quality of the obtained joined series.



**Table 5.** Result of backward join.

$t_D$	$\hat{y}$
10:00	$y_1$
10:15	$y_2$
10:30	$y_4$

## 1.2. Paper Contributions and Structure

We consider that the need to define an operational mechanism to align multiple time series with a different time base by optimizing of a cost function that can be defined by the user is not adequately addressed in the present literature. In this sense, the contributions put forward by this paper include:

- The idea that the preprocessing steps in a machine learning workflow can be subject to an optimization procedure that is similar to the one used with e.g., an empirical risk estimate in the actual model learning step.
- The idea that a join operation among tables representing time series with different time bases as operated by e.g., a SQL database engines can be learned based on previous data records.
- A specific algorithm and implementation for a method meant to align multiple time series with different time bases.

The rest of the paper is structured as follows. Section 2 introduces the state of the art approaches. The methodology and a proposed solution are explained in Section 3. Section 4 provides a description of the case studies, whereas Section 5 shows the results of those case studies. Finally, conclusions and future work are presented in Section 6.

## 2. State of the Art

A number of contributions have been put forward in the literature that deal with the need to align of time series. On the one hand, such a need could stem from the fact that the time series described related phenomena with “warped” temporal aspects (as in Dynamic Time Warping). On the other hand, such a need could depend on the fact that the time series suffer from the effects of different decimation processes (as in the literature related to Dynamic Processes).

In the first group, Folgado et al. [2] considered an extension of Dynamic Time Warping based on a distance which characterized the degree of time warping between two sequences meant for applications where the timing factor is essential, and proposed a Time Alignment Measurement, which delivered similarity information on the temporal domain.

Morel et al. [3] extended Dynamic Time Warping to sets of signals. A significant point with respect to the topic of the present paper is the definition of a tolerance that takes into account the admissible variability around the average signal.

One of the nearest related topics is trying to solve, at the same time, several goals, or to deal with several constraints in parallel. In this sense, there are some works which tackle scheduling problems; a review of this type of models in a practical problem related to flow shop scheduling is presented by Sun et al. [4]. The authors stated that that heuristic and meta-heuristic methods and hybrid procedures were proven much more useful than other methods in large and complex situations.

Tawhid and Savsani [5] proposed a novel multi-objective optimization algorithm named multi-objective sine–cosine algorithm (MO-SCA) which was based on the search technique of the sine–cosine algorithm. They ended obtaining different non-discriminatory levels to preserve the diversity among the set of solutions.

Task scheduling is another problem related to this paper requiring multi-objective optimization paradigms. Zuo et al. [6] presented a solution based on an Ant Colony approach to deal with Cloud Computing computational load and storage minimization. In the same direction, Zahedi et al. [7] presented an approach related to vehicle routing for goods distributions in emergency situations.

The data from a 2017 big earthquake in India was used, considering the demands heterogeneity and dynamics, distribution planning of goods and routing of vehicles simultaneously by means of a genetic algorithm.

Finally, regarding forecasting, Yang et al. [8] presented a system based on a dual decomposition strategy and multi-objective optimization for electricity price forecasting with the goal of balancing electricity generation and consumption. Data pre-processing was fundamental in the selected time window.

### 3. Proposed Solution: Smart Join

In this paper, a smart join method based on an optimization process is proposed. The aim of this optimization problem is to select the method that minimizes the errors of the resampling process for each feature.

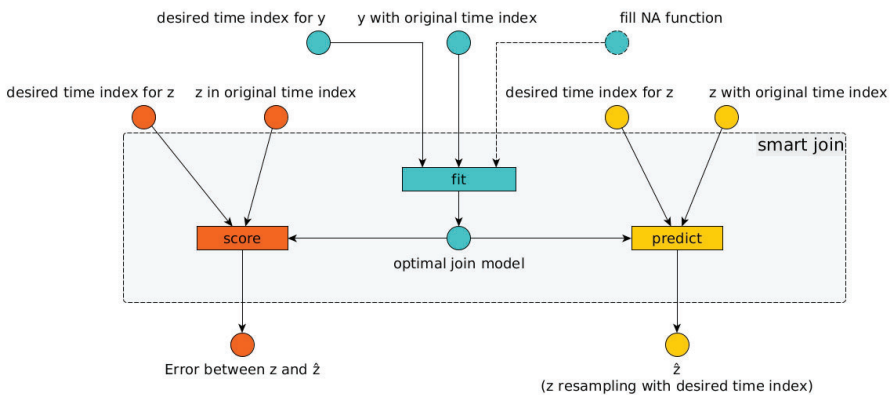
First, a detailed explanation is presented in Section 3.1 and an example of application is shown afterwards in Section 3.2.

#### 3.1. Description of The Methodology

The general concept of the methodology of the smart join is explained next:

1. First, the joining model is fitted using training data; in other words, the optimal joining solution of the process is obtained. This needs to be done for each feature separately.
2. Then, resampled data is predicted by applying the selected join method to the test data.
3. Finally, the model is validated using resampling error.

Suppose we have a time series slice  $y$  of the selected feature that needs to be resampled to be joined with a desired time index. First of all, the fit method is used in order to obtain the “optimal” join method. The inputs needed for the join are the original time series slice ( $y$  with the original time index) and the desired time index. Other optional parameter can be a fill NA function as it can affect selecting the “optimal” method. Then, another slice of the same feature ( $z$ ) is used for the testing by the use of the method score. Finally, the optimal joined method is used for resampling other time slices of the features with the predict method. The structure of the different methods can be depicted as in Figure 1.



**Figure 1.** Smart join methodology implementation structure. Firstly fit method is used for the selection of the optimal join method and then, predict and score methods are used to resample other slices of the time series and in order to control the error produced by the join.

The fitting process to find an optimal joining model could be mathematically represented as follows:

Suppose we have the time series  $y(t_O)$  where  $t_O = [t_{O1}, t_{O2}, \dots, t_{Om}]$  is the initial temporal reference system. Let  $j$  be a join method from the available methods set  $J$  (*left, backward, forward, nearest*). We need to obtain a new time series  $\hat{y}(t_D; j, y)$  with the desired temporal reference system  $t_D = [t_{D1}, t_{D2}, \dots, t_{Dn}]$ . The smart join algorithm aims to find the optimal join method  $j \in J = \{\textit{left, backward, forward, nearest}\}$  that minimizes an error function  $E(y, \hat{y})$ . The parameters for applying the smart join method are the function meant to fill unavailable measurement values  $f \in F = \{\textit{None, bfill, ffill, nearest}\}$ . In case of not being specified, default values will be used (in which case  $f = \textit{None}$ ). The possible values of the imputation function  $f$  are *None* (not filling NA values), *bfill* (using subsequent value that is nearest) and *ffill* (using prior value that is nearest). The optimization problem is defined as:

$$\arg \min_{j \in J} E(y, \hat{y}) \tag{1}$$

With respect to the second contribution put forward by the present paper, the error function  $E(y, \hat{y})$  proposed is defined by Equation (2).

$$\begin{aligned} E(y, \hat{y}) = & w_1 \cdot NaEl(\hat{y}) + w_2 \cdot MissEl(y, \hat{y}) + w_3 \cdot DelEl(y, \hat{y}) \\ & + w_4 \cdot DelT(y, \hat{y}) + w_5 \cdot AntEl(y, \hat{y}) + w_6 \cdot AntT(y, \hat{y}) \\ & + w_7 \cdot Diff(y, \hat{y}), \end{aligned} \tag{2}$$

where  $w_i > 0$  with  $i \in \{1, 2, \dots, 7\}$  and  $\sum_{i=1}^7 w_i = 1$  are the weights for the total error calculation and, in case of not being specified, their default value is  $w_i = 1/7 \forall i$ .

In the following paragraphs, each function that takes part in the error  $E(y, \hat{y})$  is presented. Suppose  $k \in \{1, 2, \dots, n\}$  and  $l \in \{1, 2, \dots, m\}$  indicate the index of elements in  $\hat{y}$  and  $y$  respectively.

$NaEl(\hat{y})$  represents the percentage of NA elements of  $\hat{y}$  after the application of  $f$ . NA values can be problematic in machine learning applications implying for example the need to remove data points with NA value on the model training process or the impossibility to predict an output value using the trained model.

$$NaEl(\hat{y}) = \frac{\sum_{k=1}^n s_k}{n}, \text{ where } s_k = \begin{cases} 1 & \text{if } \hat{y}_k \text{ is NA} \\ 0 & \text{otherwise} \end{cases} \tag{3}$$

$MissEl(y, \hat{y})$  is the percentage of elements from  $y$  that are not used in  $\hat{y}$ . This value is related to the lost of information from the original time series due to the resampling needed.

$$MissEl(y, \hat{y}) = \frac{\sum_{l=1}^m x_l}{m}, \text{ where } x_l = \begin{cases} 1 & \text{if } y_l \notin \hat{y} \\ 0 & \text{otherwise} \end{cases} \tag{4}$$

$DelEl(y, \hat{y})$  indicates the percentage of delayed elements. If most of the data points from  $y$  are delayed, the reality for the machine learning model is displaced. Depending on the application environment, taking decisions supported by the machine learning system that could not adequately represent the current situation can be problematic.

$$DelEl(y, \hat{y}) = \frac{\sum_{k=1}^n d_k}{n}, \text{ where } d_k = \begin{cases} 1 & \text{if } (\hat{y}_k = y_l) \text{ and } (t_{Dk} > t_{Ol}) \forall l \\ 0 & \text{otherwise} \end{cases} \tag{5}$$

$DelT(y, \hat{y})$  is the maximum difference in time between a delayed element used in  $\hat{y}$  and its original time position normalized by the time window of  $y$ . Whereas the previous case considers the frequency of delayed elements,  $DelT(y, \hat{y})$  takes into account the magnitude of the displacement.

$$DelT(y, \hat{y}) = \frac{\max(e_k)}{t_{Om} - t_{O1}}, \text{ where } e_k = \begin{cases} t_{Dk} - t_{Ol} & \text{if } (\hat{y}_k = y_l) \text{ and } (t_{Dk} > t_{Ol}) \forall l \\ 0 & \text{otherwise} \end{cases} \tag{6}$$

$AntEI(y, \hat{y})$  and  $AntT(y, \hat{y})$  are equivalent functions but in this case for anticipated elements.

$$AntEI(y, \hat{y}) = \frac{\sum_{k=1}^n a_k}{n}, \text{ where } a_k = \begin{cases} 1 & \text{if } (\hat{y}_k = y_k) \text{ and } (t_{Dk} < t_{O1}) \forall l \\ 0 & \text{otherwise} \end{cases} \quad (7)$$

and

$$AntT(y, \hat{y}) = \frac{\max(b_k)}{t_{Om} - t_{O1}}, \text{ where } b_k = \begin{cases} t_{O1} - t_{Dk} & \text{if } (\hat{y}_k = y_k) \text{ and } (t_{Dk} < t_{O1}) \forall l \\ 0 & \text{otherwise} \end{cases} \quad (8)$$

On the one side, the use of anticipated data is equivalent to the use of future information for prediction and results can be misleading and the used approximation should be sound enough to deal with value forecasting. On the other side, using future data could imply a need to wait for the arrival of a new observation to be able to make a prediction, or a correction would be needed once the predicted value and the real one are compared.

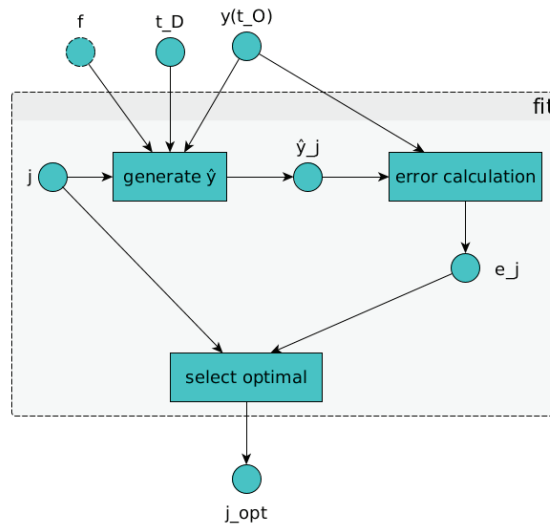
Finally  $Diff(y, \hat{y})$  calculates the difference between the two time series (original and resampled). This value could represent the magnitude of the distortion committed due to the need of a joined data with synchronized temporal reference system.

$$Diff(y, \hat{y}) = \frac{\text{mean}(\text{abs}(y_{inter} - \hat{y}_{inter}))}{\max(y) - \min(y)}, \quad (9)$$

where  $y_{inter}$  and  $\hat{y}_{inter}$  are obtained by means of linear interpolation of time series  $y$  and  $\hat{y}$  respectively for time values in  $t_O \cup t_D$ .

Each part of the sum of the error calculation Equations (3)–(9) is normalized to guarantee that the result is in range [0, 1] so different errors are comparable between them.

The fitting method can be seen graphically in Figure 2.



**Figure 2.** Fitting method diagram. First,  $\hat{y}_j$  resampled time series are generated from each joining method ( $j \in J$ ). Using the generated resampled time series, error is calculated in each case and the optimal solution is selected  $j_{opt}$ .

Validating the joined method in different time slices of the time series is crucial. If the slice of data used to train the joining model is adequately selected, the errors should be similar in different time

windows. Depending on the stability of the feature, retraining may be required as the optimal join method could not be the most adequate during all time period. Furthermore, selecting the desired temporal reference system ( $t_D$ ) has equal importance as it should be the same for all the features, in order to be able to construct a database with all the features used by the model. Although the error calculation and the optimal joining methodology is chosen separately per feature, the desired temporal reference system is a common input of all the optimization problems and its selection affects to all the features.

### 3.2. Application Example

The current subsection introduces an illustrative example of the application of the proposed method to a dataset from a simple piecewise function. Suppose that the piecewise function is sampled irregularly in order to save memory applying two criteria:

- The system checks every minute if the value of the data point has changed enough according to a pre-established criterion (in this particular case, a difference with the prior data point higher than 0.5) to save that data point.
- Every minute the system also checks the difference in time with the last saved data point and if this difference is greater than or equal to four minutes it saves the last available data point.

The original piecewise function and the saved data using these criteria are shown in Figure 3.

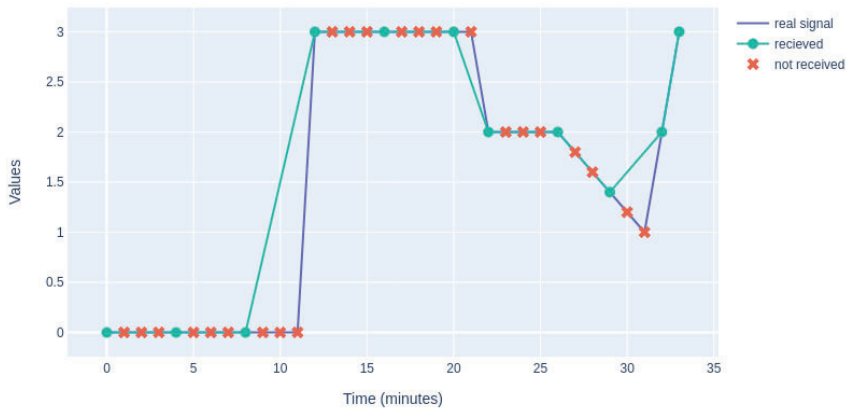
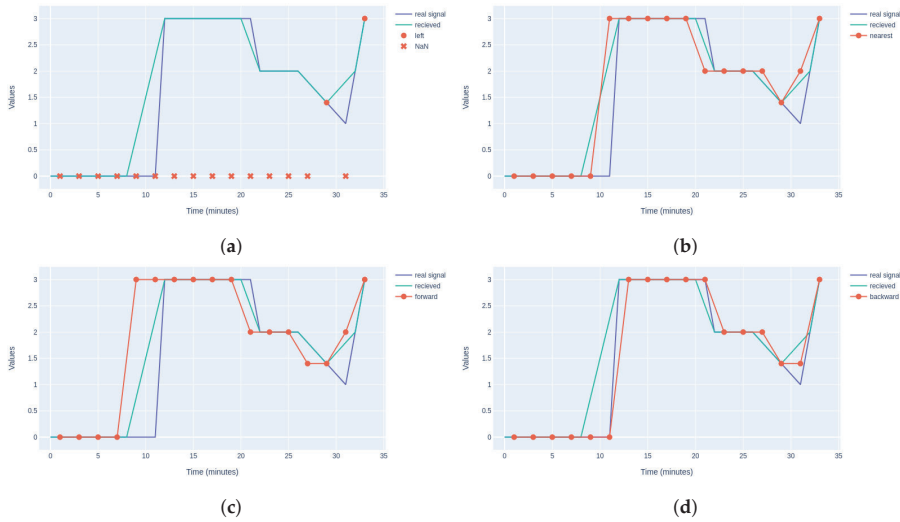


Figure 3. Application example problem.

Suppose that the desired time reference system corresponds to  $t_d = \{1, 3, 5, \dots, 33\}$ . Results after the application of different joining methods are shown in Figure 4. Error values used in the optimization of the Smart Join methodology are shown in Table 6.

Because the input for the algorithm is the received data, when default weights in the error function are used ( $w_i = 1/7 \forall i$ ), the minimal error is obtained by the nearest join (see Figure 4b). However, if knowledge about the irregular sampling approach used by the system is introduced by penalizing the anticipation of data points (for example with  $w_5 = w_6 = 2/9$  and  $w_1 = w_2 = w_3 = w_4 = w_7 = 1/9$ ), the optimal join method is backward join. Figure 4d shows that the data points obtained by the backward join as a result of taking into account this extended description of the data sampling mechanism are the ones that are the closest to the real piecewise function.



**Figure 4.** Application example results for different joining methods. (a) Left join. (b) Nearest join. (c) Forward join. (d) Backward join.

**Table 6.** Error values for different join methods in the application example.

Method	$NaEI(\hat{y})$	$MissEI(y, \hat{y})$	$DelEI(y, \hat{y})$	$DelT(y, \hat{y})$	$AntEI(y, \hat{y})$	$AntT(y, \hat{y})$	$Diff(y, \hat{y})$
left	0.882	0.818	0.0	0.0	0.0	0.0	1.0
nearest	0.0	0.0	0.471	0.031	0.412	0.031	0.045
forward	0.0	0.091	0.0	0.0	0.882	0.094	0.092
backward	0.0	0.091	0.882	0.094	0.0	0.0	0.084

Having established the significance of the measure of quality of a joining method, in the remainder of this contribution we leverage mathematical optimization techniques on training data to automatically determine which of the joining methods is most adequate for a given time series.

#### 4. Experimental Setup

Two experiments were used in order to show the usefulness of the proposed smart join methodology.

The first one is a controlled application from simulated data and working with a unique time series to resample. Different distortion methods were applied to the data in order to have a practical use case with known theoretical result.

The second case is an application from an industrial chemical process. The aim of showing this example is to demonstrate the performance of the smart join method in a real scenario and the importance of adequately selecting the joining method and its implications.

##### 4.1. Experiments on Synthetic Data

The experiments on synthetic data are carried out on the  $x, y, z$  3D curve generated in time  $t$  by a Lorenz system (originally a simplified model for atmospheric convection) [9].

$$\begin{aligned}
 \frac{dx}{dt} &= \sigma(y - x) \\
 \frac{dy}{dt} &= x(\rho - z) - y \\
 \frac{dz}{dt} &= xy - \beta z
 \end{aligned}
 \tag{10}$$

with parameters  $\sigma = 10, \rho = 28$  and  $\beta = 8/3$  and initial conditions  $x(0) = y(0) = z(0) = 1$  and  $t \in [0, 40]$ . The time sampling interval selected for the time series was 0.1 time units.

The simulated data can be observed in Figure 5a. To apply the smart join methodology only dimension  $x$  was used. The time series is shown in Figure 5b.

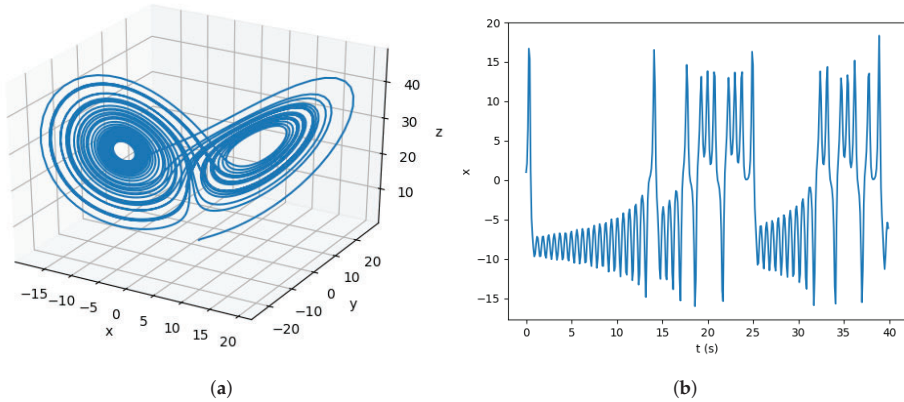


Figure 5. Lorenz system result data. (a) Three-dimensional data. (b) First dimension time series.

In order to generate a distorted version of this time series in a controlled manner, some distortion methods were applied, inspired by from the work of Kreindler and Lumsden [10], which will be described later in the section.

This controlled experiment setup was used to demonstrate how errors change depending on the join method and on the type of distortion that is applied to each time window. The distortions have been selected in order to represent usual problems such as missing data or delays in receiving data points.

The series was divided into four parts of equal size. In the first part ( $t \in (0, 10]$ ) the time series remains unaltered. In the range  $t \in (10, 20]$ , 20% randomly selected data points were removed. This distortion can be seen in Figure 6a. In the remaining part of the time series, 20% of data were shifted forward (in  $t \in (20, 30]$ ) or backward (in  $t \in (30, 40]$ ). The shifted quantity was selected by a random uniform variable, guaranteeing that data points cannot be disordered. In other words, the maximum possible shifted quantity was set by the sampling frequency value of the original simulation data. The distortion effect generated in the time series can be observed in Figure 6b.

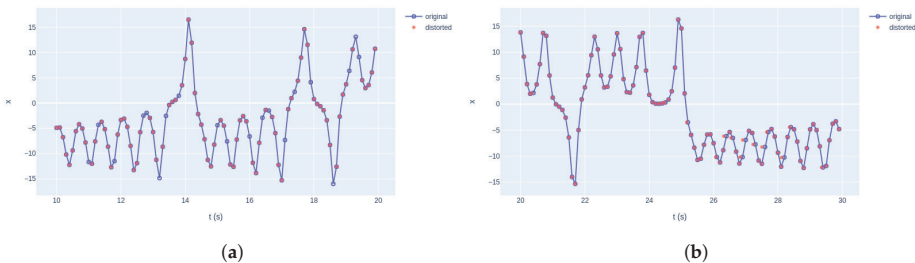


Figure 6. Zoomed distortions of Lorenz first dimension. (a) Removed data. (b) Shifted data.

The difference between modified and original data can be seen in Figure 7.



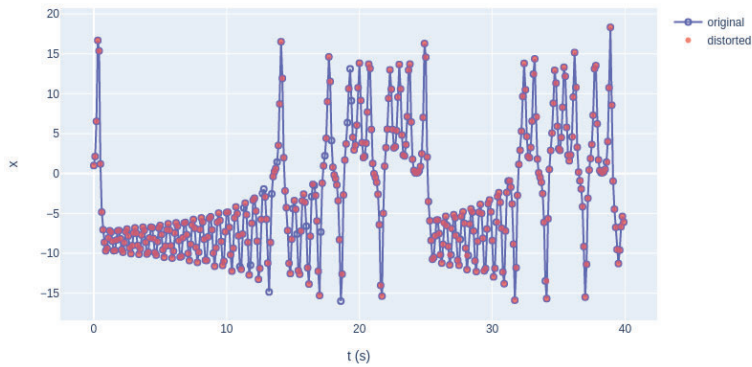


Figure 7. Original vs. distorted Lorenz first dimension.

4.2. Experiments on Real Industrial Dataset

The efficient management and the energy optimization of distillation units [11,12] in terms of both product quality and energy efficiency in both the petro-chemical and in the sustainable sector pose a great challenge to process and control engineers because of their complexity. The management, optimization and fault analysis of such units all require accurate process models, which in recent years have started to be generated directly from the data available in SCADA Historian databases by using machine learning methods [13–15] whose performance depends on the availability of properly pre-processed multi-variate data.

Suppose that the system captures and stores real-time sensor-based data. In this particular case, each sensor writes values in the database only when there is a significant change in the values of data. The decision on the significance of the difference between data points is based on the scale of each feature. The aim of this data recording strategy is reducing data volume. Consequently, if a feature becomes unstable the writing frequency augments drastically.

For machine learning applications, an alignment between features is needed. Each feature should be resampled to obtain a common desired temporal reference system previous to any feature extraction/selection algorithm application. Depending on the feature and the application system, the optimal joining method can be different.

Figure 8 shows the initial sampling for different features. Each column represents a feature and each row an hour time window. The number of samples is counted per hour and feature, and represented by the colour.

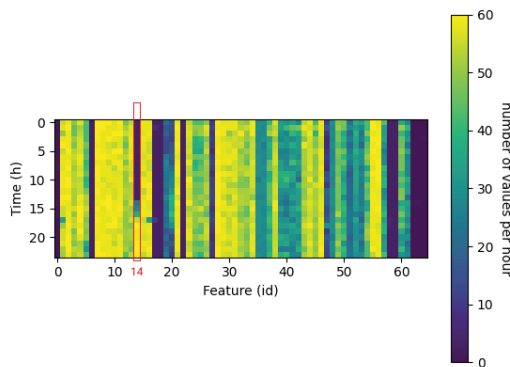


Figure 8. Original sampling of real industrial data from a distillation unit.

Figure 8 shows how, depending on the feature the frequency of data availability can be constant or variable, and the quantity of samples can be very different among features. For example, the feature with id 14 changes drastically from very low frequency to high frequency only in a couple of hours. The data points for this particular feature are shown in Figure 9, where the frequency change is observable.

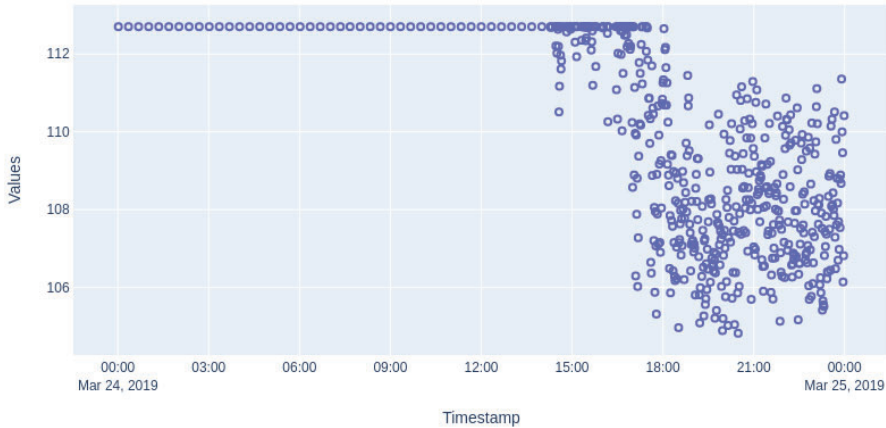


Figure 9. Original data of the feature with id 14 from Figure 8.

In this particular case, the desired time sampling interval is selected to be 15 min.

## 5. Experimental Results

This section presents experimental results for the aforementioned case studies.

### 5.1. Results on Synthetic Data

For synthetic data, different time series joining methods were used separately and the error, defined by Equation (2), was calculated for each method using windows of  $t \in (p, p + 2]$  with  $p \in \{0, 2, \dots, 38\}$ . The selected values for the parameters of smart join methodology were  $w_i = 1/7 \forall i$  (i.e., the same importance for all different functions taking part in the error calculation) and the imputation function was  $f = \text{ffill}$ .

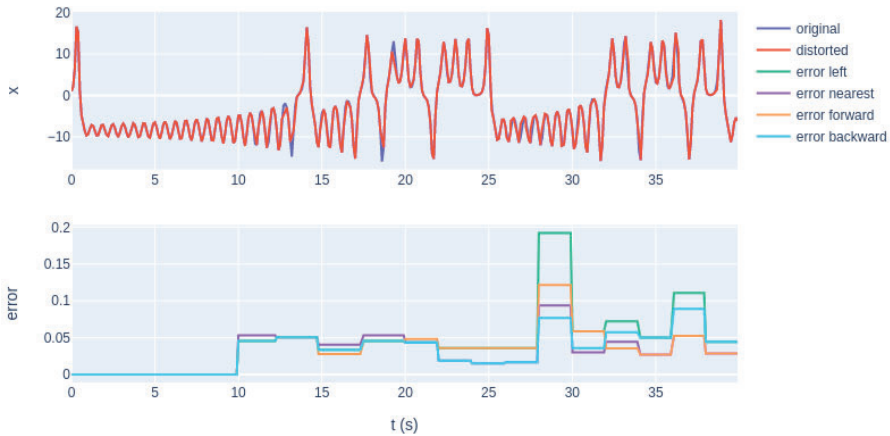
Table 7 shows the error values per method and time window. The optimal solution (minimum error) is marked in bold. An additional column labelled “theoretical” represents the theoretically optimal solution. Thus, the obtained optimal solution in each time window can be compared and contrasted with the theoretical solution. In Figure 10 numerical results are shown graphically.

As per Table 7, the optimal joining method (the one that has minimal error in each window) depends on the controlled distortion introduced. The proposed methodology is capable of obtaining as one of the optimal available results the theoretical solution. On the one hand, for  $t \in (0, 10]$ , the data was already available for the needed temporal reference system and for that reason all the methods were able to obtain a 0.0 value error. On the other hand, for  $t \in (10, 20]$ , as data points are removed randomly, there was no optimal theoretical solution, as from known data points the joining method should not be able to reconstruct the time series. For this range, the optimal solution for the joining method depends on  $\text{Diff}(y, \hat{y})$ , i.e., the distortion introduced is comparable to the one obtained with the lineal interpolation result. For  $t \in (20, 30]$  and  $t \in (30, 40]$  the optimal theoretical solutions were backward and forward join, respectively. However, in multiple windows, the nearest join method obtained the same solution as the theoretically optimal method, as the shifted data points

( $t_{O_i}$  introduced in the smart join system) are the nearest ones to the desired data points ( $t_{D_k}$  output temporal reference system).

**Table 7.** Results on synthetic data, in bold the method with minimal error (multiple solutions are possible).

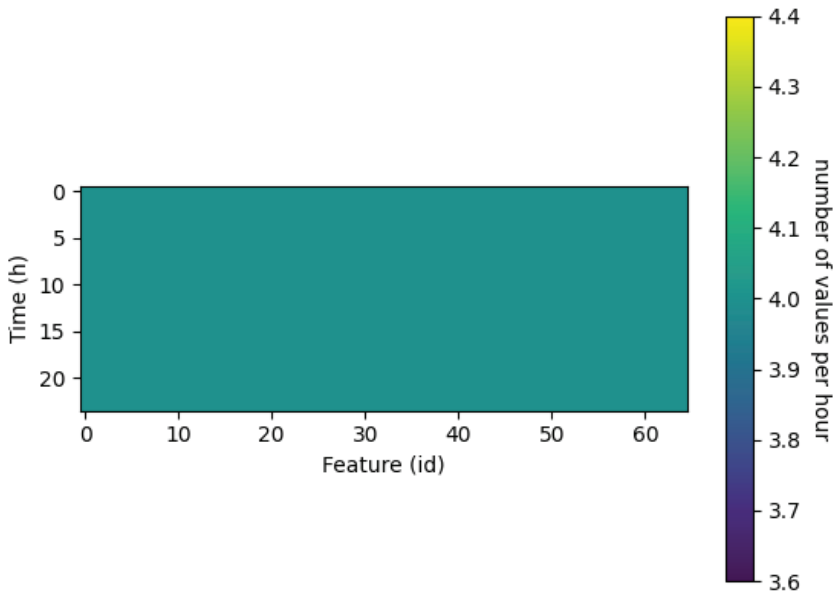
$t$ Range	Backward Join	Forward Join	Left Join	Nearest Join	Theoretical
0–2	$0.00 \times 10^0$	$0.00 \times 10^0$	$0.00 \times 10^0$	$0.00 \times 10^0$	all
2–4	$0.00 \times 10^0$	$0.00 \times 10^0$	$0.00 \times 10^0$	$0.00 \times 10^0$	all
4–6	$0.00 \times 10^0$	$0.00 \times 10^0$	$0.00 \times 10^0$	$0.00 \times 10^0$	all
6–8	$0.00 \times 10^0$	$0.00 \times 10^0$	$0.00 \times 10^0$	$0.00 \times 10^0$	all
8–10	$0.00 \times 10^0$	$0.00 \times 10^0$	$0.00 \times 10^0$	$0.00 \times 10^0$	all
10–12	$4.56 \times 10^{-2}$	$4.56 \times 10^{-2}$	$4.56 \times 10^{-2}$	$5.32 \times 10^{-2}$	none
12–14	$5.08 \times 10^{-2}$	$5.08 \times 10^{-2}$	$5.08 \times 10^{-2}$	$5.08 \times 10^{-2}$	none
14–16	$3.35 \times 10^{-2}$	$2.77 \times 10^{-2}$	$3.35 \times 10^{-2}$	$4.06 \times 10^{-2}$	none
16–18	$4.55 \times 10^{-2}$	$4.55 \times 10^{-2}$	$4.55 \times 10^{-2}$	$5.35 \times 10^{-2}$	none
18–20	$4.36 \times 10^{-2}$	$4.83 \times 10^{-2}$	$4.36 \times 10^{-2}$	$4.36 \times 10^{-2}$	none
20–22	$1.89 \times 10^{-2}$	$3.61 \times 10^{-2}$	$3.61 \times 10^{-2}$	$1.89 \times 10^{-2}$	backward
22–24	$1.50 \times 10^{-2}$	$3.61 \times 10^{-2}$	$3.61 \times 10^{-2}$	$1.50 \times 10^{-2}$	backward
24–26	$1.67 \times 10^{-2}$	$3.61 \times 10^{-2}$	$3.61 \times 10^{-2}$	$1.67 \times 10^{-2}$	backward
26–28	$7.69 \times 10^{-2}$	$1.21 \times 10^{-1}$	$1.92 \times 10^{-1}$	$9.39 \times 10^{-2}$	backward
28–30	$3.61 \times 10^{-2}$	$5.86 \times 10^{-2}$	$5.86 \times 10^{-2}$	$3.00 \times 10^{-2}$	backward
30–32	$5.75 \times 10^{-2}$	$3.55 \times 10^{-2}$	$7.22 \times 10^{-2}$	$4.44 \times 10^{-2}$	forward
32–34	$5.04 \times 10^{-2}$	$2.72 \times 10^{-2}$	$5.04 \times 10^{-2}$	$2.72 \times 10^{-2}$	forward
34–36	$8.90 \times 10^{-2}$	$5.26 \times 10^{-2}$	$1.11 \times 10^{-1}$	$5.26 \times 10^{-2}$	forward
36–38	$4.44 \times 10^{-2}$	$2.86 \times 10^{-2}$	$4.44 \times 10^{-2}$	$2.86 \times 10^{-2}$	forward
38–40	$3.61 \times 10^{-2}$	$1.99 \times 10^{-2}$	$3.61 \times 10^{-2}$	$1.99 \times 10^{-2}$	forward



**Figure 10.** Synthetic data results.

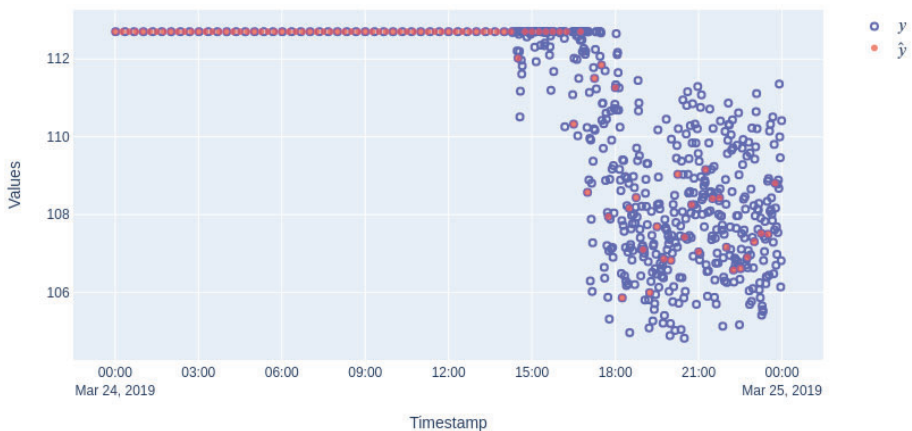
5.2. Results on Real Industrial Datasets

With respect to the real dataset, all features had a common sampling distribution after the joining as per Figure 11. In this case, the common sampling distribution was represented by having the same colour by row for all the features (represented by columns). Furthermore, as the selected temporal reference system ( $t_D$ ) had a constant sampling interval, the figure results in constant colour (four data points for each feature each hour).



**Figure 11.** Result after the use of smart join. After the joining, all features have a common sampling distribution.

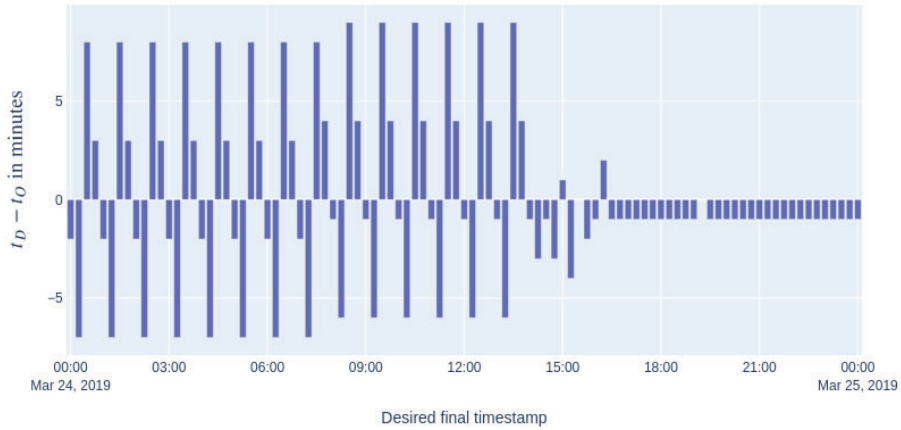
In Figure 12 the original time series ( $y$ ) and the one obtained from the joining methodology ( $\hat{y}$ ) are shown for making a visual comparison. Both time series ( $y$  and  $\hat{y}$ ) had similar appearance until 16:00 where the feature became unstable. Due to the selected time sampling and the joins considered for finding the optimum being the ones operated by SQL database engines, only a data point near the needed sampling was selected.



**Figure 12.** Comparison between original time series and after the use of smart join for the feature with id 14.

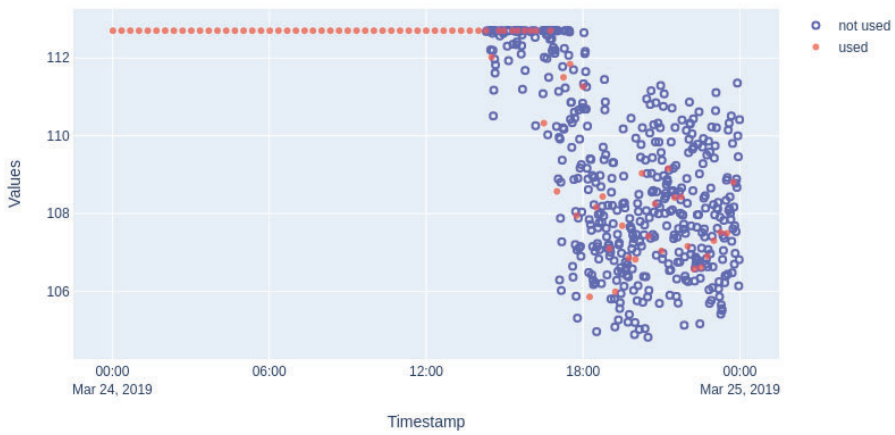
Figure 13 shows the alignment distortion for the feature with id 14. Negative values in this misalignment imply that anticipated time data were used in the join, whereas positive values imply

delayed time. The difference in the alignment could imply delays in prediction if anticipated data were used in the join or did not really have updated information of the process in order to make an adequate decision. In this particular case as the original time sampling initially writes nearly each 20 min and the desired time sampling is every 15 min, delays or anticipations of nearly 8 min become common. In the last part of the original time series, as data were available every minute or two, the delays or anticipations are drastically reduced for the joined time series.



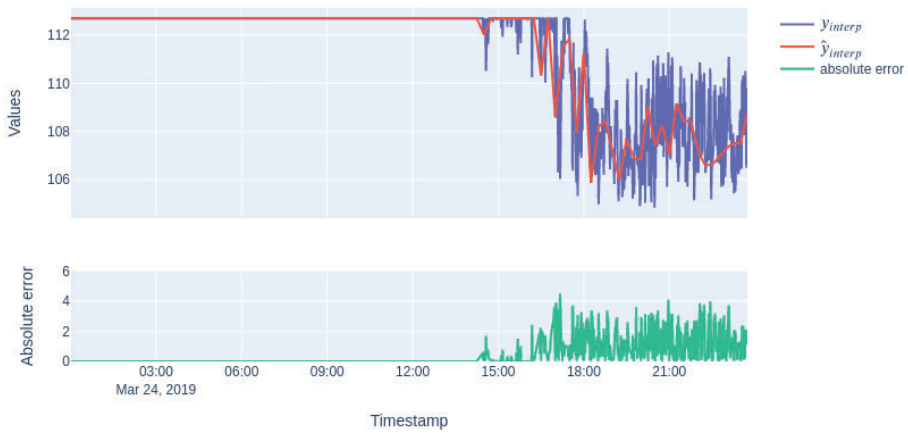
**Figure 13.** Alignment distortion for feature 14 between  $y$  and  $\hat{y}$ . Negative values in this misalignment imply that anticipated time data has been used in the join, whereas positive values imply delayed time. The difference in the alignment could imply delays in prediction if anticipated data was used in the join or did not really have updated information of the process in order to make an adequate decision.

Figure 14 shows used and unused points from the original time series in the join time series. Depending on the application the lost information could have a great impact. For time later that 16:00, as the selected time sampling ( $t_D$ ) is slower than the dynamic of the original time series, a lot of data points are unused in the joining process, losing the information provided by those data points showed in blue in the figure. In some cases, different aggregation methods or rolling windows could be more adequate to use the data that otherwise will be lost.

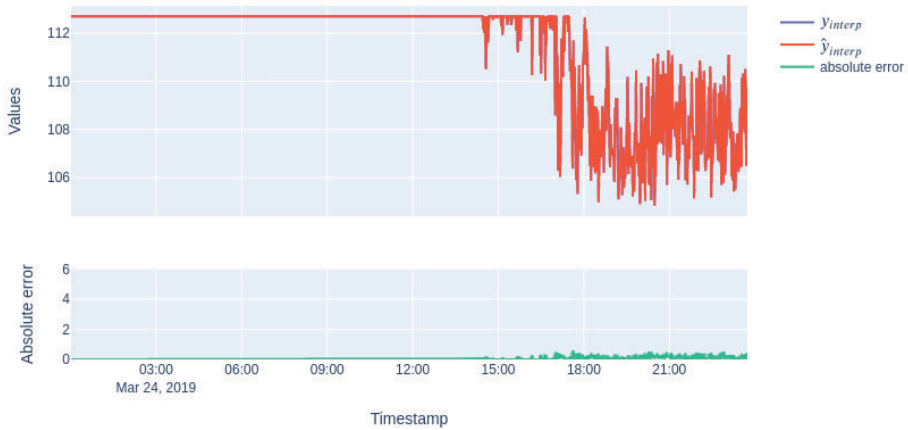


**Figure 14.** Data used and not used from  $y$  to generate  $\hat{y}$ .

The difference between the original time series and the joined one can help optimize the time sampling for a specific application. At the top of Figure 15 both the time series used for error calculation in part of  $Diff(y, \hat{y})$  ( $\hat{y}_{interp}$  and  $y_{interp}$ , i.e., generated by linear interpolation of time series  $y$  and  $\hat{y}$  in order to have common data sampling distribution ( $t_O \cup t_D$ )) are shown, while the lower diagram shows the absolute error value calculated at each point. For a comparison of how the frequency selection can affect the desired time sampling, a similar diagram with a desired sampling frequency modified from 15 min to one minute is shown in Figure 16. In both figures, as initially the original time series has constant values, there is no difference between both interpolated time series. However, as time passes by and the time series becomes unstable, the difference is remarkable. This error is greater in Figure 15, as the desired time sampling frequency is slower than the real dynamic of the feature and data is not linear.



**Figure 15.** Difference between original data and joined data with a desired time sampling frequency 15 min.



**Figure 16.** Difference between original data and joined data with desired time sampling frequency 1 min.

In Table 8 the effect in the error of different selections of desired sampling frequency are shown for comparison.

**Table 8.** Error values for the nearest joining method for different requested sampling frequencies.

Frequency	$NaEl(\hat{y})$	$MissEl(y, \hat{y})$	$DelEl(y, \hat{y})$	$DelT(y, \hat{y})$	$AntEl(y, \hat{y})$	$AntT(y, \hat{y})$	$Diff(y, \hat{y})$
1	0.0	0.494	0.333	0.007	0.667	0.006	0.004
5	0.0	0.851	0.322	0.007	0.678	0.005	0.036
10	0.0	0.903	0.315	0.007	0.685	0.005	0.054
15	0.0	0.921	0.333	0.007	0.667	0.004	0.058

## 6. Conclusions and Future Work

Standard data analysis pipelines often include resampling, interpolation and aggregation steps that are not optimized in the model learning procedure.

This paper introduced the definition of an optimization problem for data preprocessing, and in particular for data joining processes that imply a need for data resampling. The defined problem has been addressed by a method designed to efficiently solve it. The case studies introduced have demonstrated the applicability of the proposed method to time series data, using standard SQL-like data joining primitives as a basis to be optimized upon. The first case study, with simulated data and controlled distortions, means to provide insight into the methodology and its applicability. In the second experiment, the proposed methodology is applied in a real scenario, showing the impact of the decisions taken in the preprocessing step on the learning of data-based models.

Furthermore, the paper proposed an error function for its use in the optimization problem of joining time series. This error function allows comparisons across different features and time slices, which is needed to select among different join methods or to monitor their quality on different time series slices. As errors are comparable, selecting the optimal solution or knowing when there is a need for retraining is possible. Moreover, using the input parameters ( $w$  and  $f$ ) of the proposed error function allows adapting the function to an adequate solution for different applications.

The approach presented in this paper has several new paths to follow as future works: on the one hand, the approach could be improved, adding automatic selection of the time window size, or applying B-Spline mode approximations of the missing values; on the other hand, the benefits of the proposed Smart Join method should be quantified on a diverse range of real world applications. Energy consumption, storage and production, supply transportation and storage management are candidates towards this end.

**Author Contributions:** A.G. and M.Q. designed and implemented the experimental testbed and algorithm, B.S. and I.G.O. supervised the experimental design and managed the project. M.Q. and B.S. reviewed the new approach of this research. A.G. performed the experimental phase. All authors contributed to the writing and reviewing of the present manuscript. All authors read and agreed to the published version of the manuscript.

**Funding:** This research has been partially funded by the 3KIA project (ELKARTEK, Basque Government).

**Conflicts of Interest:** The authors declare no conflict of interest.

## References

- Codd, E.F. *The Relational Model for Database Management*; Addison-Wesley Publishing Company: Boston, MA, USA, 1990.
- Folgado, D.; Barandas, M.; Matias, R.; Martins, R.; Carvalho, M.; Gamboa, H. Time alignment measurement for time series. *Pattern Recognit.* **2018**, *81*, 268–279.
- Morel, M.; Achard, C.; Kulpa, R.; Dubuisson, S. Time-series averaging using constrained dynamic time warping with tolerance. *Pattern Recognit.* **2018**, *74*, 77–89.
- Sun, Y.; Zhang, C.; Gao, L.; Wang, X. Multi-objective optimization algorithms for flow shop scheduling problem: A review and prospects. *Int. J. Adv. Manuf. Technol.* **2011**, *55*, 723–739, doi:10.1007/s00170-010-3094-4.
- Tawhid, M.A.; Savsani, V. Multi-objective sine-cosine algorithm (MO-SCA) for multi-objective engineering design problems. *Neural Comput. Appl.* **2019**, *31*, 915–929, doi:10.1007/s00521-017-3049-x.
- Zuo, L.; Shu, L.; Dong, S.; Zhu, C.; Hara, T. A Multi-Objective Optimization Scheduling Method Based on the Ant Colony Algorithm in Cloud Computing. *IEEE Access* **2015**, *3*, 2687–2699.



7. Zahedi, A.; Kargari, M.; Husseinzadeh Kashan, A. Multi-objective decision-making model for distribution planning of goods and routing of vehicles in emergency multi-objective decision-making model for distribution planning of goods and routing of vehicles in emergency. *Int. J. Disaster Risk Reduct.* **2020**, *48*, 101587, doi:10.1016/j.ijdrr.2020.101587.
8. Yang, W.; Wang, J.; Niu, T.; Du, P. A hybrid forecasting system based on a dual decomposition strategy and multi-objective optimization for electricity price forecasting. *Appl. Energy* **2019**, *235*, 1205–1225, doi:10.1016/j.apenergy.2018.11.034.
9. Lorenz, E.N. Deterministic Nonperiodic Flow. *J. Atmos. Sci.* **1963**, *20*, 130–141, doi:10.1175/1520-0469(1963)020<0130:DNF>2.0.CO;2.
10. Guastello, S.J.; Gregson, R.A. (Eds.) *Nonlinear Dynamical Systems Analysis for the Behavioral Sciences Using Real Data*; CRC Press Taylor & Francis Group: Abingdon, UK, 2011.
11. Ciric, A.R.; Miao, P. Steady state multiplicities in an ethylene glycol reactive distillation column. *Ind. Eng. Chem. Res.* **1994**, *33*, 2738–2748.
12. Kumar, A.; Daoutidis, P. Modeling, analysis and control of ethylene glycol reactive distillation column. *AIChE J.* **1999**, *45*, 51–68.
13. Osulale, F.N.; Zhang, J. Energy efficiency optimisation for distillation column using artificial neural network models. *Energy* **2016**, *106*, 562–578.
14. Tehlah, N.; Kaewpradit, P.; Mujtaba, I.M. Artificial neural network based modeling and optimization of refined palm oil process. *Neurocomputing* **2016**, *216*, 489–501.
15. Mirakhorli, E. Fault diagnosis in a distillation column using a support vector machine based classifier. *Int. J. Smart Electr. Eng.* **2020**, *8*, 105–113.



© 2020 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<http://creativecommons.org/licenses/by/4.0/>).

Article

# An Effective Multi-Label Feature Selection Model Towards Eliminating Noisy Features

Jun Wang <sup>1</sup>, Yuanyuan Xu <sup>2</sup>, Hengpeng Xu <sup>3</sup>, Zhe Sun <sup>4</sup>, Zhenglu Yang <sup>2</sup> and Jinmao Wei <sup>2,\*</sup>

<sup>1</sup> College of Mathematics and Statistics Science, Ludong University, Yantai 264025, China; junwang@mail.nankai.edu.cn

<sup>2</sup> College of Computer Science, Nankai University, Tianjin 300071, China; xuyuanyuan@mail.nankai.edu.cn (Y.X.); yangzlj@nankai.edu.cn (Z.Y.)

<sup>3</sup> Tianjin Key Laboratory of Wireless Mobile Communications and Power Transmission, College of Electronic and Communication Engineering, Tianjin Normal University, Tianjin 300387, China; xuhp@tjnu.edu.cn

<sup>4</sup> RIKEN National Science Institute, Wako, Saitama 351-0198, Japan; zhe.sun.vk@riken.jp

\* Correspondence: weijm@nankai.edu.cn

Received: 21 October 2020; Accepted: 12 November 2020; Published: 15 November 2020



**Abstract:** Feature selection has devoted a consistently great amount of effort to dimension reduction for various machine learning tasks. Existing feature selection models focus on selecting the most discriminative features for learning targets. However, this strategy is weak in handling two kinds of features, that is, the irrelevant and redundant ones, which are collectively referred to as noisy features. These features may hamper the construction of optimal low-dimensional subspaces and compromise the learning performance of downstream tasks. In this study, we propose a novel multi-label feature selection approach by embedding label correlations (dubbed ELC) to address these issues. Particularly, we extract label correlations for reliable label space structures and employ them to steer feature selection. In this way, label and feature spaces can be expected to be consistent and noisy features can be effectively eliminated. An extensive experimental evaluation on public benchmarks validated the superiority of ELC.

**Keywords:** feature selection; noise elimination; space consistency; label correlations

## 1. Introduction

For pattern recognition, feature selection is important for its effectiveness in reducing dimensionality. Feature selection methods are divided into supervised, semi-supervised, and unsupervised ones, according to whether the instances are labeled, partially labeled, or not [1–4]. For supervised cases, class labels are employed for measuring features' discriminative abilities. Many popular and efficient feature selection methods belong to this group [5–10]. Supervised methods are further categorized into three well-known models: filter, wrapper, and embedded [11]. In recent years, some hybrid methods have emerged that combine filter and wrapper processes for enhancing performance and reducing computational cost [12,13].

In another categorization view, existing feature selection approaches can also be grouped to single-label and multi-label ones, whose difference lies in the size of labels that each instance is related with [14]. In single-label FS, instances and labels hold many-to-one connections and the target separability is emphasized in this learning task. With the great potential and success of multi-label learning in many machine learning fields, such as text categorization [15], content annotation [16], and protein location prediction [17], multi-label feature selection has received considerable attention in recent years. We approach the supervised multi-label feature selection in this study.

In multi-label learning, label correlations are the key to combining the complicated relationships among instances, which are typically annotated with multiple labels [18,19]. The mainstream multi-label feature selection strategy is to extract label correlations (via statistical or information-based measurements) and employ them to help find the most remarkable features. A critical issue is, however, this strategy would be trapped by two kinds of features, that is, irrelevant and redundant ones. Irrelevant features represent those lowly discriminative ones. Features of this kind are loosely correlated with learning targets and even may provide misleading information. Compared with irrelevant features, redundant features seem more deceptive. They may exhibit excellent (or comparably superior) performances and mix with remarkable features. Nevertheless, redundant features also lowly contribute to enhancing the discriminative ability of the constructed low-dimensional subspace, because the learning information they provide is redundant with the already distilled information. In general, we regard both irrelevant and redundant features as noisy ones, which may confuse selection processes and compromise the learning performance of downstream tasks.

In this paper, we present an effective multi-label feature selection model by embedding label correlations to eliminate noisy features, named ELC. Our major strategy is to keep feature-label space consistent and explore reliable label structures to drive feature selection. Concretely, we qualitatively assess label correlations in the label space and embed them in feature selection. In this way, the label structure information can be maximally preserved in the constructed low-dimensional subspace, and eventually the consistency between feature and label spaces can be achieved. Furthermore, we devise an efficient framework base on the sparse multi-task learning to optimize ELC, which can help ELC find globally optimal solutions and efficiently converge.

The major contributions of this paper are as follows:

- We present a novel multi-label feature selection model to address the issue of noisy features. This model qualitatively measures label correlations and employs feature-label space consistency to steer feature selection.
- We devised a compact framework to optimize the proposed model. This framework resorts to the multi-task learning strategy and promises globally optimal solutions and efficient convergence.
- Comprehensive experiments on openly available benchmarks were conducted to validate the performance of the proposed model in feature selection and noise elimination.

The remaining parts of this paper are arranged as follows: related works are reviewed in Section 2; the proposed model ELC and its optimization framework are respectively introduced in Section 3 and Section 4; the experimental comparisons of ELC with several popular feature selection approaches are presented in Section 5; finally, conclusions are drawn in Section 6.

## 2. Related Work

Feature selection approaches are commonly specified to a certain recognition scenario, i.e., single-label learning or multi-label learning, because of the different concerns of the two recognition tasks. The issue of noisy feature elimination is firstly raised in single-label feature selection, focusing on removing irrelevant features and picking out discriminative ones. For example, the popular single-label feature selection family by preserving instance similarity [20] directly highly scores the most discriminative features under various statistical metrics, such as the Laplacian score [7,21], the Fisher score [6], the Hilbert–Schmidt independence criterion [22], and the trace ratio [23], just to name a few. In addition to the above similarity preservation approaches, some traditional distance or instance difference based ones can also be deemed as simply pursuing “target-specific features,” such as ReliefF [10], SPEC [24,25], and SPFS [20]. This denotation arises from the fact that target-specific features are picked based only on whether they are strongly correlated with the learning targets. In other words, those features that have excellent discriminative abilities for targets will prevail. The aforementioned approaches have generally achieved excellent performance in eliminating

irrelevant features, while may experience difficulties in improving learning performance due to their scarce attention on removing redundant features.

Recently, some remarkable neural networks-based and fuzzy logic-based feature selection works have been presented, which have received extensive attention due to their excellent feature selection performances [26–28]. For example, Verikas and Bacauskiene [26] proposed a feedforward neural network-based approach to find the salient features and remove those yielding the least accurate classifications. Arefnezhad et al. [27] highly scored the features most related to the drowsiness level via an adaptive neuro-fuzzy inference system, which was devised by combining filter and wrapper feature selection approaches. Cateni et al. [28] selected the mostly relevant features for better binary classification by combining several filter approaches through a fuzzy inference system. Generally speaking, the above studies serve as excellent examples of picking out target-specific features, while still leaving aside the underlying negative effects of noisy features.

A salient but redundant feature provides little valuable learning information if selected. Although this issue is ignored by a majority of feature selection approaches, it gains attention from some information-based ones. Among them, the family based on mutual information is regarded as the mainstream redundancy removing approach. The classical mutual information [9] and its variants (e.g., conditional mutual information) [5,29] can effectively position the redundant features and remove them via a greedy search. Nevertheless, an inevitable problem is that the performances of these approaches heavily depend on their probability estimation accuracy. This problem is more complicated in high-dimensional space.

In terms of multi-label feature selection approaches, they can be roughly categorized into two families. The first family directly divides the multi-label learning into multiple subproblems and utilizes single-label feature evaluation metrics to tackle them [4]. For instance, ReliefF is tailed for multi-label learning by dividing its estimations of nearest misses and hits to eight subproblems [30]. In addition, some single-label feature evaluation strategies are also reformulated to the multi-label ones by enforcing on each subgroup, such as class separability and linear discriminant analysis [31,32]. A major drawback of the above subproblem division strategy is that it ignores label correlations, which encode the underlying label structures for recognition and play critical roles in multi-label learning.

On the other hand, the second family of multi-label feature selection can better fix this issue since it incorporates label correlations into model construction. A common strategy of this family is to evaluate instance-label pairs via specific label ranking metrics and select the features by minimizing loss functions [33–36]. While real-world label relations could be beyond pairwise situations, some high-order correlation approaches have been proposed to model complicated label structures. A feasible solution is to build a common space shared among various labels [16,33,37], which typically suffers from high costs and complex computation. It is noteworthy that in contrast to single-label feature selection approaches, the multi-label ones rarely have the issue of noisy feature elimination. A few approaches specific to ruling out irrelevant features are based on sparse regularization [38]. These approaches neglect the negative effects of redundant features and are not competent in completely removing noisy features.

To comprehensively address the above issues, we will introduce a novel multi-label feature selection model in Section 3, which can effectively filter both kinds of noisy features (including irrelevant and redundant ones) and select the remarkable ones. The proposed model adopts a statistical metric to measure target-related feature redundancy and dispense with any probability estimation. Furthermore, this model extracts label correlations and keeps feature-label space consistency to guide feature selection, which facilitates irrelevant feature exclusion and remarkable feature domination.

### 3. The Methodology: ELC

#### 3.1. Model Description

In this paper, we use  $\{\mathbf{x}_i, \mathbf{y}_i\}_{i=1}^n$  to denote the data set, where  $\mathbf{X} = [\mathbf{x}_1; \dots; \mathbf{x}_n] \in \mathbb{R}^{n \times d}$  represents the instance matrix and instances are characterized by  $d$  features in the feature set  $\mathbf{F} = \{\mathbf{f}_1, \dots, \mathbf{f}_d\}$ .  $\mathbf{Y} = [\mathbf{y}_1, \dots, \mathbf{y}_l] \in \{0, 1\}^{n \times l}$  denotes the target label matrix, where  $y_{ij} = 1$  represents a positive label and  $y_{ij} = 0$  corresponds to a negative one.

Then, we formulate the multi-label feature selection by embedding label correlation (ELC) as follows:

$$\min_{\mathbf{W}} \frac{1}{2} \left\| \hat{\mathbf{Y}}^T \hat{\mathbf{Y}} - \mathbf{S} \right\|_F^2, \text{ s.t. } \hat{\mathbf{Y}} = \frac{1}{n} (\mathbf{X}\mathbf{W})^T \mathbf{Y}, \mathbf{W} \in \{0, 1\}^{d \times l}, \|\mathbf{W}\|_{2,0} = k, \quad (1)$$

where  $\mathbf{S} \in \mathbb{R}^{l \times l}$  represents the label correlation matrix calculated over the initial label matrix, and  $k$  is the number of selected features.  $\mathbf{W} \in \mathbb{R}^{d \times l}$  is the feature selection matrix, where  $w_{ij}$  indicates the importance (also known as weight) of the  $i$ -th feature to the  $j$ -th label.

Equation (1) is actually the feature evaluation function of ELC, which is essentially a Frobenius-norm quadratic model. The matrix  $\mathbf{S}$  represents the label correlations extracted from the label space, and its each element describes a relation between two target labels. These correlations can be easily obtained by some quantitative measurements, including RBF kernel function, Pearson correlation coefficient, etc.  $\hat{\mathbf{Y}}^T \hat{\mathbf{Y}}$  represents the label correlations extracted from the reduced feature space.  $\hat{\mathbf{Y}}^T \hat{\mathbf{Y}}$  is differentiated from  $\mathbf{S}$  on account of the disturbance of noisy features. As described in Section 1, noisy features may distort the structure of the feature space and provide negative learning information. Considering this, ELC evaluates features based on their abilities of preserving label correlations in the feature space, that is, keeping feature-label space consistency. The features that can minimize the discrepancy between  $\hat{\mathbf{Y}}^T \hat{\mathbf{Y}}$  and  $\mathbf{S}$  will be highly scored by ELC. In this way, ELC can be expected to construct an optimal feature subspace with eliminating different kinds of noisy features.

Under the constraint of the  $\ell_{2,0}$ -norm in Equation (1), only  $k$  row in  $\mathbf{W}$  is nonzero. This corresponds to the  $k$  selected features for  $l$  target labels, where 1 represents selected and 0 represents none. Note that  $k$  is most likely to be unequal to  $l$ . That is, more than one feature may be selected responsible for discriminating the same label, or only one feature is discriminative for more than one label. In the former case, multiple features are unified to recognize one target, while one feature deals with multiple recognition sub-tasks in the latter case.

#### 3.2. Property Analysis

The feature subset  $\hat{\mathbf{F}} = \{\hat{\mathbf{f}}_1, \hat{\mathbf{f}}_2, \dots, \hat{\mathbf{f}}_k\}$  that is selected by ELC can be considered as maximally maintaining feature-label space consistency.  $\hat{\mathbf{F}}$  is expected to be constituted by the remarkable features and exclude the noisy ones. In this subsection, we will further analyze the properties of ELC and reveal its underlying characteristics.

Suppose that each feature in  $\mathbf{F}$  has been standardized to have mean zero and unit length. Then, the following things hold for Equation (1):

$$\left\| \hat{\mathbf{Y}}^T \hat{\mathbf{Y}} - \mathbf{S} \right\|_F^2 = \left\| \frac{1}{n^2} \left( \mathbf{Y}^T (\mathbf{X}\mathbf{W}) (\mathbf{X}\mathbf{W})^T \mathbf{Y} \right) - \mathbf{S} \right\|_F^2.$$

This is the objective of ELC. For more clearly illustrating its properties, let  $\hat{\mathcal{S}} = n^2 \mathbf{S}$  and  $\mathcal{H} = \mathbf{Y}^T (\mathbf{X}\mathbf{W}) (\mathbf{X}\mathbf{W})^T \mathbf{Y}$ . Then,

$$\left\| \hat{\mathbf{Y}}^T \hat{\mathbf{Y}} - \mathbf{S} \right\|_F^2 = \frac{1}{n^2} \left( \text{tr}(\mathcal{H}^T \mathcal{H}) + \text{tr}(\hat{\mathcal{S}}^T \hat{\mathcal{S}}) - 2\text{tr}(\hat{\mathcal{S}}^T \mathcal{H}) \right).$$

Three terms are involved in this equation. Clearly,  $\text{tr}(\hat{\mathcal{S}}^T \hat{\mathcal{S}})$  represents the label correlation information extracted from the label space and is constant in the selection process. Thus, it is easy to

conclude that  $\min_{\mathbf{W}} \|\hat{\mathbf{Y}}^T \hat{\mathbf{Y}} - \mathbf{S}\|_F^2$  is equivalent to  $\min_{\mathbf{W}} \text{tr}(\mathcal{H}^T \mathcal{H})$  and  $\max_{\mathbf{W}} \text{tr}(\hat{\mathcal{S}}^T \mathcal{H})$ . Then, two properties of ELC are given as follows:

**Property 1.** Label correlation information can be maximally embedded in feature selection by ELC.

**Proof.**  $\text{tr}(\hat{\mathcal{S}}^T \mathcal{H}) = \text{tr}((\mathbf{XW})^T \mathbf{Y} \hat{\mathcal{S}} \mathbf{Y}^T (\mathbf{XW})) = \sum_{i=1}^k \hat{\mathbf{f}}_i^T (\mathbf{Y} \hat{\mathcal{S}} \mathbf{Y}^T) \hat{\mathbf{f}}_i = \sum_{i=1}^k \hat{\mathbf{f}}_i^T \left( \sum_{c1=1}^l \sum_{c2=1}^l \mathbf{y}_{c1} s_{c1,c2} \mathbf{y}_{c2}^T \right) \hat{\mathbf{f}}_i$ , where  $s_{c1,c2}$  is the correlation degree of the labels  $\mathbf{y}_{c1}$  and  $\mathbf{y}_{c2}$ , and  $\mathbf{XW}$  indicates the selected features. Then, the following things holds:  $\min_{\mathbf{W}} \|\hat{\mathbf{Y}}^T \hat{\mathbf{Y}} - \mathbf{S}\|_F^2 \propto \max_{\mathbf{W}} \sum_{i=1}^k \hat{\mathbf{f}}_i^T \left( \sum_{c1=1}^l \sum_{c2=1}^l \mathbf{y}_{c1} s_{c1,c2} \mathbf{y}_{c2}^T \right) \hat{\mathbf{f}}_i$ .  $\sum_{c1=1}^l \sum_{c2=1}^l \mathbf{y}_{c1} s_{c1,c2} \mathbf{y}_{c2}^T$  can be regarded as the correlation information of pairwise labels. Therefore, ELC can maximally embed label correlations in its feature selection process.  $\square$

Label correlation information is important for multi-label learning. For example, the images about seas may share some common labels for recognition, such as ship, fish, and seagull, and their close correlations may help us distinguish the image category and find their shared features. The existing multi-label learning methods are categorized on the basis of the label correlation orders they consider [39]. Their correlation modeling capabilities directly affect their discriminative performance. As demonstrated in Property 1, ELC can measure the pairwise label correlations. Furthermore, it can also preserve this correlation information in its constructed feature subspace, which is crucial for ELC to eliminate noisy features. In other words, the features that can maximally preserve label correlation information are preferred by ELC. This strategy facilitates ELC building a low-dimensional feature space that is consistent with the label space and also suitable for multi-label learning.

In addition to the above property with respect to maximally embedding label correlations, another important property of ELC is illustrated as follows:

**Property 2.** Feature redundancy can be minimized by ELC.

**Proof.**  $\text{tr}(\mathcal{H}^T \mathcal{H}) = \sum_{i,j=1}^k \left( (\hat{\mathbf{f}}_i^T \mathbf{Y}) (\hat{\mathbf{f}}_j^T \mathbf{Y})^T \right)^2 = \sum_{i,j=1}^k \sum_{c=1}^l \left( \langle \hat{\mathbf{f}}_i, \mathbf{y}_c \rangle \langle \hat{\mathbf{f}}_j, \mathbf{y}_c \rangle \right)^2$   
 $= \sum_{i,j=1}^k \sum_{c=1}^l n^4 \sigma_{y_c}^4 \rho_{\hat{\mathbf{f}}_i, y_c}^2 \rho_{\hat{\mathbf{f}}_j, y_c}^2$ , where  $\sigma_{y_c}$  is the standard deviation of the label  $\mathbf{y}_c$ , and  $\rho_{\hat{\mathbf{f}}_i, y_c}$  and  $\rho_{\hat{\mathbf{f}}_j, y_c}$  are the Pearson correlation coefficients of  $\mathbf{y}_c$  with the features  $\hat{\mathbf{f}}_i$  and  $\hat{\mathbf{f}}_j$ , respectively. Then, we have  $\min_{\mathbf{W}} \|\hat{\mathbf{Y}}^T \hat{\mathbf{Y}} - \mathbf{S}\|_F^2 \propto \min_{\mathbf{W}} \sum_{i,j=1}^k \sum_{c=1}^l n^4 \sigma_{y_c}^4 \rho_{\hat{\mathbf{f}}_i, y_c}^2 \rho_{\hat{\mathbf{f}}_j, y_c}^2$ .

Clearly,  $n$  and  $\sigma_{y_c}$  are constant in the feature selection process.  $\sum_{c=1}^l \rho_{\hat{\mathbf{f}}_i, y_c} \rho_{\hat{\mathbf{f}}_j, y_c}$  can be regarded as the shared label dependency of the features  $\hat{\mathbf{f}}_i$  and  $\hat{\mathbf{f}}_j$ , that is, the feature redundancy for recognizing the target  $\mathbf{y}_c$ . Therefore, ELC can minimize feature redundancy in its feature selection process.  $\square$

Note that the term  $\sum_{c=1}^l \rho_{\hat{\mathbf{f}}_i, y_c} \rho_{\hat{\mathbf{f}}_j, y_c}$  in Property 2 is obtained by introducing the label correlation information. This is a completely novel estimation for the label-specific feature redundancy. The most majority of existing feature selection approaches (including the single-label and multi-label ones) adopt a univariate measurement criterion and merely the top- $k$  features have opportunities to prevail. This strategy largely increases the redundant recognition information shared between features. For example, if we select the genes that are all discriminative for the diabetes type 1, we probably cannot give an accurate diagnosis since these features may be less aware of other types of diabetes. This is why we have to reduce recognition redundancy and enrich recognition information. Some approaches are able to reduce feature redundancy, while their focus is not the label-specific redundancy. For example,  $\sum_{i,j=1}^k \rho_{\hat{\mathbf{f}}_i, \hat{\mathbf{f}}_j}$  is actually reduced in SPFS [20]. This term includes an additional information irrelevant to recognition, and correspondingly, it is inappropriate. In contrast, ELC removes label-specific feature redundancy and is more suitable for multi-label learning with eliminating noisy features.

As discussed above, ELC processes two properties, i.e., maximally preserving label correlation information and minimizing label-specific feature redundancy. These characteristics account for the superior ability of ELC in eliminating noisy features and picking out remarkable ones.

#### 4. Multi-Task Optimization for ELC

Equation (1) describes an integer programming problem, which is NP-hard and complicated to solve. Moreover, the  $\ell_{2,0}$ -norm constraint in Equation (1) is non-smooth, which leads to a slow convergence rate. In this section, we devise an efficient framework to address this problem by using the sparse multi-task learning technology in the proximal alternating direction method (PADM) framework [40].

Suppose the spectral decomposition of the correlation matrix  $\mathbf{S}$  can be denoted as

$$\mathbf{S} = \Phi \Sigma \Phi^T = \Phi \text{diag}(\sigma_1, \dots, \sigma_l) \Phi^T, \sigma_1 \geq \dots \geq \sigma_l,$$

where  $\Phi$  and  $\Sigma$  are respectively the eigenvector and eigenvalue matrices of  $\mathbf{S}$ . Then, Equation (1) can be reformulated as

$$\min_{\mathbf{W}, \mathbf{p}} \frac{1}{2} \left\| \mathbf{Y}^T \mathbf{X} \text{diag}(\mathbf{p}) \mathbf{W} - \Gamma^* \right\|_F^2, \text{ s.t. } \mathbf{W} \in \mathbb{R}^{d \times l}, \|\mathbf{W}\|_{2,1} \leq t, \mathbf{p} \in \{0, 1\}^d, \mathbf{p}^T \mathbf{1} = k, \quad (2)$$

where  $\Gamma^* = n\Phi\Sigma^{1/2}$ ,  $t$  is a hyperparameter to constrain  $\|\mathbf{W}\|_{2,1}$  to a convex solution,  $\mathbf{p}$  is a feature indicator vector that reflects whether the corresponding features are selected or not (1 for selected and 0 for otherwise), and  $\mathbf{1}$  is the vector with all ones.

On the basis of Equation (2), ELC is actually reformulated as a multivariate regression problem, which enables the multi-task learning technology [41]. This technology aims to learn a common set of features to tackle multiple relevant tasks and excels at various sparse learning formulations, including the optimization problem in Equation (1). Based on the multi-task learning technology, we then obtain the equivalent form of ELC as follows:

$$\min_{\mathbf{W}, \mathbf{p}} \frac{1}{2} \left\| \hat{\mathbf{A}} \text{diag}(\mathbf{p}) \mathbf{W} - \Gamma^* \right\|_F^2 + \lambda \|\mathbf{W}\|_{2,1}, \text{ s.t. } \mathbf{p} \in \{0, 1\}^d, \mathbf{p}^T \mathbf{1} = k, \quad (3)$$

where  $\hat{\mathbf{A}} = \mathbf{Y}^T \mathbf{X}$ , and  $\lambda > 0$  is the regularization parameter. Clearly, we can apply the augmented Lagrangian method to solve this problem. Then, Equation (3) is further reformulated as

$$\min_{\mathbf{U}, \mathbf{W}, \mathbf{p}} \frac{1}{2} \left\| \hat{\mathbf{A}} \text{diag}(\mathbf{p}) \mathbf{W} - \Gamma^* \right\|_F^2 + \lambda \|\mathbf{U}\|_{2,1}, \text{ s.t. } \mathbf{U} = \mathbf{W}, \mathbf{p} \in \{0, 1\}^d, \mathbf{p}^T \mathbf{1} = k. \quad (4)$$

The Lagrangian function can be defined as

$$\mathcal{L}(\mathbf{U}, \mathbf{W}, \mathbf{p}, \mathbf{V}) = \frac{1}{2} \left\| \hat{\mathbf{A}} \text{diag}(\mathbf{p}) \mathbf{W} - \Gamma^* \right\|_F^2 + \frac{\beta}{2} \|\mathbf{W} - \mathbf{U}\|^2 + \lambda \|\mathbf{U}\|_{2,1} - \text{tr}(\mathbf{V}^T (\mathbf{W} - \mathbf{U})), \quad (5)$$

where  $\mathbf{V} = (\mathbf{v}_1^T, \dots, \mathbf{v}_d^T)^T \in \mathbb{R}^{d \times l}$  is the Lagrangian multiplier, and  $\beta > 0$  is the penalty parameter.

Equation (5) involves four variables, that is, the auxiliary variable  $\mathbf{U}$ , the feature weight matrix  $\mathbf{W}$ , the feature indicator vector  $\mathbf{p}$ , and the Lagrangian multiplier  $\mathbf{V}$ . Clearly, simultaneously optimizing four variables is impractical. Accordingly,  $\mathbf{V}$  is temporarily fixed for simplification in the following analysis. Then, minimizing  $\mathcal{L}(\mathbf{U}, \mathbf{W}, \mathbf{p}, \mathbf{V})$  is equivalent to the following two subproblems; i.e.,

- $\min_{\mathbf{U}} \mathcal{L}_1(\mathbf{U}) = \min_{\mathbf{U}} \frac{\beta}{2} \|\mathbf{W} - \mathbf{U}\|^2 + \lambda \|\mathbf{U}\|_{2,1} + \text{tr}(\mathbf{V}^T \mathbf{U});$
- $\min_{\mathbf{W}, \mathbf{p}} \mathcal{L}_2(\mathbf{W}, \mathbf{p}) = \min_{\mathbf{W}, \mathbf{p}} \frac{1}{\beta} \left\| \hat{\mathbf{A}} \text{diag}(\mathbf{p}) \mathbf{W} - \Gamma^* \right\|_F^2 + \|\mathbf{W} - \mathbf{U}\|^2 - \frac{2}{\beta} \text{tr}(\mathbf{V}^T \mathbf{W}).$



As to  $\mathcal{L}_1(\mathbf{U})$ , the following holds:

$$\mathcal{L}_1(\mathbf{U}) = \sum_{i=1}^d \left( \frac{\beta}{2} \left\| \mathbf{w}^i - \mathbf{u}^i \right\|^2 + \lambda \left\| \mathbf{u}^i \right\| + \text{tr}(\mathbf{v}_i^T \mathbf{u}^i) \right), \quad (6)$$

where  $\mathbf{w}^i$  and  $\mathbf{u}^i$  are the  $i$ -th row vectors of  $\mathbf{W}$  and  $\mathbf{U}$ , respectively. Then, we reformulate  $\min_{\mathbf{U}} \mathcal{L}_1(\mathbf{U})$  to its close form [41] as

$$\min_{\mathbf{u}^i} \sum_{i=1}^d \left( \frac{\beta}{2} \left\| \mathbf{w}^i - \mathbf{u}^i + \frac{1}{\beta} \mathbf{v}_i \right\|^2 + \lambda \left\| \mathbf{u}^i \right\| \right). \quad (7)$$

Conducting gradient descent on Equation (7) yields the following optimal solution as

$$\mathbf{u}^i = \max \left\{ \left\| \mathbf{w}^i + \frac{1}{\beta} \mathbf{v}_i \right\| - \frac{\lambda}{\beta}, 0 \right\} \frac{\mathbf{w}^i + \frac{1}{\beta} \mathbf{v}_i}{\left\| \mathbf{w}^i + \frac{1}{\beta} \mathbf{v}_i \right\|}. \quad (8)$$

Then, the optimal  $\mathbf{U}$  in iteration  $[t + 1]$  can be denoted as

$$\mathbf{U}^{[t+1]} = \max \left\{ \left\| \mathbf{W}^{[t]} + \frac{1}{\beta} \mathbf{V}^{[t]} \right\| - \frac{\lambda}{\beta}, 0 \right\} \frac{\mathbf{W}^{[t]} + \frac{1}{\beta} \mathbf{V}^{[t]}}{\left\| \mathbf{W}^{[t]} + \frac{1}{\beta} \mathbf{V}^{[t]} \right\|}. \quad (9)$$

In terms of  $\min_{\mathbf{W}, \mathbf{p}} \mathcal{L}_2(\mathbf{W}, \mathbf{p})$ , we let  $\mathcal{P} = \{\mathbf{p} | \mathbf{p} \in \{0, 1\}^d, \mathbf{p}^T \mathbf{1} = k\}$ . The dual problem of  $\min_{\mathbf{W}, \mathbf{p}} \mathcal{L}_2(\mathbf{W}, \mathbf{p})$  is

$$\min_{\mathbf{p} \in \mathcal{P}} \max_{\mathbf{W}} \mathcal{L}_2(\mathbf{W}, \mathbf{p}). \quad (10)$$

Since simultaneously solving the both variables  $\mathbf{p}$  and  $\mathbf{W}$  is still tough, we first fix  $\mathbf{p}$  to optimize  $\mathbf{W}$ . Then, the solution of  $\mathbf{W}$  can be obtained as

$$\left( \text{diag}(\mathbf{p}) \hat{\mathbf{A}}^T \hat{\mathbf{A}} \text{diag}(\mathbf{p}) - \beta \mathbf{I} \right) \mathbf{W} = \text{diag}(\mathbf{p}) \hat{\mathbf{A}}^T \Gamma^* + \beta \mathbf{U} + \mathbf{V}, \quad (11)$$

where  $\mathbf{I}$  is the identity matrix. The structure of  $\hat{\mathbf{A}}^T \hat{\mathbf{A}}$  is commonly not circulant, and therefore the computation of Equation (11) is involved [42]. Considering this, an approximate term is added to  $\mathcal{L}_2(\mathbf{W}, \mathbf{p})$  as follows:

$$\begin{aligned} \tilde{\mathcal{L}}_2(\mathbf{W}, \mathbf{p}) &= \frac{1}{\beta \tau} \left\| \mathbf{W} - \mathbf{W}^{[t]} + \tau \Omega^{[t]} \right\| - \frac{2}{\beta} \text{tr}(\mathbf{V}^T \mathbf{W}) + \left\| \mathbf{W} - \mathbf{U} \right\|^2, \\ \Omega^{[t]} &= \text{diag}(\mathbf{p}^{[t]}) \hat{\mathbf{A}}^T \left( \hat{\mathbf{A}} \text{diag}(\mathbf{p}^{[t]}) \mathbf{W}^{[t]} - \Gamma^* \right), \end{aligned} \quad (12)$$

where  $\tau > 0$ , and  $\mathbf{W}^{[t]}$  is the optimal value of  $\mathbf{W}$  in iteration  $[t]$ . Then, the solution of  $\mathbf{W}^{[t+1]}$  is

$$\mathbf{W}^{[t+1]} = \left( \frac{\tau}{\beta \tau + 1} \right) \left( \beta \mathbf{U}^{[t+1]} + \mathbf{V}^{[t]} + \frac{1}{\tau} (\mathbf{W}^{[t]} - \tau \Omega^{[t]}) \right). \quad (13)$$

The detailed inference can be found in the Appendix A.

Similarly, we can easily obtain the optimal  $\mathbf{p}$  by fixing  $\mathbf{W}$ . Equation (10) is then equivalent to the following minimization problem in this case as follows:

$$\min_{\mathbf{p} \in \mathcal{P}} \left\| \hat{\mathbf{A}} \text{diag}(\mathbf{p}) \mathbf{W} - \Gamma^* \right\|_F^2 = \min_{\mathbf{p} \in \mathcal{P}} \left\| \mathbf{Y}^T \sum_{i=1}^d p_i \mathbf{f}_i \mathbf{w}^i - \Gamma^* \right\|_F^2. \quad (14)$$

Apparently, the top- $k$  features that minimize  $\left\| \mathbf{Y}^T \mathbf{f}_i \mathbf{w}^i - \Gamma^* \right\|_F^2$  can be regarded as the remarkable ones. Their corresponding values in  $\mathbf{p}$  are assigned as 1.

Note that the Lagrangian multiplier  $\mathbf{V}$  is fixed through the above analysis, mainly for simplifying the solution process. We further tackle this problem in the popular PADM framework as illustrated in Algorithm 1. In this framework,  $\mathbf{V}$  can be updated as

$$\mathbf{V}^{[t+1]} = \mathbf{V}^{[t]} - \beta (\mathbf{W}^{[t+1]} - \mathbf{U}^{[t+1]}). \tag{15}$$

---

**Algorithm 1** ELC.

---

**input:**  $\mathbf{F} = \{\mathbf{f}_1, \dots, \mathbf{f}_d\}, \mathbf{Y}, \mathbf{S}, k, \beta, \tau, \lambda$

**output:**  $\mathbf{p}^{[t]}$

```

1: begin
2:  $t = 0, \mathbf{W}^{[0]} = \mathbf{0}_{d \times l}, \mathbf{U}^{[0]} = \mathbf{0}_{d \times l}, \mathbf{V}^{[0]} = \frac{1}{d} \mathbf{1}_{d \times l}$ ;
3: find top- $k$  features  $\hat{\mathbf{f}}_1^{[0]}, \dots, \hat{\mathbf{f}}_k^{[0]}$  that minimize Equation (1), and set  $p_i^{[0]} = \begin{cases} 1, \mathbf{f}_i \in \{\hat{\mathbf{f}}_1^{[0]}, \dots, \hat{\mathbf{f}}_k^{[0]}\} \\ 0, \text{otherwise} \end{cases}$ ;
4: while “not converged” do
5:   optimize  $\mathbf{U}^{[t+1]}$  according to Equation (9);
6:   optimize  $\mathbf{W}^{[t+1]}$  according to Equation (13);
7:   find top- $k$  features  $\hat{\mathbf{f}}_1^{[t+1]}, \dots, \hat{\mathbf{f}}_k^{[t+1]}$  which minimize Equation (14), and set  $p_i^{[t+1]} = \begin{cases} 1, \mathbf{f}_i \in \{\hat{\mathbf{f}}_1^{[t+1]}, \dots, \hat{\mathbf{f}}_k^{[t+1]}\} \\ 0, \text{otherwise} \end{cases}$ ;
8:   update  $\mathbf{V}^{[t+1]}$  according to Equation (15);
9:    $t = t + 1$ ;
10: end while;
11: return  $\mathbf{p}^{[t]}$ ;
12: end;
```

---

ELC in Algorithm 1 is implemented in the regression framework PADM, which is a fast alternating approach for the well-known alternating direction method (ADM) framework. PADM is effective and efficient in solving the minimization problem of the augmented Lagrangian function, and is able to converge to a certain solution  $\{\mathbf{W}^*, \mathbf{U}^*\}$  from any starting point  $\{\mathbf{W}^{[0]}, \mathbf{U}^{[0]}\}$  for any  $\beta > 0$  [40].

In terms of the complexity of ELC, it only takes  $O(k \log d)$  time to find  $k$  remarkable features from the  $d$  candidates. Thus, the time consumption for line 3 is  $O(ndl^2 + k \log d)$ . The cost of the while loop in Algorithm 1 mainly lies in lines 6 and 7, which is  $O(d^2l^2 + ndl^2 + k \log d)$ . As this iteration process is repeated for  $t$  times, its total cost is  $O(t(d^2l^2 + ndl^2 + k \log d))$ . Suppose  $t \gg 1$ . Then, the total complexity of ELC is approximately equal to  $O(t(d^2l^2 + ndl^2 + k \log d))$ , where  $d, n, l, k, t$  are the numbers of features, instances, labels, selected features, and iterations for convergence, respectively.

### 5. Experimental Evaluation

Fourteen groups of multi-label data sets fetched from the Mulan library (<http://mulan.sourceforge.net/datasets-mlc.html>) are taken as the benchmarks in this section, which are shown in Table 1. We compare ELC (the source code is available at <https://github.com/wangjuncs/ELC>) with the following state-of-the-art multi-label feature selection methods:

- MIFS (multi-label informed feature selection) [33]: a label correlation-based multi-label feature selection approach, which maps label information into a low-dimensional subspace and captures the correlations among multiple labels;

- CMFS (correlated and multi-label feature selection) [35]: a feature selection approach based on non-negative matrix factorization, which exploits the label correlation information in features, labels, and instances to select the relevant features and remove the noisy ones;
- LLSF (learning label-specific features) [36]: a unified multi-label learning framework for both feature selection and classification, which models high-order label correlations to select label-specific features.

**Table 1.** Benchmarks for multi-label feature selection.

Data Set	#Features	#Instances	#Labels	Domain
emotions	72	539	6	music
yeast	103	2417	14	biology
birds	260	645	19	audio
enron	1001	1702	53	text
genbase	1186	662	27	biology
business	21,924	11,214	30	text
arts	23146	7484	26	text
education	27,534	12,030	33	text
reaction	30,324	12,828	22	text
health	30,605	9205	32	text
computers	34,096	12,444	33	text
science	37,187	6428	40	text
reference	39,679	8027	33	text
society	49,060	14,512	22	text

More detailed experimental configurations can be found in the Appendix B.

### 5.1. Example 1: Classification Performance

The average classification performance of each feature selection approach is recorded in Table 2 and the pairwise *t*-tests at 5% significance level were conducted to validate the statistical significance. In addition to the traditional precision and AUC metrics, hamming loss penalizes incorrect the recognitions of instances to each target label, ranking loss penalizes the misordered labels in pairs, and one-error penalizes the instances whose top-ranked predicted labels are not in the ground-truth label set. Five metrics evaluated the multi-label classification performance from different aspects.

A single metric is insufficient to illustrate the general classification performance on a dataset. For example, the overall performance of ML-KNN classifier [43] on birds is worse than that on enron under the precision metric, while it shows a better performance on birds than on enron under the AUC metric. Therefore, we extensively used five metrics to compare the performances of the compared approaches. As shown in Table 2, ELC outperforms MIFS, CMFS, and LLSF under various metrics. This superiority is attributed to two reasons. That is, ELC can effectively eliminate noisy features from the candidate feature subsets and maximally embed label correlation information into its selection process. The first term rules out the selection disturbance in the feature space, and the second term promises the proper guiding information extracted from the label space. By seamlessly fusing these two terms, ELC is able to find discriminative features for the downstream learning tasks. This point will be further validated in Sections 5.2 and 5.3.

**Table 2.** Average multi-label classification performance (mean  $\pm$  std.): the best results and those not significantly worse than it are highlighted in bold (pairwise *t*-test at 5% significance level).

Approaches	Data Sets										AVG.		
	Emotions	Yeast	Birds	Enron	Genbase	Business	Arts	Education	Reaction	Health		Computers	Reference
MIFS	0.6667 $\pm$ 0.04	0.7520 $\pm$ 0.02	0.3938 $\pm$ 0.02	0.6139 $\pm$ 0.03	0.7361 $\pm$ 0.15	0.8812 $\pm$ 0.00	0.5108 $\pm$ 0.01	0.6506					
CMFS	0.7221 $\pm$ 0.02	0.7464 $\pm$ 0.01	0.4116 $\pm$ 0.03	0.6206 $\pm$ 0.01	0.7342 $\pm$ 0.15	0.892 $\pm$ 0.00	0.5611 $\pm$ 0.00	0.6697					
LLSF	0.7016 $\pm$ 0.02	0.7532 $\pm$ 0.01	0.4231 $\pm$ 0.07	0.6197 $\pm$ 0.04	0.7352 $\pm$ 0.15	0.8924 $\pm$ 0.00	0.5615 $\pm$ 0.01	0.6695					
ELC	<b>0.7306 <math>\pm</math> 0.02</b>	<b>0.7564 <math>\pm</math> 0.01</b>	<b>0.4671 <math>\pm</math> 0.08</b>	<b>0.6347 <math>\pm</math> 0.01</b>	<b>0.9868 <math>\pm</math> 0.00</b>	<b>0.8931 <math>\pm</math> 0.00</b>	<b>0.5646 <math>\pm</math> 0.00</b>	<b>0.7190</b>					
MIFS	0.5129 $\pm$ 0.01	0.5836 $\pm$ 0.01	0.7391 $\pm$ 0.02	0.6629 $\pm$ 0.01	0.4592 $\pm$ 0.02	0.6267 $\pm$ 0.01	0.5899 $\pm$ 0.01	0.5963					
CMFS	0.6164 $\pm$ 0.01	0.5883 $\pm$ 0.01	0.7437 $\pm$ 0.01	0.6931 $\pm$ 0.00	0.5477 $\pm$ 0.01	0.6718 $\pm$ 0.01	0.6463 $\pm$ 0.01	0.6439					
LLSF	0.6163 $\pm$ 0.01	0.5880 $\pm$ 0.01	0.7435 $\pm$ 0.01	0.6932 $\pm$ 0.00	0.5478 $\pm$ 0.01	0.6720 $\pm$ 0.00	0.6460 $\pm$ 0.01	0.6438					
ELC	<b>0.6213 <math>\pm</math> 0.00</b>	<b>0.5952 <math>\pm</math> 0.00</b>	<b>0.7469 <math>\pm</math> 0.01</b>	<b>0.6962 <math>\pm</math> 0.00</b>	<b>0.5565 <math>\pm</math> 0.00</b>	<b>0.6742 <math>\pm</math> 0.00</b>	<b>0.6500 <math>\pm</math> 0.00</b>	<b>0.6486</b>					

Approaches	Data Sets										AVG.		
	Emotions	Yeast	Birds	Enron	Genbase	Business	Arts	Education	Reaction	Health		Computers	Reference
MIFS	0.6601 $\pm$ 0.53	0.6554 $\pm$ 0.03	0.6497 $\pm$ 0.02	0.5968 $\pm$ 0.03	0.7886 $\pm$ 0.10	0.6371 $\pm$ 0.01	0.6100 $\pm$ 0.01	0.6568					
CMFS	0.7307 $\pm$ 0.03	0.6473 $\pm$ 0.03	0.6403 $\pm$ 0.04	0.6194 $\pm$ 0.01	0.7883 $\pm$ 0.10	0.6821 $\pm$ 0.01	0.6606 $\pm$ 0.01	0.6812					
LLSF	0.7069 $\pm$ 0.02	0.6601 $\pm$ 0.02	0.6793 $\pm$ 0.05	0.6092 $\pm$ 0.03	0.7887 $\pm$ 0.10	0.6824 $\pm$ 0.01	0.6608 $\pm$ 0.01	0.6839					
ELC	<b>0.7513 <math>\pm</math> 0.02</b>	<b>0.6706 <math>\pm</math> 0.02</b>	<b>0.7018 <math>\pm</math> 0.06</b>	<b>0.6385 <math>\pm</math> 0.01</b>	<b>0.9663 <math>\pm</math> 0.00</b>	<b>0.6834 <math>\pm</math> 0.00</b>	<b>0.6659 <math>\pm</math> 0.00</b>	<b>0.7254</b>					
MIFS	0.5830 $\pm$ 0.02	0.7065 $\pm$ 0.01	0.6994 $\pm$ 0.01	0.6364 $\pm$ 0.01	0.6109 $\pm$ 0.02	0.6246 $\pm$ 0.01	0.5964 $\pm$ 0.01	0.6423					
CMFS	0.6753 $\pm$ 0.00	0.7111 $\pm$ 0.01	0.7014 $\pm$ 0.01	0.6864 $\pm$ 0.01	0.6732 $\pm$ 0.01	0.6674 $\pm$ 0.01	0.6430 $\pm$ 0.01	0.6867					
LLSF	0.6761 $\pm$ 0.01	0.7114 $\pm$ 0.01	0.7032 $\pm$ 0.01	0.6859 $\pm$ 0.01	0.6723 $\pm$ 0.01	0.6672 $\pm$ 0.01	0.6427 $\pm$ 0.01	0.6867					
ELC	<b>0.6779 <math>\pm</math> 0.00</b>	<b>0.7188 <math>\pm</math> 0.01</b>	<b>0.7053 <math>\pm</math> 0.01</b>	<b>0.6905 <math>\pm</math> 0.00</b>	<b>0.6789 <math>\pm</math> 0.00</b>	<b>0.6681 <math>\pm</math> 0.01</b>	<b>0.6465 <math>\pm</math> 0.00</b>	<b>0.6911</b>					

Table 2. Cont.

(c) Hamming loss (the lower the better).

Approaches	Data Sets										AVG.			
	Emotions	Yeast	Birds	Enron	Genbase	Business	Arts	Education	Reaction	Health		Computers	Science	Reference
MIFS	0.2865 ± 0.02	0.2006 ± 0.01	0.0538 ± 0.00	0.0544 ± 0.00	0.0303 ± 0.02	0.0270 ± 0.00	0.0610 ± 0.00	0.2865 ± 0.02	0.2006 ± 0.01	0.0538 ± 0.00	0.0544 ± 0.00	0.0303 ± 0.02	0.0270 ± 0.00	0.0610 ± 0.00
CMFS	0.2600 ± 0.01	0.2031 ± 0.01	0.0535 ± 0.00	0.0527 ± 0.00	0.0303 ± 0.02	0.0253 ± 0.00	0.0568 ± 0.00	0.2600 ± 0.01	0.2031 ± 0.01	0.0535 ± 0.00	0.0527 ± 0.00	0.0303 ± 0.02	0.0253 ± 0.00	0.0568 ± 0.00
LLSF	0.2697 ± 0.01	0.1999 ± 0.01	0.0532 ± 0.00	0.0539 ± 0.00	0.0303 ± 0.02	0.0253 ± 0.00	0.0568 ± 0.00	0.2697 ± 0.01	0.1999 ± 0.01	0.0532 ± 0.00	0.0539 ± 0.00	0.0303 ± 0.02	0.0253 ± 0.00	0.0568 ± 0.00
ELC	<b>0.2517 ± 0.02</b>	<b>0.1982 ± 0.01</b>	<b>0.0518 ± 0.00</b>	<b>0.0523 ± 0.00</b>	<b>0.0049 ± 0.00</b>	<b>0.0251 ± 0.00</b>	<b>0.0567 ± 0.00</b>	<b>0.2517 ± 0.02</b>	<b>0.1982 ± 0.01</b>	<b>0.0518 ± 0.00</b>	<b>0.0523 ± 0.00</b>	<b>0.0049 ± 0.00</b>	<b>0.0251 ± 0.00</b>	<b>0.0567 ± 0.00</b>
MIFS	0.0430 ± 0.00	0.0539 ± 0.00	0.0373 ± 0.00	0.0379 ± 0.00	0.0353 ± 0.00	0.0317 ± 0.00	0.0557 ± 0.00	0.0430 ± 0.00	0.0539 ± 0.00	0.0373 ± 0.00	0.0379 ± 0.00	0.0353 ± 0.00	0.0317 ± 0.00	0.0557 ± 0.00
CMFS	0.0371 ± 0.00	0.0536 ± 0.00	0.0368 ± 0.00	0.0350 ± 0.00	0.0322 ± 0.00	0.0280 ± 0.00	0.0511 ± 0.00	0.0371 ± 0.00	0.0536 ± 0.00	0.0368 ± 0.00	0.0350 ± 0.00	0.0322 ± 0.00	0.0280 ± 0.00	0.0511 ± 0.00
LLSF	0.0371 ± 0.00	0.0536 ± 0.00	0.0369 ± 0.00	0.0349 ± 0.00	0.0322 ± 0.00	0.0280 ± 0.00	0.0512 ± 0.00	0.0371 ± 0.00	0.0536 ± 0.00	0.0369 ± 0.00	0.0349 ± 0.00	0.0322 ± 0.00	0.0280 ± 0.00	0.0512 ± 0.00
ELC	<b>0.0368 ± 0.00</b>	<b>0.0531 ± 0.00</b>	<b>0.0366 ± 0.00</b>	<b>0.0348 ± 0.00</b>	<b>0.0319 ± 0.00</b>	<b>0.0278 ± 0.00</b>	<b>0.0508 ± 0.00</b>	<b>0.0368 ± 0.00</b>	<b>0.0531 ± 0.00</b>	<b>0.0366 ± 0.00</b>	<b>0.0348 ± 0.00</b>	<b>0.0319 ± 0.00</b>	<b>0.0278 ± 0.00</b>	<b>0.0508 ± 0.00</b>

(d) Ranking loss (the lower the better).

Approaches	Data Sets										AVG.			
	Emotions	Yeast	Birds	Enron	Genbase	Business	Arts	Education	Reaction	Health		Computers	Science	Reference
MIFS	0.3154 ± 0.05	0.1767 ± 0.01	0.2988 ± 0.01	0.0986 ± 0.01	0.0586 ± 0.03	0.0377 ± 0.00	0.1510 ± 0.05	0.3154 ± 0.05	0.1767 ± 0.01	0.2988 ± 0.01	0.0986 ± 0.01	0.0586 ± 0.03	0.0377 ± 0.00	0.1510 ± 0.05
CMFS	0.2404 ± 0.02	0.1810 ± 0.01	0.2887 ± 0.02	0.0959 ± 0.00	0.0594 ± 0.03	0.0336 ± 0.00	0.1341 ± 0.00	0.2404 ± 0.02	0.1810 ± 0.01	0.2887 ± 0.02	0.0959 ± 0.00	0.0594 ± 0.03	0.0336 ± 0.00	0.1341 ± 0.00
LLSF	0.2711 ± 0.02	0.1756 ± 0.01	0.2777 ± 0.05	0.0962 ± 0.01	0.0591 ± 0.03	0.0335 ± 0.00	0.1341 ± 0.00	0.2711 ± 0.02	0.1756 ± 0.01	0.2777 ± 0.05	0.0962 ± 0.01	0.0591 ± 0.03	0.0335 ± 0.00	0.1341 ± 0.00
ELC	<b>0.2379 ± 0.02</b>	<b>0.1733 ± 0.01</b>	<b>0.2568 ± 0.05</b>	<b>0.0924 ± 0.00</b>	<b>0.0065 ± 0.00</b>	<b>0.0334 ± 0.00</b>	<b>0.1332 ± 0.00</b>	<b>0.2379 ± 0.02</b>	<b>0.1733 ± 0.01</b>	<b>0.2568 ± 0.05</b>	<b>0.0924 ± 0.00</b>	<b>0.0065 ± 0.00</b>	<b>0.0334 ± 0.00</b>	<b>0.1332 ± 0.00</b>
MIFS	0.0988 ± 0.00	0.1459 ± 0.01	0.0498 ± 0.00	0.0820 ± 0.00	0.1351 ± 0.01	0.0849 ± 0.00	0.1372 ± 0.00	0.0988 ± 0.00	0.1459 ± 0.01	0.0498 ± 0.00	0.0820 ± 0.00	0.1351 ± 0.01	0.0849 ± 0.00	0.1372 ± 0.00
CMFS	0.0768 ± 0.00	0.1438 ± 0.01	0.0492 ± 0.00	0.0735 ± 0.00	0.1112 ± 0.00	0.0734 ± 0.00	0.1175 ± 0.00	0.0768 ± 0.00	0.1438 ± 0.01	0.0492 ± 0.00	0.0735 ± 0.00	0.1112 ± 0.00	0.0734 ± 0.00	0.1175 ± 0.00
LLSF	0.0768 ± 0.00	0.1438 ± 0.01	0.0492 ± 0.00	0.0736 ± 0.00	0.1111 ± 0.00	0.0736 ± 0.00	0.1176 ± 0.00	0.0768 ± 0.00	0.1438 ± 0.01	0.0492 ± 0.00	0.0736 ± 0.00	0.1111 ± 0.00	0.0736 ± 0.00	0.1176 ± 0.00
ELC	<b>0.0759 ± 0.00</b>	<b>0.1405 ± 0.00</b>	<b>0.0486 ± 0.00</b>	<b>0.0728 ± 0.00</b>	<b>0.1085 ± 0.00</b>	<b>0.0731 ± 0.00</b>	<b>0.1161 ± 0.00</b>	<b>0.0759 ± 0.00</b>	<b>0.1405 ± 0.00</b>	<b>0.0486 ± 0.00</b>	<b>0.0728 ± 0.00</b>	<b>0.1085 ± 0.00</b>	<b>0.0731 ± 0.00</b>	<b>0.1161 ± 0.00</b>

Table 2. Cont.

(e) One error (the lower the better).

Approaches	Data Sets										AVG.			
	Emotions	Yeast	Birds	Enron	Genbase	Business	Arts	Education	Reaction	Health		Computers	Science	Reference
MIFS	0.4451 ± 0.05	0.2403 ± 0.01	0.7226 ± 0.03	0.3230 ± 0.03	0.3698 ± 0.21	0.1187 ± 0.00	0.6215 ± 0.01	0.4451 ± 0.05	0.2403 ± 0.01	0.7226 ± 0.03	0.3230 ± 0.03	0.3698 ± 0.21	0.1187 ± 0.00	0.6215 ± 0.01
CMFS	0.3871 ± 0.02	0.2445 ± 0.01	0.6968 ± 0.04	0.3093 ± 0.02	0.3719 ± 0.21	0.1061 ± 0.00	0.5518 ± 0.01	0.3871 ± 0.02	0.2445 ± 0.01	0.6968 ± 0.04	0.3093 ± 0.02	0.3719 ± 0.21	0.1061 ± 0.00	0.5518 ± 0.01
LLSF	0.3986 ± 0.03	0.2387 ± 0.01	0.6879 ± 0.08	0.3158 ± 0.05	0.3707 ± 0.21	0.1058 ± 0.00	0.5509 ± 0.01	0.3986 ± 0.03	0.2387 ± 0.01	0.6879 ± 0.08	0.3158 ± 0.05	0.3707 ± 0.21	0.1058 ± 0.00	0.5509 ± 0.01
ELC	<b>0.3664 ± 0.03</b>	<b>0.2361 ± 0.01</b>	<b>0.6192 ± 0.11</b>	<b>0.2988 ± 0.01</b>	<b>0.0123 ± 0.00</b>	<b>0.1050 ± 0.00</b>	<b>0.5464 ± 0.00</b>	<b>0.3664 ± 0.03</b>	<b>0.2361 ± 0.01</b>	<b>0.6192 ± 0.11</b>	<b>0.2988 ± 0.01</b>	<b>0.0123 ± 0.00</b>	<b>0.1050 ± 0.00</b>	<b>0.5464 ± 0.00</b>
MIFS	0.6452 ± 0.02	0.5292 ± 0.02	0.3335 ± 0.03	0.4041 ± 0.01	0.6761 ± 0.02	0.4743 ± 0.01	0.4684 ± 0.01	0.6452 ± 0.02	0.5292 ± 0.02	0.3335 ± 0.03	0.4041 ± 0.01	0.6761 ± 0.02	0.4743 ± 0.01	0.4684 ± 0.01
CMFS	0.4989 ± 0.01	0.5234 ± 0.02	0.3266 ± 0.02	0.3738 ± 0.01	0.5605 ± 0.02	0.4208 ± 0.01	0.3933 ± 0.01	0.4989 ± 0.01	0.5234 ± 0.02	0.3266 ± 0.02	0.3738 ± 0.01	0.5605 ± 0.02	0.4208 ± 0.01	0.3933 ± 0.01
LLSF	0.4993 ± 0.01	0.5237 ± 0.02	0.3267 ± 0.02	0.3735 ± 0.00	0.5605 ± 0.02	0.4202 ± 0.01	0.3937 ± 0.01	0.4993 ± 0.01	0.5237 ± 0.02	0.3267 ± 0.02	0.3735 ± 0.00	0.5605 ± 0.02	0.4202 ± 0.01	0.3937 ± 0.01
ELC	<b>0.4923 ± 0.00</b>	<b>0.5147 ± 0.01</b>	<b>0.3223 ± 0.02</b>	<b>0.3699 ± 0.00</b>	<b>0.5490 ± 0.01</b>	<b>0.4172 ± 0.00</b>	<b>0.3885 ± 0.00</b>	<b>0.4923 ± 0.00</b>	<b>0.5147 ± 0.01</b>	<b>0.3223 ± 0.02</b>	<b>0.3699 ± 0.00</b>	<b>0.5490 ± 0.01</b>	<b>0.4172 ± 0.00</b>	<b>0.3885 ± 0.00</b>

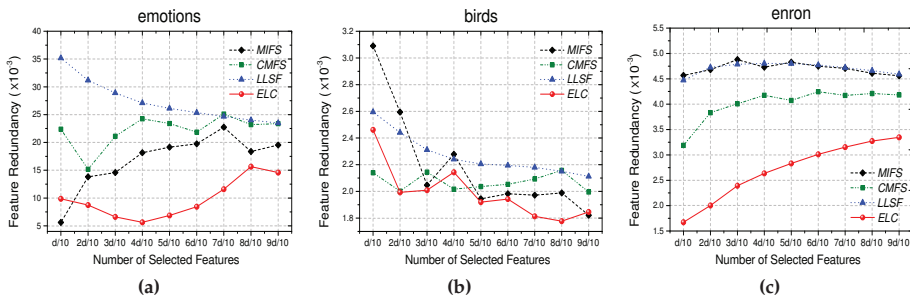
5.2. Example 2: Eliminating Noisy Features

In this section, we evaluate the performances of the compared approaches in eliminating noisy features. We take emotions, birds, and enron as the benchmarks, and measure the residual feature redundancy in the selected feature subset  $\hat{\mathbf{F}}$  as follows:

$$R(\hat{\mathbf{F}}) = \frac{1}{k'(k' - 1)l} \sum_{\hat{\mathbf{f}}_i, \hat{\mathbf{f}}_j \in \hat{\mathbf{F}}} \sum_{c=1}^l \rho_{\hat{\mathbf{f}}_i, y_c}^2 \rho_{\hat{\mathbf{f}}_j, y_c}^2 \tag{16}$$

where  $\rho_{\hat{\mathbf{f}}_i, y_c}$  and  $\rho_{\hat{\mathbf{f}}_j, y_c}$  are the Pearson correlation coefficients of the features  $\hat{\mathbf{f}}_i$  and  $\hat{\mathbf{f}}_j$  with the target label  $y_l$ , and  $k'$  and  $l$  are the numbers of the selected features and labels, respectively. When  $R(\hat{\mathbf{F}})$  reaches its maximum value, the maximal redundant information exists in  $\hat{\mathbf{F}}$ , which interprets as the inferior ability of the selection approach in removing noisy features.

The feature redundancy of  $k'$  selected features for each approach is demonstrated in Figure 1, where  $k' \in \{d/10, 2d/10, \dots, 9d/10\}$  and  $d$  is the total number of original features. It illustrates that ELC is superior in reducing feature redundancy. In other words, ELC can effectively remove redundant features in its multi-label feature selection process. This is one of the crucial factors leading to the excellent discriminative ability of ELC. It should be pointed out that in contrast to the case of single-label feature selection, eliminating noisy features has not received sufficient attention from existing multi-label feature selection approaches. While the issue of noisy features is an obstacle of yielding high selection performance not only for the single-label learning but also for the multi-label cases, we devised ELC to comprehensively tackle this problem. Moreover, the reduced feature redundancy in the majority of redundancy elimination-based approaches is not directly relevant to the target labels. In contrast, ELC quantitatively reduces target-relevant redundancy without any prior probability knowledge, which is conducive to its superiority in multi-label feature selection.



**Figure 1.** Classification redundancy: (a–c) are the classification redundancies produced by the feature selection approaches on the emotions, birds, and enron datasets, and the lower of the redundancy is the better.

5.3. Example 3: Embedding Label Correlations

Label correlation information is important for multi-label learning. In the following experiments, we estimate the preserved label correlation information of the selected feature subset  $\hat{\mathbf{F}}$  as follows:

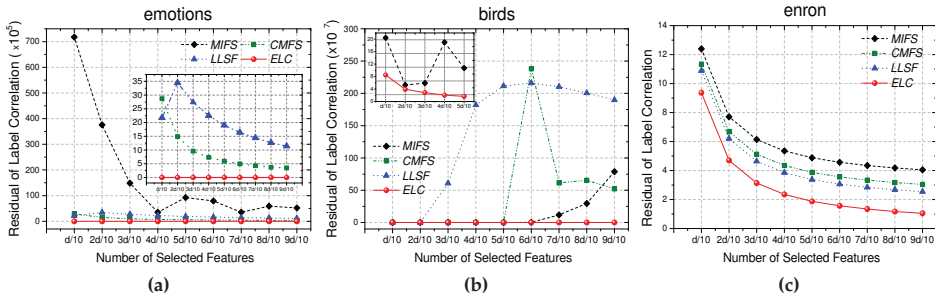
$$C(\hat{\mathbf{F}}) = \frac{1}{k'(k' - 1)} \left\| \frac{1}{n^2} \mathbf{Y}^T \mathbf{X}_{\hat{\mathbf{F}}} \mathbf{X}_{\hat{\mathbf{F}}}^T \mathbf{Y} - \mathbf{S} \right\|_F^2, \tag{17}$$

where  $\mathbf{X}_{\hat{\mathbf{F}}}$  denotes the instances characterized by  $\hat{\mathbf{F}}$  and  $\mathbf{S}$  is the label correlation matrix of the original data. Intuitively, Equation (17) measures the residue scale of label correlation information in the original



and reduced feature spaces. A lower value indicates more information preserved. In other words, more label correlation information can be embedded in the feature selection process in this situation.

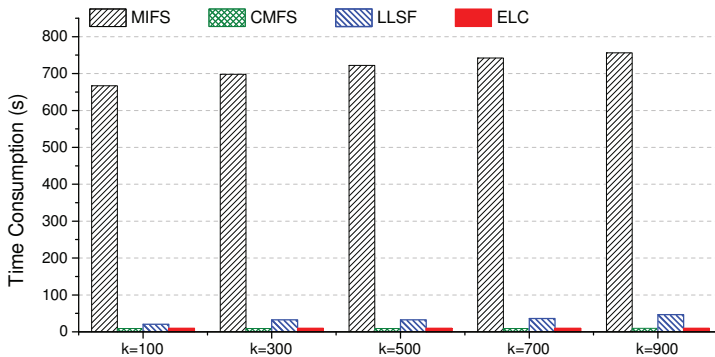
Similarly to the configuration in Section 5.2, we take emotions, birds, and enron as the benchmarks and record  $C(\hat{\mathbf{F}})$  of the  $k'$  features selected by each approach, where  $k' \in \{d/10, 2d/10, \dots, 9d/10\}$ . As shown in Figure 2, ELC is better at preserving the class correlation information than the other multi-label feature selection approaches. Actually, the majority of the existing multi-label feature selection approaches take the label correlation information into consideration to some extent. In contrast to these approaches, ELC quantitatively measures this correlation information and maximally embeds it into the feature selection process. This characteristic, which has already been proved in Property 2, can be further revealed by the experimental results in this section.



**Figure 2.** Residual label correlation information: (a–c) are the residual scales of the label correlation information that are not embedded by the feature selection approaches on the emotions, birds, and enron datasets, and the lower of the residual scale is the better.

5.4. Example 4: Time Consumption

In this section, we compare the approaches in terms of their feature selection efficiency. The time consumption here merely records the feature selection time, excluding the classification cost. All of the tests were implemented in Matlab on an Intel Core i7-4790 CPU (@3.6GHz) with 32GB memory (Intel Corp., Santa Clara, CA, USA). We respectively selected  $k'$  ( $k' \in \{100, 300, 500, 700, 900\}$ ) features on the enron dataset and recorded the time consumption of each compared approach. As illustrated in Figure 3, ELC and CMFS are comparably efficient to converge, while MIFS is most time-consuming, which may be mainly attributed to its involved label clustering process.



**Figure 3.** Time consumption of each multi-label feature selection approach on the enron dataset.

## 6. Conclusions

A novel multi-label feature selection method called ELC is proposed in this paper. ELC embeds label correlation information in reduced feature subspace to eliminate noisy features. In this way, irrelevant and redundant features can be expected to be removed and a discriminative feature subset is constructed for the downstream learning tasks. These advantages help ELC yield good feature selection performance on a wide broad of multi-label data sets under various evaluation metrics.

In terms of optimizing ELC, we can feed it to some gradient descent frameworks to efficiently yield its optimal values, such as Adam with a self-adaptive learning rate [44]. Another interesting and possible exploration would be the consideration of noisy labels, which would induce negative effects on estimating label correlations. According to our pilot study, noisy labels may distort the label space and provide inaccurate guide information for feature selection. How to eliminate noisy labels may inspire our future work.

**Author Contributions:** Each author greatly contributed to the preparation of this manuscript. J.W. (Jun Wang) and J.W. (Jinmao Wei) wrote the paper; Y.X. and H.X. designed and performed the experiments; Z.S. and Z.Y. devised the optimization algorithms. All authors have read and agreed to the published version of the manuscript.

**Funding:** This research was funded by the National Natural Science Foundation of China (number 61772288), the Natural Science Foundation of Tianjin City (number 18JCZDJC30900), the Ministry of Education of Humanities and Social Science Project (number 16YJC790123), the National Natural Science Foundation of Shandong Province (number ZR2019MA049), and the Cooperative Education Project of the Ministry of Education of China (number 201902199006).

**Acknowledgments:** The authors are very grateful to the anonymous reviewers and editor for their helpful and constructive comments and suggestions.

**Conflicts of Interest:** The authors declare no conflict of interest.

## Appendix A

After adding an approximate term to  $\mathcal{L}_2(\mathbf{W}, \mathbf{p})$  and reformulating it to  $\tilde{\mathcal{L}}_2(\mathbf{W}, \mathbf{p})$ , we take the derivative of  $\tilde{\mathcal{L}}_2(\mathbf{W}, \mathbf{p})$  with respect to  $\mathbf{W}$  as follows:

$$\frac{\partial \tilde{\mathcal{L}}_2}{\partial \mathbf{W}} = \beta(\mathbf{W} - \mathbf{U}) - \mathbf{V} + \frac{1}{\tau}(\mathbf{W} - \mathbf{W}^{[t]} + \tau\mathbf{\Omega}^{[t]}), \mathbf{\Omega}^{[t]} = \text{diag}(\mathbf{p}^{[t]})\mathbf{A}^T \left( \mathbf{A} \text{diag}(\mathbf{p}^{[t]})\mathbf{W}^{[t]} - \mathbf{\Gamma}^* \right).$$

To induce the optimal solution of  $\mathbf{W}$ , we make  $\frac{\partial \tilde{\mathcal{L}}_2}{\partial \mathbf{W}}$  equal to 0 and obtain:

$$\left(\beta + \frac{1}{\tau}\right)\mathbf{W} = \beta\mathbf{U} + \mathbf{V} + \frac{1}{\tau}(\mathbf{W}^{[t]} - \tau\mathbf{\Omega}^{[t]}).$$

Then, the optimal solution of  $\mathbf{W}$  in the iteration  $[t + 1]$  can be represented as

$$\mathbf{W}^{[t+1]} = \left(\frac{\tau}{\beta\tau + 1}\right) \left(\beta\mathbf{U}^{[t+1]} + \mathbf{V}^{[t]} + \frac{1}{\tau}(\mathbf{W}^{[t]} - \tau\mathbf{\Omega}^{[t]})\right).$$

## Appendix B. Experimental Configuration

The correlation (or similarity) matrices involved in experiments are all calculated based on the RBF kernel function. Specifically, the label correlation matrix  $\mathbf{S}$  in ELP is defined as

$$\mathbf{S}_{ij} = \begin{cases} \exp\left(-\frac{\|\mathbf{y}_i - \mathbf{y}_j\|^2}{2\delta^2}\right), & \langle \mathbf{y}_i, \mathbf{y}_j \rangle \neq 0 \\ 0, & \text{otherwise} \end{cases}, \text{ where } \delta^2 = \text{mean}(\|\mathbf{y}_i - \mathbf{y}_j\|^2), i, j = 1, \dots, l. \text{ The instance}$$

similarity matrix in SPFS and CMFS is calculated as  $\mathbf{K}_{ij} = \begin{cases} \exp\left(-\frac{\|\mathbf{x}_i - \mathbf{x}_j\|^2}{2\delta^2}\right), & \mathbf{y}_i = \mathbf{y}_j \\ 0, & \text{otherwise} \end{cases}$ ,

where  $\delta^2 = \text{mean}(\|\mathbf{x}_i - \mathbf{x}_j\|^2)$ . The affinity graph in MIFS is constructed as  $\mathbf{K}_{ij} =$

$$\begin{cases} \exp\left(-\frac{\|\mathbf{x}_i - \mathbf{x}_j\|^2}{2\delta^2}\right), & \mathbf{x}_i \in \mathcal{N}_p(\mathbf{x}_j) \text{ or } \mathbf{x}_j \in \mathcal{N}_p(\mathbf{x}_i) \\ 0; & \text{otherwise} \end{cases}$$

instance  $\mathbf{x}_i$ .

SPFS is implemented via the sequential forward selection (SFS) strategy. For a fair comparison, we tune the regularization parameter for all approaches via a grid search from  $\{10^{-3}, 10^{-2}, 10^{-1}, 1, 10\}$ . For ELC, the parameter  $\beta$  is fixed to  $\beta = 10^8$ , and  $\tau$  is set to the spectral radius of  $\hat{\mathbf{A}}^T \hat{\mathbf{A}}$  in the initial state and updated as  $\tau^{[t]} = \frac{1}{\max(\|\psi^i\|)}$  in the  $t$ -th iteration, where  $\psi^i$  is the  $i$ -th row vector of  $\Psi$  and  $\Psi = \hat{\mathbf{A}}^T \hat{\mathbf{A}} \mathbf{V}^{[t]}$ . The convergence state is reached when any of the following two conditions is satisfied: (1)  $t_{max} = 10^3$ ; and (2)  $\|\mathbf{W}^{[t+1]} - \mathbf{W}^{[t]}\| \leq 10^{-4}$ .

Multi-label k-nearest neighbor (ML-kNN) classifier [43] is built on the  $k'$  features selected by each compared approach, when  $k' \in \{d/10, 2d/10, \dots, 9d/10\}$  and  $d$  is the total number of features. All of the numerical features are normalized to zero mean and unit variance, and we employ the excellent features selected by the compared approaches to construct the ML-kNN classifiers and compare their classification performances. The 5-fold cross-validation is conducted, and we report the average performance of the ML-kNN classification under five metrics, i.e., precision, AUC, Hamming loss, ranking loss, and one error [39].

## References

1. Tang, J.; Alelyani, S.; Liu, H. Feature felection for classification: A review. In *Data Classification: Algorithms and Applications*; CRC Press: Chapman, CA, USA, 2014.
2. Wang, J.; Wei, J.; Yang, Z. Supervised feature selection by preserving class correlation. In Proceedings of the 25th ACM International Conference on Information and Knowledge Management, Indianapolis, IN, USA, 24–28 October 2016; pp. 1613–1622.
3. Cai, D.; Zhang, C.; He, X. Efficient and robust feature selection via joint l2,1-norms minimization. In Proceedings of the KDD '10: The 16th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Sydney, Australia, 10–13 August 2015; pp. 333–342.
4. Xu, Y.; Wang, J.; An, S.; Wei, J.; Ruan, J. Semi-supervised multi-label feature selection by preserving feature-label space consistency. In Proceedings of the CIKM '18: Proceedings of the 27th ACM International Conference on Information and Knowledge Management, Orino, Italy, 22–26 October 2018; pp. 783–792.
5. Brown, G.; Pocock, A.; Zhao, M.; Luján, M. Conditional Likelihood Maximisation: A unifying framework for information theoretic feature selection. *J. Mach. Learn. Res.* **2012**, *12*, 27–66.
6. Gu, Q.; Li, Z.; Han, J. Generalized fisher score for feature selection. In Proceedings of the Twenty-Seventh Conference on Uncertainty in Artificial Intelligence, Barcelona, Spain, 14–17 July 2011; pp. 266–273.
7. He, X.; Cai, D.; Niyogi, P. Laplacian score for feature selection. In Proceedings of the 18th International Conference on Neural Information Processing Systems, Shanghai, China, 13–17 November 2011; pp. 507–514.
8. Lin, D.; Tang, X. Conditional Infomax Learning: An Integrated Framework for Feature Extraction and Fusion. In Proceedings of the Computer Vision—ECCV 2006, 9th European Conference on Computer Vision, Graz, Austria, 7–13 May 2006; pp. 68–82.
9. Peng, H.; Long, F.; Ding, C. Feature selection based on mutual information criteria of max-dependency, max-relevance, and min-redundancy. *IEEE Trans. Pattern Anal. Mach. Intell.* **2005**, *27*, 1226–1238.
10. Robnik-Šikonja, M.; Kononenko, I. Theoretical and empirical analysis of ReliefF and RReliefF. *Mach. Learn.* **2003**, *53*, 23–69.
11. Guyon, I.; Elisseeff, A. An introduction to variable and feature selection. *J. Mach. Learn. Res.* **2003**, *3*, 1157–1182.
12. Bermejo, P.; Gámez, J.A.; Puerta, J.M. Speeding up incremental wrapper feature subset selection with Naive Bayes classifier. *Knowl.-Based Syst.* **2014**, *55*, 140–147.
13. Gütlein, M.; Frank, E.; Hall, M.; Karwath, A. Large-scale attribute selection using wrappers. In Proceedings of the 2009 IEEE Symposium on Computational Intelligence and Data Mining, CIDM 2009, Nashville, TN, USA, 30 March–2 April 2009; pp. 332–339.

14. Xu, Y.; Wang, J.; Wei, J. To avoid the pitfall of missing labels in feature selection: A generative model gives the answer. In Proceedings of the AAAI Conference on Artificial Intelligence 2020, New York, NY, USA, 7–12 February 2020; pp. 6534–6541.
15. Chen, W.; Yan, J.; Zhang, B.; Chen, Z.; Yang, Q. Document transformation for multi-label feature selection in text categorization. In Proceedings of the 2007 Seventh IEEE International Conference on Data Mining, Washington, DC, USA, 28–31 October 2007; pp. 451–456.
16. Ma, Z.; Nie, F.; Yang, Y.; Uijlings, J.R.; Sebe, N. Web image annotation via subspace-sparsity collaborated feature selection. *IEEE Trans. Multimedia* **2012**, *14*, 1021–1030.
17. Wang, X.; Li, G.Z. Multilabel learning via random label selection for protein subcellular multilocations prediction. *IEEE/ACM Trans. Comput. Biol. Bioinform.* **2013**, *10*, 436–446.
18. Zhang, Z.L.; Zhou, Z.H. A review on multi-label learning algorithms. *IEEE Trans. Knowl. Data Eng.* **2014**, *26*, 1819–1837.
19. Rivolli, A.; J, J.R.; Soares, C.; Pfahringer, B.; de Carvalho, A.C. An empirical analysis of binary transformation strategies and base algorithms for multi-label learning. *Mach. Learn.* **2020**, *9*, 1–55.
20. Zhao, Z.; Wang, L.; Liu, H.; Ye, J. On similarity preserving feature selection. *IEEE Trans. Knowl. Data Eng.* **2013**, *25*, 619–632.
21. Zhao, J.; Lu, K.; He, X. Locality sensitive semi-supervised feature selection. *Neurocomputing* **2008**, *71*, 1842–1849.
22. Zhang, Y.; Zhou, Z.H. Multi-label dimensionality reduction via dependence maximization. *ACM Trans. Knowl. Discovery Data* **2010**, *4*, 1503–1505.
23. Nie, F.; Xiang, S.; Jia, Y. Trace ratio criterion for feature selection. In Proceedings of the Twenty-Third AAAI Conference on Artificial Intelligence, AAAI 2008, Chicago, IL, USA, 13–17 July 2008; pp. 671–676.
24. Zhao, Z.; Liu, H. Spectral feature selection for supervised and unsupervised learning. In Proceedings of the 24th International Conference on Machine Learning, ICML 2007, Corvallis, OR, USA, 20–24 June 2007; pp. 1151–1157.
25. Zhao, Z.; Wang, L.; Liu, H. Efficient spectral feature selection with minimum redundancy. In Proceedings of the Twenty-Fourth AAAI Conference on Artificial Intelligence, Atlanta, GA, USA, 11–15 July 2010; pp. 673–678.
26. Verikas, A.; Bacauskiene, M. Feature selection with neural networks. *Pattern Recog. Lett.* **2002**, *23*, 1323–1335.
27. Arefnezhad, S.; Samiee, S.; Eichberger, A.; Nahvi, A. Driver drowsiness detection based on steering wheel data applying adaptive neuro-fuzzy feature selection. *Sensors* **2019**, *14*, 943.
28. Cateni, S.; Colla, V.; Vannucci, M. A fuzzy system for combining filter features selection methods. *Int. J. Fuzzy Syst.* **2017**, *19*, 1168–1180.
29. Wang, J.; Wei, J.M.; Yang, Z.; Wang, S.Q. Feature selection by maximizing independent classification information. *IEEE Trans. Knowl. Data Eng.* **2017**, *29*, 828–841.
30. Kong, D.; Ding, C.; Huang, H.; Zhao, H. Multi-label ReliefF and F-statistic feature selections for image annotation. In Proceedings of the 2012 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Providence, RI, USA, 16–21 June 2012; pp. 2352–2359.
31. Ji, S.; Ye, J. Linear dimensionality reduction for multi-label classification. In Proceedings of the 21st International Joint Conference on Artificial Intelligence, Pasadena, CA, USA, 11–17 July 2009; pp. 1077–1082.
32. Wang, H.; Ding, C.; Huang, H. Multi-label linear discriminant analysis. In Proceedings of the 11th European Conference on Computer Vision, Heraklion, Crete, Greece, 5–11 September 2010; pp. 126–139.
33. Jian, L.; Li, J.; Shu, K.; Liu, H. Multi-label informed feature selection. In Proceedings of the Twenty-Fifth International Joint Conference on Artificial Intelligence, New York, NY, USA, 9–15 July 2016; pp. 1627–1633.
34. Huang, J.; Li, G.; Huang, Q.; Wu, X. Joint feature selection and classification for multilabel learning. *IEEE Trans. Cybern.* **2018**, *48*, 876–889.
35. Braytee, A.; Liu, W.; Catchpoole, D.R.; Kennedy, P.J. Multi-label feature selection using correlation information. In Proceedings of the 2017 ACM Conference on Information and Knowledge Management, Singapore, 6–10 November 2017; pp. 1649–1656.
36. Huang, J.; Li, G.; Huang, Q.; Wu, X. Learning label-specific features and class-dependent labels for multi-label classification. *IEEE Trans. Knowl. Data Eng.* **2016**, *28*, 3309–3323.

37. Ji, S.; Tang, L.; Yu, S.; Ye, J. Extracting shared subspace for multi-label classification. In Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Las Vegas, NV, USA, 24–27 August 2008; pp. 381–389.
38. Nie, F.; Huang, H.; Cai, X.; Ding, C.H. Efficient and robust feature selection via joint  $2,1$ -norms minimization. In Proceedings of the 4th Annual Conference on Neural Information Processing Systems 2010, Vancouver, BC, Canada, 6–9 December 2010; pp. 1813–1821.
39. Zhang, M.L.; Wu, L. LIFT: Multi-label learning with label-specific features. *IEEE Trans. Pattern Anal. Mach. Intell.* **2015**, *37*, 107–119.
40. Xiao, Y.H.; Song, H.N. An inexact alternating directions algorithm for constrained total variation regularized compressive sensing problems. *J. Math Imaging Vision* **2012**, *44*, 114–127.
41. Gong, P.; Zhou, J.; Fan, W.; Ye, J. Efficient multi-task feature learning with calibration. In Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, New York, NY, USA, 10–13 August 2014; pp. 761–770.
42. Horn, R.A.; Johnson, C.R. *Matrix Analysis*, 2nd ed.; Cambridge University: Cambridge, UK, 2012.
43. Zhang, M.L.; Zhou, Z.H. ML-kNN: A lazy learning approach to multi-label learning. *Pattern Recog.* **2007**, *40*, 2038–2048.
44. Kingma, D.K.; Ba, J. Adam: A method for stochastic optimization. In Proceedings of the 3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, 7–9 May 2015.




**Publisher's Note:** MDPI stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.



© 2020 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<http://creativecommons.org/licenses/by/4.0/>).

Article

# Automatic Classification of Morphologically Similar Fish Species Using Their Head Contours

Pere Marti-Puig <sup>1,\*</sup>, Amalia Manjabacas <sup>2</sup> and Antoni Lombarte <sup>2,\*</sup>

<sup>1</sup> Data and Signal Processing Group, University of Vic—Central University of Catalonia, 08500 Vic, Catalonia, Spain

<sup>2</sup> Institut de Ciències del Mar, ICM (CSIC), 08003 Barcelona, Catalonia, Spain; manjabacas@icm.csic.es

\* Correspondence: pere.marti@uvic.cat (P.M.-P.); toni@icm.csic.es (A.L.)

Received: 26 March 2020; Accepted: 12 May 2020; Published: 14 May 2020



**Abstract:** This work deals with the task of distinguishing between different Mediterranean demersal species of fish that share a remarkably similar form and that are also used for the evaluation of marine resources. The experts who are currently able to classify these types of species do so by considering only a segment of the contour of the fish, specifically its head, instead of using the entire silhouette of the animal. Based on this knowledge, a set of features to classify contour segments is presented to address both a binary and a multi-class classification problem. In addition to the difficulty present in successfully discriminating between very similar forms, we have the limitation of having small, unreliably labeled image data sets. The results obtained were comparable to those obtained by trained experts.

**Keywords:** open contours; similarly shaped fish species; Discrete Cosine Transform (DCT); Discrete Fourier Transform (DFT); Extreme Learning Machines (ELM); feature engineering; small data-sets

## 1. Introduction

Being able to determine the differences in shape between closely related fish species not only has an impact on fisheries and their management but also is essential for a variety of studies on marine ecology, genetics, phylogeny, biological invasion, and anthropogenic impact on the environment, among others [1]. All of these studies require the correct identification of a wide range of species, and the automatic classification of such species from images is still an open problem. The correct identification of many specimens of these sympatric species requires an experienced observer and is a highly time-consuming task; otherwise, many misclassification errors occur.

This paper focuses on two cases in which this problem appears. The first case deals with the separation between Red Mulletts, (*Mullus barbatus* and *Mullus surmuletus*), and it will be treated as a binary classification problem. The second case is a multiclassification problem in which different species of Gurnard *Chelidonichthys cuculus*, *Chelidonichthys lucerna*, *Chelidonichthys obscurus*, *Eutrigla gurnardus*, *Lepidotrigla cavillone*, *Lepidotrigla dieuzeidei*, *Peristedion cataphractum*, *Trigla lyra*, and *Trigloporus lastoviza* are involved. Red Mulletts and Gurnards are commonly caught and sold in fish markets as a mixture of species (Figure 1).

The observation is that experts who can correctly classify individuals of these species of fish by using shape characteristics—instead of considering the entire silhouette, represented as a closed contour—identify them by focusing on a specific contour fragment. They disregard the remaining part of the contour as they realise that it is irrelevant for classification purposes. This is because the discarded contour part provides information on variations within individuals of the same species with respect to age, size, sex or body condition, but does not provide information on the intrinsic characteristics involved in the distinction between species. The challenge for automated identification

is to take these species-specific open contour fragments pointed out by the specialists to develop features that correctly identify the species automatically. In both cases considered in this work, experts mark the segment of interest of the profile as being the part that goes from the nose to the beginning of the dorsal fin of the fish.

The first case addressed deals with the binary classification of the *Mullus* network. Red Mulletts (*M. barbatus* and *M. surmuletus*) are a major target as demersal fish species for the Mediterranean fish industry [2,3]. There are many comparative studies of these two species focusing on a range of topics, such as pollution bio-indicators, and fisheries or studies on age composition, growth, life history, feeding, genetics, ecology, and distribution [4–9]. As mentioned above, the evaluation of living resources requires the correct identification of species.

Searches for morphological features to help with the task of identifying these sympatric species have focused on characteristics, such as head and body measurements [10,11], barbells [12], dentition [13] or otoliths [14]. Amongst all of these biometric features, the angle of the head has been found to be one of the best diagnostic characteristics to distinguish between the two species, as the angle that the *M. barbatus* head contour forms is more marked than that of *M. surmuletus* [10,15]. Even in fishery studies, however, the individual variability within Red Mulletts causes identification problems and the only reliable distinction is the color pattern of their bent dorsal fin. However, it is not easy to obtain images of this feature from boat moorings in fishing harbours as the fish's dorsal fin is folded. Furthermore, in most cases, the information on the fish caught that is given upon the arrival of the boat does not discriminate between both species of Red Mulletts.

Thus, it is not possible to determine the statistics of individualized species of fish caught from the fishing industry, which adds error in the mathematical models of each species. In consequence, the correct identification of many specimens of these sympatric species requires an experienced observer and, as mentioned before, is a highly time-consuming task. A more complex case occurred in Mediterranean Gurnard fish (in the family of triglids and peristids), consisting of nine sympatric species (*C. cuculus*, *C. lucerna*, *C. obscurus*, *E. gurnardus*, *L. cavillone*, *L. dieuzeidei*, *P. cataphractum*, *T. lyra*, and *T. lastoviza*), which are fished through trawl fishing on continental shelves and upper slopes, and are usually considered as only one species or an unidentified mix of species in commercial capture statistics (Fisheries Database of the Catalan Government). Morphological differences between species are based on trends of the lateral line, scale distribution, head shape and preorbital bones [15], and a high degree of expertise is required to classify triglid species.

To solve both classification problems, we used a traditional machine learning (ML) workflow to create the models, which are the trained programs that predict the outputs based on a set of given inputs, also called features. That implies the design of such features as a first step, followed by the task of training the model. Conventional ML techniques include support vector machines (SVM), ensemble methods, decision trees, and also the extreme learning machines (ELM) used in this work, among others. In this context, ELM performs very similarly to SVM, with the added advantage of having a much quicker training process in comparison. In contrast to ML, an approach based on deep learning (DL) techniques exists in which the algorithms automatically learn which of the features are useful. The term *deep* refers to the multiple *layers* the networks have between inputs and outputs. In order to train models with DL techniques, a large amount of tagged data is required, which is complicated in our case as we work with a small dataset.

Several parameterization strategies have been previously explored for closed contours, and therefore an abundance of literature is available. An important part of the transformed methods dealing with closed contours is based on the Fourier transform. Among those methods, *elliptic Fourier descriptors* (EDF) is one of the most popular ones. This method was introduced by Reference [16], and some variations have been introduced since. As applications that require shape quantification are present in a wide range of different fields, an extensive amount of these can be found [17–20].

Fourier methods applied to periodic sequences (which can be associated with closed contours) exhibit good information-packing properties in the sense that they concentrate most of the signal



energy in a few coefficients. Those few coefficients are employed as features for classifying purposes. The problem under discussion cannot be solved if the specimens are treated by considering their entire silhouette independently of the contour description technique employed, as is the case of multiple spatial scales methods curvature scale space (CSS) [21–23], or from the *discrete wavelet transform* (DWT) [24].

Therefore, knowing how the specialists solve the problem, the approach that will be used will imitate the one used by these experts and thus only consider the contour fragments of interest. Unfortunately, in the context of shape analysis, there is less work done on open contours and discrete transforms. Something that is well known, however, is that Fourier-based methods do not work well when the sequences are not periodic [25].

Different approaches can be used to deal with segment contours (open contours). In Reference [26], a wavelet analysis with cubic B-spline bases was used in order to reduce disturbances and confine them around the discontinuities, and in Reference [27], the windowed short time Fourier transform (STFT) was employed. The main problem with most transformed methods, especially those using Fourier, is the characterisation of the ends of the segments, which, in the case of open contours, are seen as discontinuities. In Reference [28], a large number of open contour segment descriptors were compiled and compared with each other, but none were found to be related to discrete transforms. Certain *discrete cosine transform* (DCT) formulations work better when it comes to the ends of the segments, and therefore seem to be the most appropriate discrete transformation to use when dealing with open contours.

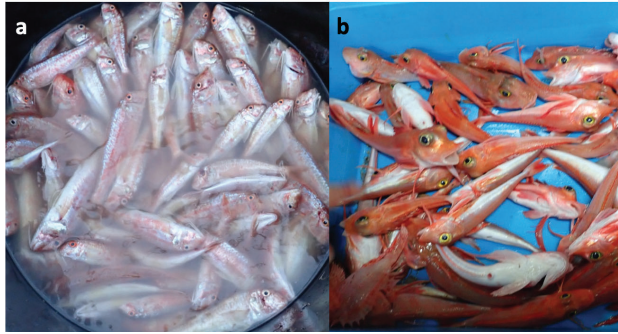
As far as we know, the DCT used in the morphometric analysis was first employed in Reference [29] for the analysis of ammonite ribs (shell features of an extinct order of cephalopods). In Reference [30], the DCT was also used for symmetric closed contours that were simultaneously represented with closed and open contours (by taking half of the closed contour) and analyzed by EFD and DCT, respectively. Recently, in Reference [31], DCT was employed to quantify open contours of organic shapes and thus evaluate the morphological disparity of male genital evolution in sibling species of *Drosophila*. The point (which will be developed later) is that some formulations of the *discrete cosine transform* (DCT) can deal with segment extremes better and can also simultaneously compact the energy of the signal in very few coefficients.

In this paper, an ML method is proposed to automatically classify very similar species with an accuracy comparable to that achieved by experts when they use images to identify the correct species. The method is used both in a binary and in a multi-class classification problem. The images contain two landmarks, one in the snout and one at the beginning of the dorsal fin, to limit the segment used by the experts. An added issue to the classification, which is very difficult if we do not have the information that the specialist uses to discriminate, is that we have very few correctly tagged images. We find this problem even more pronounced in the multi-class classification problem.

The addressed task is established under the framework of supervised classification with databases, as mentioned, of reduced dimensions, which implies that there are few available cases to train the classifiers. The proposed method takes advantage of the knowledge of specialists in the field. As the main contributions, we highlight the development of features to represent open contours in a compact way, and by taking advantage of the appropriate contour normalization, we show that it is possible to eliminate the Gibbs effect that appears with the discrete Fourier transform (DFT). The proposed segment standardization makes it possible—both for the DCT and the FFT—to describe open contours with a single sequence instead of two (one for each coordinate) while, in addition, allowing us to compare the segments independently of rotations and scale variations. It is also robust to the small imprecisions or the noise that can be introduced when finding the contour.

The work continues as follows. In Section 2, Materials and Methods, the following points are presented: the databases and the discrete transforms as well as the corresponding implicit sequence extension and its relation in the representation of open contours, the contour extraction procedure and normalization and the development of the features used. Additionally in that section,

the training, classification, and cross-validation methods are explained. Section 3 is devoted to the Results. In Section 3.2, we show how with a convolutional neural network (CNN) trained to detect the region of interest determined by the points of the expert, the region of interest can also be determined automatically, thus demonstrating that the points are not required. Finally, Section 4 is devoted to our Conclusions.



**Figure 1.** A presentation of the traditional mixture that is made by fishermen in the Catalan coast: (a) red mullets, and (b) gurnards.

## 2. Materials and Methods

### 2.1. Discrete Transforms and Signal Reconstructions from a Reduced Set of Coefficients

In the following two subsections, we show the direct and reverse expressions of the DCT and DFT transforms, and their particularities when reconstructing in an approximated way the original sequence using a limited number  $L$  of coefficients. The small set of those coefficients, used properly, will be employed to develop the features for the classifier. Intuitively, the set of  $L$   $X_k$  coefficients will be a good set of contour descriptors as long as  $L \ll N$  and the samples of  $\tilde{x}_n$  are very close to the original  $x_n$ .

#### 2.1.1. Discrete Cosine Transform (Type-II)

The DCT-II forward and backward expressions that relate the points  $x_n \in \mathbb{R}$  with the transformed coefficients  $X_k \in \mathbb{R}$  take the form:

$$X_k = \sum_{n=0}^{N-1} c_k x_n \cos \left( \frac{\pi}{N} \left( n + \frac{1}{2} \right) k \right); k = 0, \dots, N - 1 \tag{1}$$

$$x_n = \sum_{k=0}^{N-1} c_k X_k \cos \left( \frac{\pi}{N} \left( n + \frac{1}{2} \right) k \right); n = 0, \dots, N - 1, \tag{2}$$

where  $c_k = \sqrt{\frac{1}{N}}$  if  $k = 0$  and  $c_k = \sqrt{\frac{2}{N}}$  if  $k \neq 0$ . The approximate reconstruction of  $x_n$  with the first  $L$  coefficients  $X_k$ ,  $L < N$  is:

$$\tilde{x}_n = \sum_{k=0}^{L-1} c_k X_k \cos \left( \frac{\pi}{N} \left( n + \frac{1}{2} \right) k \right); n = 0, \dots, N - 1, \tag{3}$$

where  $\tilde{x}_n$  are  $N$  reconstructed samples.

#### 2.1.2. Discrete Fourier Transform

The discrete Fourier transform (DFT) forward and backward expressions that relate the points  $x_n \in \mathbb{R}$  with the transformed coefficients  $X_k \in \mathbb{C}$  take the form:

$$X_k = \sum_{n=0}^{N-1} x_n e^{-j\frac{2\pi}{N}kn}, \quad k = 0, 1, \dots, N - 1, \tag{4}$$

$$x_n = \frac{1}{N} \sum_{k=0}^{N-1} X_k e^{j\frac{2\pi}{N}kn}, \quad n = 0, 1, \dots, N - 1. \tag{5}$$

In the DFT case,  $X_k \in \mathbb{C}$ . Considering the well-known DFT property  $X_{N-k} = X_k^*$  (symbol \* stands for complex conjugated) which holds for  $x_n \in \mathbb{R}$ , the module and phase notation of  $X_k$  as  $X_k = |X_k| e^{j\phi_k}$ , and the Euler formula, the approximation of  $x_n$ ,  $\tilde{x}_n$ , obtained from the first  $L$  ( $L < N$ ) coefficients  $X_k$ , can be written as:

$$\tilde{x}_n = \frac{X_0}{N} + \frac{1}{N} \sum_{k=1}^{L-1} \left( X_k e^{j\frac{2\pi}{N}kn} + X_k^* e^{-j\frac{2\pi}{N}kn} \right) = \frac{X_0}{N} + \frac{2}{N} \sum_{k=1}^{L-1} |X_k| \cos \left( \frac{2\pi}{N}kn + \phi_k \right); \tag{6}$$

$n = 0, \dots, N - 1,$

showing that  $\tilde{x}_n \in \mathbb{R}$ .

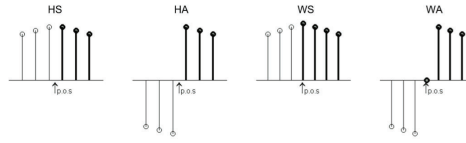
### 2.1.3. A Note on the Implicit Extension of the Sequences Depending on the Discrete Transform Used and the Case of Open Contours

For a discrete transform, the way in which the sequence remains implicitly defined outside the main range  $n \in [0, \dots, N - 1]$  is known as implicit extension. Any discrete transform has a particular way of extending the sequence. This can be seen from the backward transform expressions in (2) and in (5), for the DCT-II and the DFT, respectively. As the bases in these expressions are defined beyond  $n \in [0, N - 1]$ , it follows that  $x_n$  also remains implicitly defined beyond this range. When the approximation of  $x_n$  by  $\tilde{x}_n$  is done with few bases ( $L \ll N$ ) as in Equations (3) and (6), the samples of  $\tilde{x}_n$  near the main sequence extremes ( $n = 0$  and  $n = N - 1$ ) tend to manifest the implicit periodicity of the transform. Depending on how a particular transformation extends the sequences at both extremes, more or less transformed coefficients will be required to have a quality approximation.

Let us first consider the DCT-II, which belongs to the family of discrete trigonometric transforms (DTT), and their implicit extension properties [32–40]. To summarize the DTT sequence extensions, two relevant parameters must be considered—the *symmetry type* (ST) and the *point of symmetry* (PoS). When it comes to the ST, the replication can be performed *symmetrically* (S) or *anti-symmetrically* (A). Considering that the samples of the sequence are equispaced, the PoS can be positioned either in the middle of the space between the elements, named *half* (H) sample or *midpoint*, or just at the position of the element, named *whole* (W) sample or *meshpoint*.

Thus, when considering a single point from an edge of a sequence, there are four options to extend the sequence, commonly denoted by two letters—(HS), (HA), (WS), and (WA). Figure 2 represents those four possibilities for the left edge of a main sequence, which is depicted in bold. Then, it follows that since all finite sequence have two edges, there are eight total possible combinations, each one of which is associated with one different formulation corresponding to either the DCT or the DST (discrete sine transform) groups [34,37,41,42].

The DTTs that best approximate open contours with few coefficients are those that have HS or WS implicit extensions on both extremes. That happens for the DCT formulation types -I, -II, -V, and -VI. In those cases, the sequence considered with implicit extensions does not lose continuity. In the particular case of the DCT-II, the continuity in both extremes is defined as  $x_{-1} = x_0$  and  $x_N = x_{N-1}$ .



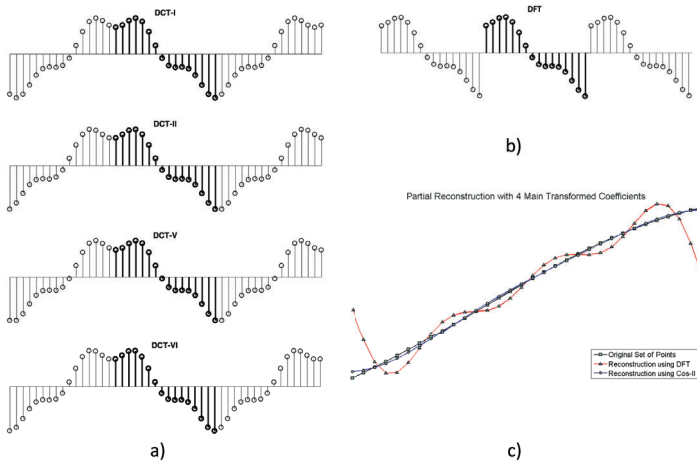
**Figure 2.** The four possible sequence extensions in discrete trigonometric transforms (DTTs) depending on the point of symmetry (POS) and the type. *HS*, *HA*, *WS*, and *WA* stand for half sample-symmetry, half sample anti-symmetry, whole sample-symmetry and whole sample anti-symmetry, respectively.

In the case of the DFT, it is easy to verify that  $x_{N+n} = x_n$ . From (5) we have:

$$x_{n+N} = \frac{1}{N} \sum_{k=0}^{N-1} X_k e^{j\frac{2\pi}{N}k(n+N)} = \frac{1}{N} \sum_{k=0}^{N-1} X_k e^{j\frac{2\pi}{N}kn} \cdot e^{j2\pi k} = x_n. \tag{7}$$

Thus, for the DFT, reconstructions with few coefficients are very efficient only when  $x_0$  and  $x_{N-1}$  are closed, as occurs in the case of objects with smooth contours. When  $x_0 \approx x_{N-1}$ , as  $x_{-1} = x_{N-1}$  and  $x_N = x_0$ , the continuity is perceived in the extended sequence. However, DFT loses the property of compacting energy in few coefficients when a discontinuity appears as  $x_0$  differs from  $x_{N-1}$ , which usually is the case of open contours [25]. See Figure 3c.

When the open contour is directly represented with the  $\tilde{x}_n$  sequence without any transformation (one sequence for each contour coordinate), the most efficient representation, which uses fewer coefficients, is obtained with the DCT-II. Figure 3a shows (with the main sequence represented in bold) the implicit extensions of the sequence outside the main range for DCT formulations -I, -II, -V, and -VI. Figure 3b shows the DFT implicit extensions for the same main sequence. Figure 3c represents the reconstruction of a set of 32 original points using four real DCT-II coefficients and seven complex DFT coefficients.

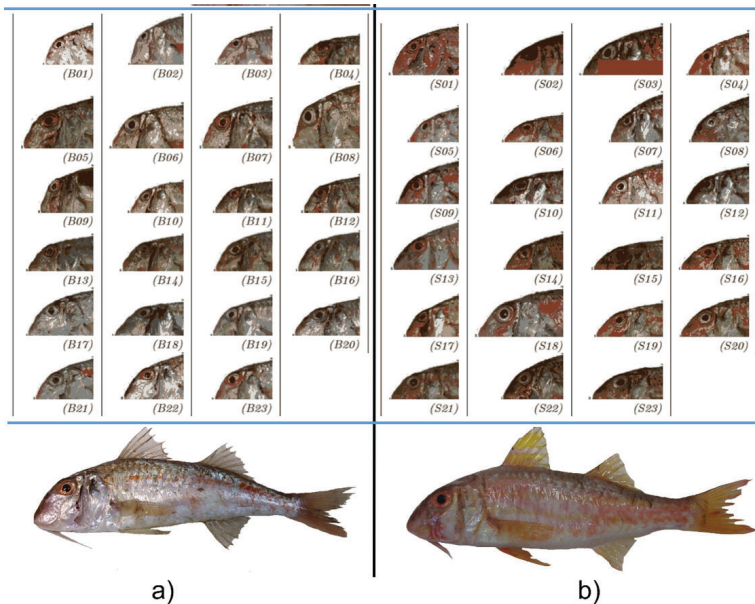


**Figure 3.** (a) Implicit extensions outside the main range for discrete cosine transform (DCT) formulations -I, -II, -V and -VI. The main sequence is in bold. (b) Discrete Fourier transform (DFT) implicit extensions for the same sequence. (c) Reconstruction of a set of 32 original points (squares) using four real DCT-II coefficients  $c_0, c_1, c_2, c_3$  (circles) and seven complex DFT coefficients  $f_0, f_1, f_2, f_3$  and  $f_{29}, f_{30}, f_{31}$  (triangles). Note that  $f_{29}, f_{30}, f_{31}$  are the complex conjugates of  $f_3, f_2, f_1$ , respectively.

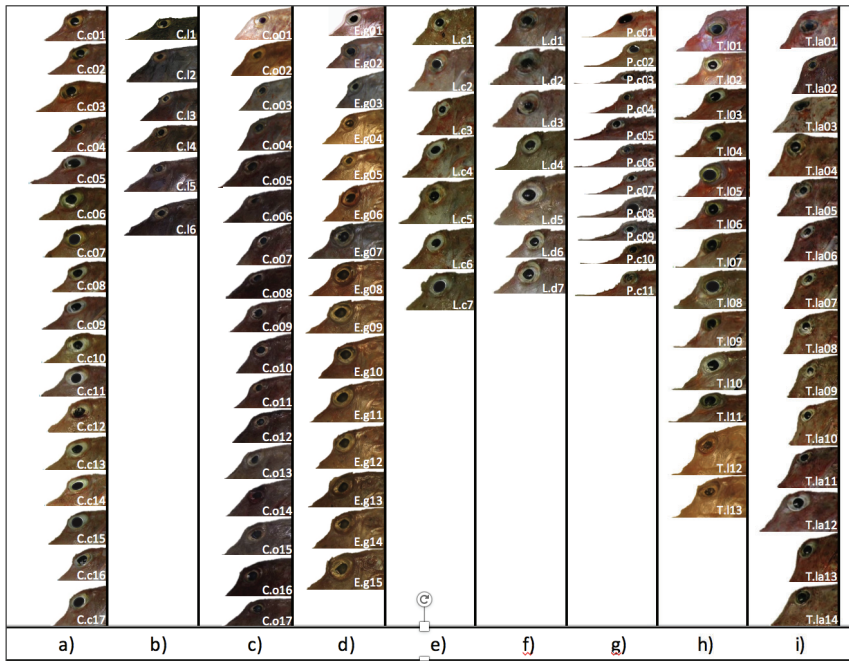
2.2. Data Used

Red Mullet specimens were collected on the trawl-fishing grounds along the island shelf and slope of Mallorca and Menorca during the 2002 survey of the BALAR project (2000–2004), which was integrated in the Spanish ‘Programa Nacional de Datos Básicos Pesqueros’ for the Balearic Islands. Trawls were carried out during daylight hours on board the oceanographic vessel R/V “Francisco de Paula Navarro” [43]. Gurnard specimens were collected on the trawl-fishing grounds along the Catalan Coast in commercial fishing boats of the ROSES and BOLINUS projects, both funded by the Catalan Government, during 2015 and 2017 [44].

Side-view images of the whole fish were taken in the field with all specimens in the same standardised position and orientation. The fish appear in the center of the image, with the head oriented to the left and the belly facing the bottom of the image. The specimens had different ages and lengths, and the distances from which the pictures were taken were not the same. An expert taxonomist added two points to each image, one to mark the tip of the snout and the other to mark the anterior insertion of the first dorsal fin. Those points indicate the beginning and the end of the contour segments used by taxonomists to identify the mullet’s species. Figure 4 shows, on the bottom, a specimen of each species and, at the top, the profile portion used by the expert to identify the species for all individuals in the database. For this study, we have a data set of 23 images for each species of Red Mullet. The left part is devoted to *M. barbatus* and the right part to *M. surmuletus*. Figure 5 shows the head contours of 108 specimens belonging to nine species of Gurnards.



**Figure 4.** (a) *Mullus barbatus* and (b) *Mullus surmuletus* specimens in the database. The images on the top show the profile of interest according to taxonomist criteria. The specimens are identified with the letter B/S plus a number.



**Figure 5.** The images show the profiles of interest of the specimens in the database according to taxonomist criteria. (a) *Chelidonichthys cuculus* (C.c.#), (b) *Chelidonichthys lucerna* (C.l.#), (c) *Chelidonichthys obscurus* (C.o.#), (d) *Eutrigla gurnardus* (E.g.#), (e) *Lepidotrigla cavillone* (L.c.#), (f) *Lepidotrigla dieuzeidei* (L.d.#), (g) *Peristedion cataphractum* (P.c.#), (h) *Trigla lyra* (T.l.#), and (i) *Triglorporus lastoviza* (T.l.#).

### 2.3. Open Contour Extraction and Normalization

As the fish were photographed following a certain protocol, they appear in the center of the image, in a horizontal position, with the dorsal fin at the top and the nose oriented to the left. To establish the fragment with relevant information, the expert introduced two marks in all images, one in the snout and the other in the starting point of the dorsal fin, indicating the limits of the contour fragment. These marks are in the background but close to the figure of the fish. Those images are our input data. At that point, the image processing is entirely conventional. Landmark detection can be done by colour. The landmark points are the references that make it possible to automatically crop the original images and thus isolate the profile of interest in a more straightforward and smaller image. In addition, the small size of the new images make the background color more uniform. Contour detection can also be established with a common algorithms presented in the *image processing toolbox* of MATLAB.

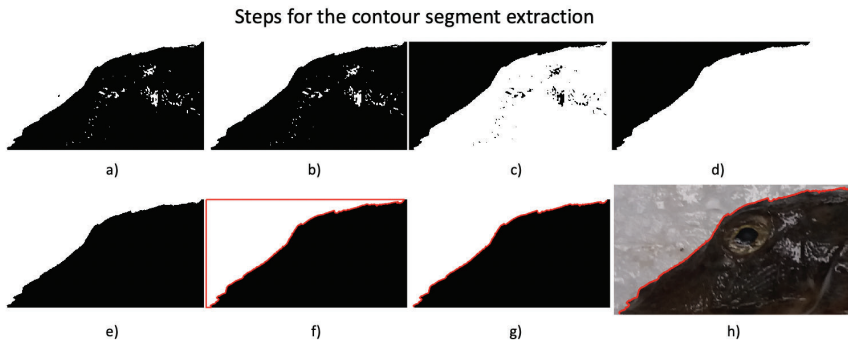
The first step is to convert the region of interest, the ROI, (a colour image) to a grey level in terms of luminance, and then binarize it. There are various methods that can be used to carry out a binarization, but they all depend on a threshold. In our particular case, we employed a global threshold computed using Otsu’s method, which chooses the threshold that minimizes the intraclass variance of black and white pixels [45]. There are more sophisticated methods that can be used to obtain the threshold, but this is the default option used by the function *imbinarize* in the MATLAB image processing toolbox.

In Figure 6, the sequential steps that are applied in order to obtain the contour segment of interest are shown, for a particular case; in Figure 6a the results of the binarization; in Figure 6b, the black holes in the white region are filled; in Figure 6c, the complement of the previous image; in Figure 6d,

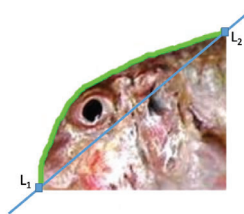


the black holes corresponding to the fish are filled; and in Figure 6e, the image is complemented once again to represent the fish in black and the background in white. In Figure 6f, the boundaries of the white region (in red) are found by exploring the image starting from the upper-left pixel and continuing in a downwards direction. In Figure 6g, the image borders are removed to find the contour of interest, which is the final result of the process. Finally, in the sub-image Figure 6h, the obtained result is shown overlapped against the original ROI.

As will be seen, the method developed for this work is very robust to the high-frequency noise that can appear in the determination of contours. Notice that the binary image in Figure 6g can be used as a segmentation template to remove the background, as done in the compositions of Figures 4 and 5. After completing this step, the contour of the specimen  $i$ ,  $C^i$ , is given by its two sequences of horizontal and vertical pixel positions in the cropped image. The origin was taken to be the first point of the open contour, which corresponds to the mark of the nose, and was assigned the (0,0) co-ordinate. The rest of the pixel positions are given relative to this first point and are grouped in the two series  $(h_k, v_k)^i$ . Each contour is represented by a different number of points  $N_i$ . Figure 7 shows a cropped image where the contour of interest is delimited by the landmarks  $L_1$  and  $L_2$ .



**Figure 6.** Steps for the contour segment extraction. In (a), binarization; in (b–e), the steps to fill the gaps in the background and in the fish regions; in (f), the boundary detection of the white region; and in (g), the contour of interest found after removing the image borders.



**Figure 7.** Fragment of interest of the fish delimited by landmarks  $L_1$  and  $L_2$ .

### 2.3.1. Open Contour Normalization

Before the development of the features can be addressed, the open contours must first be normalized. The normalization should be able to compare segments that are represented by a different number of points while also neutralising variations with respect to rotation and scale. Furthermore, care will be taken to ensure that the normalization respects the aspect ratio of the segments. In the case of open contours, the reader can see that the first,  $(h_0, v_0) = (0, 0)$ , and the last point,  $(h_{N_i-1}, v_{N_i-1})$ , of the open contour must necessarily differ and, therefore, the reconstruction done with few DFT coefficients could present strong distortions due to the Gibbs phenomenon. However, although the



elements  $x_0$  and  $x_{N_i-1}$  of the sequence differ significantly, the DCT can efficiently approximate the sequence  $x_n$  with  $\hat{x}_n$  using few coefficients ( $L$ ).

As previously demonstrated, the DFT works very well for closed contours because the natural shapes tend to have a smooth variation and, if sampled correctly, the point  $(h_0, v_0)$  is very similar to  $(h_{N_i-1}, v_{N_i-1})$ , with  $h_0$  and  $v_0$  being close to  $h_{N_i-1}$  and  $v_{N_i-1}$ , respectively. The normalization put forward is simple and takes these issues into account. It can be performed following these next two steps:

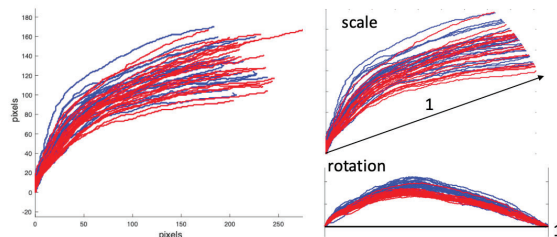
- First, for each open contour, we compute the Euclidean distance  $\alpha$  between  $(h_0, v_0)$  and  $(h_{N_i-1}, v_{N_i-1})$ , and the angle  $\theta$  that the segment defined by these two points forms with the horizontal. Seen graphically in Figure 7, the points  $(h_0, v_0)$  and  $(h_{N_i-1}, v_{N_i-1})$  correspond to the landmarks  $L_1$  and  $L_2$ , respectively. As all the segments begin at  $(h_0, v_0) = (0, 0)$ , we have:

$$\alpha = \sqrt{h_{N_i-1}^2 + v_{N_i-1}^2} \quad \text{and} \quad \theta = \tan^{-1}(v_{N_i-1}/h_{N_i-1}). \tag{8}$$

Then,  $\alpha$  and  $\theta$  are used to rotate and re-scale all the contours points according to:

$$\begin{bmatrix} p_k \\ q_k \end{bmatrix} = \frac{1}{\alpha} \begin{bmatrix} \cos(\theta) & -\sin(\theta) \\ \sin(\theta) & \cos(\theta) \end{bmatrix} \begin{bmatrix} v_k \\ h_k \end{bmatrix}. \tag{9}$$

After this operation, without changing the aspect ratio, all profiles have been re-scaled and rotated so that they start at point  $(0,0)$  and end at point  $(1,0)$ . Figure 8 separately shows both the scale and scale-and-rotation effects on the original Red Mullet contours. Note in Figure 8 that the relevant information to discriminate the contours is concentrated in  $q_k$  (the vertical axis after the rotation) while  $p_k$  (the horizontal axis after the rotation) will be very close for all contours and will always go from 0 to 1 in approximately constant increments.



**Figure 8.** Representation of the scale and scale-&-rotation effects on the original Red Mullet contours. These changes are performed during the contour normalization process.

- Second, in order to balance the number of points, the sequence of points  $q_k$  is re-sampled so that they all have 256 values. The re-sampling is performed by cubic splines in order to have values at equispaced intervals in the range going from 0 to 1. We call the re-sampled sequence  $x_k$ . Finally the contours are represented with a single sequence. In addition, the points  $x_0$  and  $x_{255}$  practically take the same value (0), which allows the DFT to be used as well.

#### 2.4. Features Development

Two types of features were developed to feed the classifiers. One is based on the DCT and the other on the DFT. Both take the normalized contour described in the previous section and apply either a DCT or a DFT. In the case of the DCT, we take the first  $L$  real coefficients as features  $F_i$ . In the case of the DFT, the coefficients are complex and the features are made of the first real parts followed by the

first imaginary parts. In that case, as the normalized contour takes real values, the first DFT coefficient is also real and therefore its imaginary part is always zero, and thus is excluded.

From the first  $N$  DFT complex coefficients  $L = 2N - 1$  features  $F_i, i \in [0, \dots, L - 1]$  are obtained. In addition, for all cases, the vectors of the features are standardised according to the *zscore* transformation. Thus, in the training step, the means  $m_i$  and standard deviations  $\sigma_i$  of feature  $F_i, i \in [0, \dots, L - 1]$  are computed. The employed features  $x_i$  will be  $x_i = (F_i - m_i)/\sigma_i$ . In the test, the averages  $m_i$  and standard deviations  $\sigma_i$  obtained in the training phase are used to form the features of the segment that is to be classified.

2.5. Extreme Learning Machines, Training and Classification

A single-hidden layer feed-forward neural (SLFN) network is trained as a classifier following the extreme learning machines (ELM) framework [46,47]. The main idea under ELM is that a SLFN network can be trained by randomly assigning both input weights and bias. Proceeding this way, the only parameters to determine in the SLFN network setting are the weights that connect the hidden layer neurons to the outlets (named output weights) and the number  $H$  of hidden neurons. This strategy changes the problem of computing the SLFN output weights to a linear problem. The resulting linear over-determined system can be quickly solved using a Moore–Penrose pseudo-inverse in one single step [46,47].

Focusing on our classification problem, each individual  $i$  in a set of  $N$  is characterized with  $L$  features organized in a vector  $\mathbf{x}_i = [x_0 \dots x_{L-1}]^T$ . The matrix  $\mathbf{X} = [\mathbf{x}_1 \dots \mathbf{x}_i \dots \mathbf{x}_N]$  of size  $L \times N$  organizes the features of all  $N$  specimens. These individuals are classified into  $C$  classes. For each class, there is a vector  $\mathbf{t}_c$  that indicates with either a 1 or a  $-1$  in the position  $i$  whether the individual  $i$  belongs (1) or does not belong ( $-1$ ) to the class  $c$ . The vectors  $\mathbf{t}_c$  are arranged in the matrix  $\mathbf{T} = [\mathbf{t}_1 \dots \mathbf{t}_c \dots \mathbf{t}_C]^T$  of size  $N \times C$  so that in each row of  $\mathbf{T}$  only one 1 appears. Considering that  $H$  is the number of hidden neurons, the input weight matrix  $\mathbf{W}$  will have a size of  $L \times H$  in which its element  $w_{lh}$  is the weight that connects the feature  $l$  with the hidden node  $h$ . As there are  $H$  hidden nodes, there is a vector of bias  $\mathbf{b}$  containing the bias  $b_h$  of each node.

Now, considering the individual  $i$ , the flow of information into the network can be written as  $\mathbf{x}_i^T \mathbf{W} + \mathbf{b}^T$ . Taking into account the activation function, which is typically a sigmoid, we have  $\text{sig}(\mathbf{x}_i^T \mathbf{W} + \mathbf{b}^T)$ , where  $\text{sig}(\mathbf{A})$  applies the sigmoid function to each  $a_{ij}$ . Then, when the resulting  $1 \times H$  row vector is multiplied by the output weight matrix  $\mathbf{B}$  of size  $H \times C$  the resulting  $1 \times C$  row vector will indicate the class to which the individual  $i$  belongs. If we group the same equation for all previously classified individuals, ideally we will have the equation:

$$\text{sig}(\mathbf{X}^T \mathbf{W} + \mathbf{u} \mathbf{b}^T) \mathbf{B} = \mathbf{T}, \tag{10}$$

where  $\mathbf{u}$  is the  $N \times 1$  all-ones vector We define the  $N \times H$  matrix  $\mathbf{H}$  as:

$$\mathbf{H} = \text{sig}(\mathbf{X}^T \mathbf{W} + \mathbf{u} \mathbf{b}^T). \tag{11}$$

Then, the ANN formulation can be compactly written as:

$$\mathbf{H} \mathbf{B} = \mathbf{T}. \tag{12}$$

Once  $H$  is decided,  $\mathbf{W}$  and  $\mathbf{b}$  are randomly selected from a Gaussian distribution function with a mean of zero and a standard deviation of one. The only unknown parameter is  $\mathbf{B}$ . Note that Equation (12) is over-determined. For ELM networks, it is common to choose a number  $H$  of hidden nodes lower than the number  $N$  of classified cases used for training, that is,  $N > H$ . Intuitively, if we see Equation (12) in terms of vectors  $\mathbf{b}_i$  and  $\mathbf{t}_i$  of  $\mathbf{B}$ , and  $\mathbf{T}$  respectively, we have  $\mathbf{H} \mathbf{b}_i = \mathbf{t}_i$ , for  $i$  going from 1 to  $C$ , meaning that for  $N$  equations, we have  $H$  unknown weights. The output weight matrix  $\mathbf{B}$ , provided that  $(\mathbf{H}^T \mathbf{H})^{-1}$  is non singular, is computed by:

$$\mathbf{B} = \left(\mathbf{H}^T \mathbf{H}\right)^{-1} \mathbf{H}^T \mathbf{T} = \mathbf{H}^\dagger \mathbf{T}, \quad (13)$$

where  $\mathbf{H}^\dagger$  is the Moore–Penrose inverse.

Once  $\mathbf{W}$ ,  $\mathbf{b}$ , and  $\mathbf{B}$  are defined, the SLFN can then be used to predict a new individual from its vector of characteristics  $\mathbf{x}$  by applying the following relationship:

$$cl = \arg \max_c \left( \text{sig} \left( \mathbf{x}^T \mathbf{W} + \mathbf{b}^T \right) \mathbf{B} \right) \quad (14)$$

where, given a vector  $\mathbf{a}^T$ , the function  $\arg \max_c (\mathbf{a}^T)$  finds the maximum value of  $\mathbf{a}^T$  and returns its position;  $cl \in [1, \dots, C]$  is the assigned class.

### 3. Results

#### 3.1. Features Based on the Dct and the Dft. a Comparative Study

##### 3.1.1. Leave-One-Out Cross-Validation

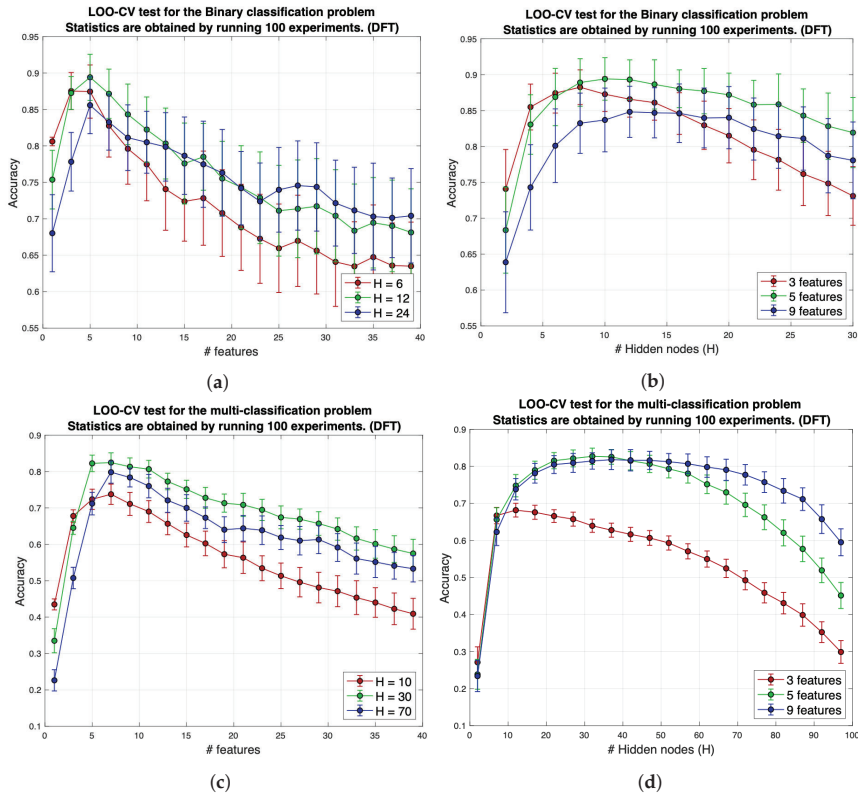
A significant limitation we deal with in this work is the limited number of classified images available for each species used for the training. In the binary classification problem, the two classes are balanced, with 23 individuals in each group. However, in the multi-classification problem, this limitation is more severe, as we have to deal with an even more reduced set of data, which is also unbalanced for the nine groups that need to be classified, meaning that there are different numbers of individuals in each of these groups, as shown in Figure 5.

Despite having these nine classes unbalanced and some of them badly under-represented, the same methodology has been applied in both problems. The *leave-one-out cross-validation* (LOO-CV) strategy is commonly used to evaluate the performance of the classifier when dealing with small data-sets [48]. This strategy uses a single observation taken from the original data-set to validate the data, while the remaining observations are used for the training phase. This way, the model is built with all the labeled elements except one, which is used for validation. Once the element has been tested, it is added to the group again, and the next element is extracted, repeating the process until all elements have been used for the test. Thus, all the elements are tested against the rest of the known classified elements. The advantage of this method is that all observations are used for both training and validation, and each observation is used for validation exactly once.

##### 3.1.2. LOO-CV Tests

The following tests are intended to determine both the size of the network and the optimal number of characteristics that achieve the best classification rate. The tests that were carried out using features coming from the DFT were then repeated; however, this time using features coming from the DCT, and their results are summarized in Figures 9 and 10.

In all these tests, LOO-CV was used to obtain realistic results, and the experiments were repeated 100 times in order to show existing tendencies. Mean values of accuracy, the quotient between the number of correct predictions divided by the number of the predictions done, are represented by color balls and their corresponding standard deviations are shown by vertical lines. The SLFN weights,  $\mathbf{W}$  and  $\mathbf{b}$ , were randomly selected according to a zero-mean Gaussian distribution with a standard deviation of one, where  $H$  indicates the number of hidden nodes and  $L$  the number of input features. In Figures 9 and 10,  $N$  number of transformed coefficients were used. In these figures, note that in the case of the DCT, the number of features ( $N$ ) is the same as the number of coefficients ( $L$ ), but in the DFT,  $L = 2N - 1$  instead.



**Figure 9.** The accuracy obtained by training a single-hidden layer feed-forward neural (SLFN) with the DFT-based features. Leave-one-out cross-validation (LOO-CV).

Figures 9 and 10 share the same structure. In both, the Figures 9a,b and 10a,b show the results of the binary classification problem and Figures 9c,d and 10c,d portray the results of the multi-class classification problem. When it comes to the binary problem, the optimal number of DFT-based features is  $L = 5$ , quite independently of the size of the network (see Figure 9a). The best accuracy values were obtained for  $H$  values of around 10 (see Figure 9b).

The best results achieved using the DCT-based features were also obtained for when  $L = 5$  independently of the size of the network (see Figure 10a). In the same way, the best accuracy results were obtained when  $H = 10$  (see Figure 10b). In both cases, the best average value for the accuracy was around 0.9 (90%). To obtain the first five features, the first three and five DFT and DCT coefficients were required, respectively. The two encoding alternatives behave in a very similar way, and to decide which one is the best one, the tests need to be refined. Regarding the multi-class classification problem, when using DFT-based features, the best results were attained when  $L = 5$ .

In this context, the optimum network size increases to values of about  $H = 30$  (see Figure 9c). Above  $H = 30$ , there is a wide range of  $H$  values for which networks work properly (see Figure 9d). When the DCT is used in this context, the results are very similar, and so from  $L = 5$  (see Figure 10c) and from  $H = 30$  (see Figure 10d) there is a wide range of values of  $L$  and  $H$  for which the networks work correctly. In both cases, independently of the use of DFT/DCT, the average accuracy values are also similar and in the better configurations appear to be slightly above 80%.

From the previous graphs, it is difficult to determine which coding (DFT/DCT) provides the best accuracy, since they behave in a very similar way. In the tests performed below, the number of

iterations is incremented to 1000 for a small range of values of  $H$  and  $L$  near the optimal point of network operation. LOO-CV is always performed. The results are presented in tables indicating the accuracy regarding the average value, standard deviation, and both the maximum and minimum values obtained. The optimal values are represented in bold.

We found that the optimum number of features was 5, independently of which transform (DCT or DFT) was used. Using more than five features reduces the performance of the networks. The only case that differs from the others in the sense of optimal number of features used is that of the binary classification case if the DFT is used, where the best accuracy scores were obtained with three features.

Table 1 shows, for the binary classification problem, a comparative summary of the network accuracy achieved for the best combinations of the DFT (three features) and the DCT (five features) depending on the hidden nodes ( $H$ ), and Table 2 shows the same information but for the multi-class classification problem, for which the best combination of features is 5 in both the DCT and DFT feature configurations. Table 2 reveals that the best DFT-based configuration exceeds the best DCT-based one by almost two percentage points.

In Table 1, however, it can be seen that the difference between using features derived from the DCT or the DFT is minimal. Regarding the case of the binary classification problem, no combination of features based on the DCT and the DFT was found that improved the results shown in Table 1. In the multi-class classification problem, the aggregation of the five features of each of the best options shown in Table 2 achieved an improvement in the results by almost one percentage point as shown in Table 3.

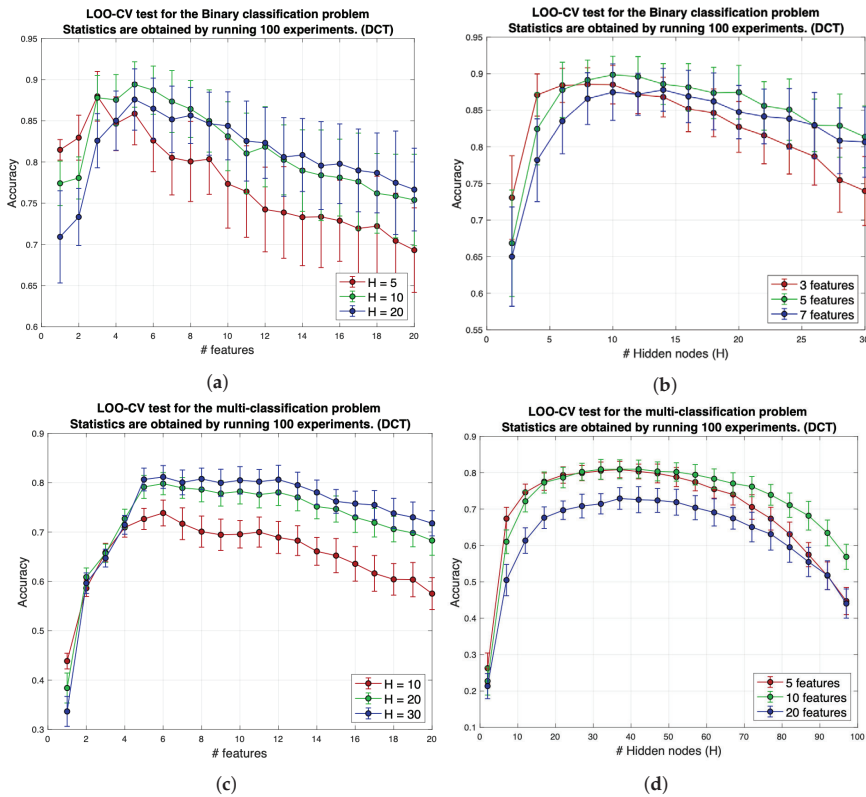


Figure 10. The accuracy obtained by training SLFN with the DCT-based features.

**Table 1.** The binary classification problem. A comparative summary of the network accuracy obtained for the best combinations of the DCT (five features) and the DFT (three features) depending on the hidden nodes used. The LOO-CV was performed and the experiments were repeated 1000 times. For each feature set, there is an optimum number of hidden nodes that provide the best results. In the table the bests accuracy results are highlighted in bold.

Binary-Class Classification Accuracy (LOO-CV 1000 Iterations)								
Hidden Nodes	DCT-5				DFT-3			
	mean	std	max	min	mean	std	max	min
7	88.44	0.031	95.65	78.26	88.46	0.032	95.65	76.09
8	89.09	0.031	95.65	78.26	89.09	0.030	95.65	78.26
9	89.37	0.030	97.83	80.44	89.29	0.031	95.65	76.09
10	89.30	0.031	95.65	78.26	<b>89.36</b>	<b>0.028</b>	<b>95.65</b>	<b>80.44</b>
11	<b>89.34</b>	<b>0.029</b>	<b>95.65</b>	<b>80.44</b>	89.19	0.030	95.65	78.26
12	89.33	0.031	95.65	80.44	89.20	0.030	95.65	78.26
13	89.04	0.031	95.65	78.26	89.10	0.030	95.65	80.44
14	88.76	0.032	95.65	78.26	88.64	0.032	95.65	78.26
15	88.80	0.031	95.65	78.26	88.57	0.031	95.65	76.09
16	88.40	0.032	95.65	78.26	88.14	0.033	95.65	73.91
17	88.05	0.033	95.65	78.26	87.92	0.033	95.65	76.09

**Table 2.** The multi-class classification problem. Comparative summary of the network accuracy obtained for the best combinations of the DCT (five features) and the DFT (five features) depending on the hidden nodes used. For each feature set, there is an optimum number of hidden nodes providing the best results. In the table the bests accuracy results are highlighted in bold. The LOO-CV was performed and the experiments were repeated 1000 times.

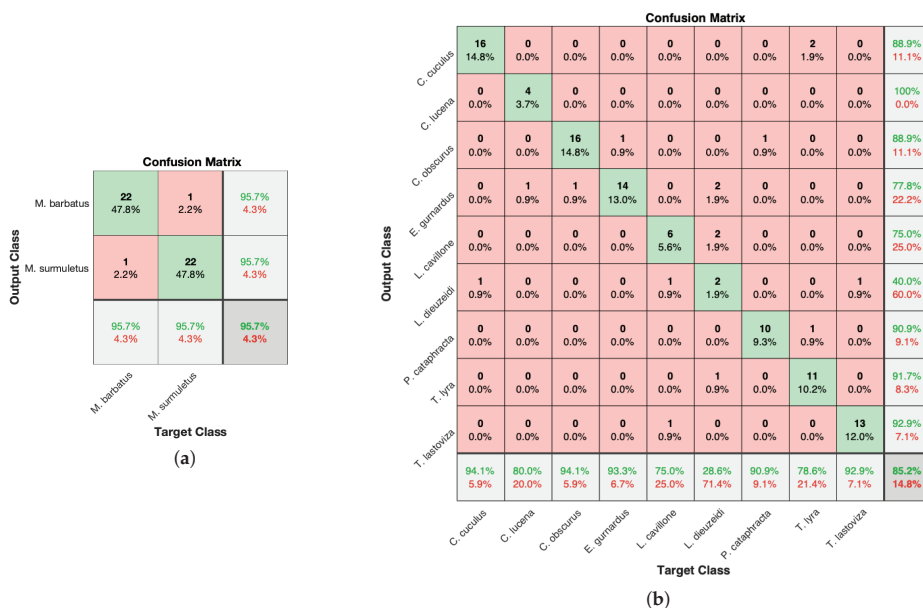
Multi-Class Classification Accuracy (LOO-CV 1000 Iterations)								
Hidden Nodes	DCT-5				DFT-5			
	mean	std	max	min	mean	std	max	min
5	72.66	0.024	79.63	64.82	56.78	0.035	66.67	44.44
10	72.66	0.024	79.63	64.82	72.29	0.029	81.48	63.89
15	76.66	0.023	84.26	68.52	77.78	0.024	85.19	69.44
20	78.62	0.022	85.19	72.22	80.52	0.024	88.89	72.22
25	79.84	0.023	87.04	72.22	82.04	0.021	87.96	74.07
30	80.42	0.023	87.04	72.22	<b>82.59</b>	<b>0.021</b>	<b>88.89</b>	<b>75.00</b>
35	<b>80.64</b>	<b>0.022</b>	<b>87.04</b>	<b>73.19</b>	82.49	0.022	87.96	75.00
40	80.54	0.023	87.96	73.19	81.94	0.022	88.89	74.07
45	80.00	0.024	87.04	72.22	81.18	0.029	87.04	74.07
50	79.38	0.025	87.04	71.30	79.79	0.024	86.11	72.22
55	78.18	0.025	87.04	69.44	78.15	0.025	85.19	68.52
60	76.71	0.026	85.18	68.52	76.26	0.027	85.19	68.52

Figure 11 shows the confusion matrices and realizations for both the binary (a) and the multi-class classification (b) problems, respectively. The results were obtained by performing LOO-CV. In both sub-figures, a confusion matrix is shown with a resulting accuracy slightly above the corresponding mean accuracy value. From Figure 11b the correct classification rates of the different classes were acceptably high compared to the ones obtained for the class of *Lepidotrigla dieuzeidei*.

As explained in this paper, the two points introduced by the specialist were used to segment the region of interest. The introduction of these points unambiguously determines the contour segment used to make the decision. However, this procedure can also be performed automatically. To show that, we conducted the following experiment. We took a set of 50 images, and we randomly took 40 of them to train a faster region-based convolutional network (Faster R-CNN) for object detection, leaving the rest of the images for the test. With this small amount of images, we were able to train a robust model that worked acceptably well. The procedure for training the model was as follows.

**Table 3.** The multi-class classification problem. Summary of the network accuracy obtained by using 10 features (five coming from the DCT and five from the DFT) depending on the hidden nodes used. The best results, highlighted in bold, are obtained when the net has 34 hidden nodes. The LOO-CV was performed and the experiments were repeated 1000 times.

Multi-Class Classification Accuracy (LOO-CV 1000 Iterations)				
Hidden nodes	DCT-5 mean	DFT-5 std	max	min
20	80.70	0.024	87.96	73.15
21	81.50	0.023	87.96	75.00
24	82.08	0.022	88.89	75.00
26	82.48	0.022	88.89	75.93
28	82.89	0.022	89.82	75.00
30	83.15	0.022	90.74	75.93
32	83.41	0.023	91.67	75.00
34	<b>83.53</b>	<b>0.022</b>	<b>89.82</b>	<b>76.85</b>
36	83.40	0.023	89.82	75.00
38	83.34	0.023	89.82	75.00
40	83.48	0.023	89.82	75.93
42	83.20	0.022	89.82	76.85
44	83.20	0.024	89.82	75.00
46	82.82	0.023	89.82	74.07
48	82.57	0.023	89.82	75.00
50	82.11	0.024	88.89	75.00



**Figure 11.** (a) The confusion matrix of binary-classification realization with an accuracy value of 95.7%. The results were obtained by performing LOO-CV using five features based on the DFT with a SLFN of 10 hidden nodes. The mean accuracy value after performing 1000 LOO-CV iterations was 89.3%. (b) The confusion matrix of multi-class classification realization with 85.2% accuracy. The results were obtained by performing LOO-CV using five features based on the DCT plus five more based on the DFT, with an SLFN of 30 hidden nodes. The mean accuracy value after performing 1000 LOO-CV iterations maintaining the same SLFN size was 83.15%.



The first step was to reduce the images to a size of  $384 \times 512$  pixels. For training, we used the 40 images without marks, and we indicated the region of interest found with the specialist’s marks. This region of interest is a rectangle that is specified with the two extreme points on its main diagonal. We employed the Faster R-CNN object detector provided by MATLAB [49,50]. A Faster R-CNN object detection network is constituted of a feature extraction network followed by two subnetworks.

The feature extraction network is usually a pre-trained CNN. The first subnetwork that supports the feature extraction network is a region proposal network (RPN) devoted to finding the areas in the image where objects are likely to exist. Those areas are known as object proposals. The second subnetwork is dedicated to predicting the class of each object proposal. In the case we are dealing with, for the first network, instead of using a pre-trained CNN as, for instance, the ResNet-50 or the Inception v3, we used a much simpler CNN that has only three convolutional layers from a total of fifteen, that was trained for another problem.

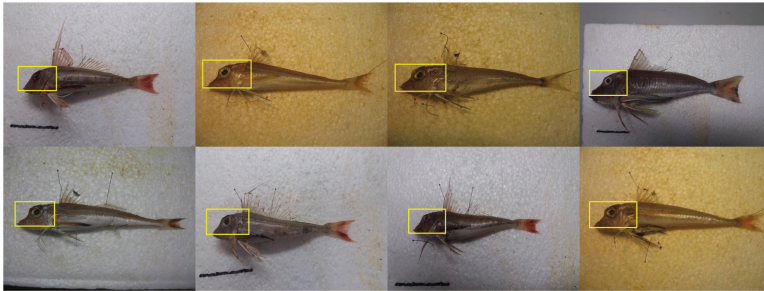
The main characteristics of the structure are summarized in Table 4. The main training options chosen for the Faster R-CNN are ‘MaxEpochs’, ‘MiniBatchSize’ and ‘InitialLearnRate’, which were set to 12, 8, and  $5 \times 10^{-4}$  with a ‘NegativeOverlapRange’ of [0 0.3] and a ‘PositiveOverlapRange’ of [0.6 1]. From these parameters, we note that the ‘MaxEpochs’ specifies the number of epochs, that is, the number of full passes of the training algorithm over the entire training set, and the learning rate that controls the speed of training. The training was performed using the CPU of a MacBook Pro with eight-core Intel i9 processor at 2.4 GHz and 64 GB of RAM running the MATLAB environment on the macOSCatalina 10.15.3 operating system.

**Table 4.** Network layers that compose the feature extraction network. The convolution put the input images through a set of convolutional filters, each of which activates certain features from the images. The pooling layer simplifies the output by performing nonlinear downsampling, reducing the number of parameters that the network needs to learn. The rectified linear unit (ReLU) allows for faster and more effective training by mapping negative values to zero and maintaining positive values. The fully connected layer “flattens” the network’s 2D spatial features into a 1D vector that represents, and, finally, the Softmax provides probabilities for each category in the dataset image-level features for classification purposes.

Feature Extraction Network		
Layer	Type	Main Characteristics
1	Image Input	$32 \times 32 \times 3$ images with ‘zerocenter’ normalization
2	Convolution	$32 \times 5 \times 5 \times 3$ convolutions with stride [1 1] and padding [2 2 2 2]
3	Max Pooling	$3 \times 3$ max pooling with stride [2 2] and padding [0 0 0 0]
4	ReLU	ReLU
5	Convolution	$32 \times 5 \times 5 \times 32$ convolutions with stride [1 1] and padding [2 2 2 2]
6	ReLU	ReLU
7	Average Pooling	$3 \times 3$ average pooling with stride [2 2] and padding [0 0 0 0]
8	Convolution	$64 \times 5 \times 5 \times 32$ convolutions with stride [1 1] and padding [2 2 2 2]
9	ReLU	ReLU
10	Average Pooling	$3 \times 3$ average pooling with stride [2 2] and padding [0 0 0 0]
11	Fully Connected	64 fully connected layer
12	ReLU	ReLU
13	Fully Connected	2 fully connected layer
14	Softmax	softmax
15	Classification Output	crossentropyex

Training lasted approximately 3 min. Due to the reasonably fast training time, many random tests on the whole set of images could be performed, and thus the solidity of the approach could be checked. The FasterRCNN object detection network output provides a rectangle by giving two points corresponding to the endpoints of one of its diagonals (specifically the one that goes from the top right to the lower-left point of the rectangle). Some ROI detection results are shown in Figure 12.

These points can be used to crop the original image and obtain the ROI, which is practically the same provided by the specialist (although the specialist selects the other diagonal because its endpoints are more near to the segment of interest).



**Figure 12.** Example of the detection results performed on the test images using the trained faster region-based convolutional network (Faster R-CNN) object detector. The obtained ROI is shown in yellow.

### 3.2. Regarding Automation of the Whole Process

A step that is required to be carried out if the ROI detection is performed without the support of the two points provided by the expert is to re-scale the original image to the FasterRCNN object detection input size, and then, once the ROI is found, it must be mapped onto the original image. This allows the extraction of the ROI from the original image. The following step is to obtain the contour, compute the features, and to introduce them in the ELM network to identify the class of the specimen. The diagram presented in Figure 13 can help to clarify the whole process, which can be described in four stages. The first stage is the ROI detection. This procedure can be supervised or it can be carried out automatically using a Faster-RCNN object detector.

Once the ROI has been determined, the second stage consists of extracting the fish contour segment from this region. The result of stage two is the horizontal and vertical pixel coordinates that describe it in relation to the first contour point, which is taken to be the origin of the coordinate system. The third stage is the 'feature extraction', in which this sequence of contour points obtained from the previous stage is used as input. These points are standardized by applying a rotation, a re-scaling and a re-sampling by using third-order splines. The result is a single sequence ( $q_k$ ) of 256 values. Then, in the same stage, DCT and DFT are taken on  $q_k$ , and we select the first five DCT and DFT coefficients to form the vector of features. In the multi-class classification problem, a features vector containing only 10 real numbers produced the best accuracy results. The vector of features was used in the last stage to add to an already trained ELM of 34 hidden nodes. Finally, the ELM output is the species that the individual belongs to.

Figure 14 shows the details of a feature extraction stage and of a classification stage corresponding to a model that solves the binary classification problem. It is an SLFN network of 10 hidden nodes trained with features based on the DFT. Figure 14a–c graphically show the steps required to obtain the sequence of 256 points in vector  $\mathbf{q}$  that represent the contour after the standardization, and Figure 14a shows the way to obtain the vector  $\mathbf{x}$  of features. Note that  $\mathbf{F}$  is the Fourier matrix and  $\mathbf{f}$ , the DFT of  $\mathbf{q}$ . In this case, the vector of features  $\mathbf{x}$  has five real elements obtained from the first three Fourier coefficients, as  $f_0 \in \mathbb{R}$ . Figure 14e shows the classification stage, which begins with a Z-score standardization, being the vectors  $\mathbf{m}$  and  $\boldsymbol{\sigma}$ , the means and standard deviations of  $\mathbf{x}$ . The normalized features after the Z-score are the inputs of the SLFN network.  $\mathbf{W}$ ,  $\mathbf{b}$  and  $\mathbf{B}$  are the input weights, the bias, and the output weights of the trained SLFN net, respectively. Finally, the decision is taken on the values achieved in the output nodes. This simple architecture reached accuracy values up to 95.65% and, in 1000 LOO-CV experiments a mean accuracy of 89.36% (see the column *DFT-3* of Table 1).

Note that, in this case, the fish appears in the images following a standardization, and the problem is not overly complicated. In a random disposition of the objects of interest in the image, one would have to train the model with many more images, using a data augmentation strategy and likely a more complex feature extraction network. Additionally, the substitution of Faster-RCNN by Mask-RCNN [51,52] could be explored to see if the contour segment of interest could even be detected automatically in the ROI directly.

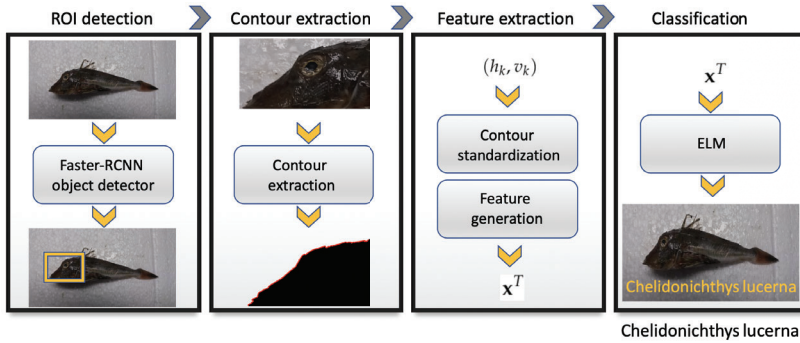


Figure 13. Simplified block diagram of the process chain when the ROI detection is done automatically.

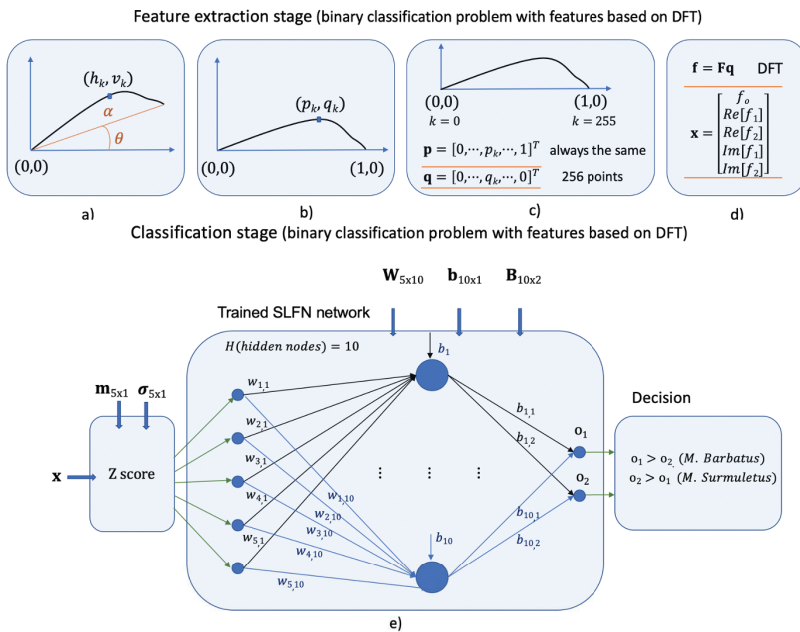


Figure 14. Detail at the parameter level of the feature extraction and classification stages of the binary classification solution working with features based on DFT. (a–c) show the steps required to obtain the vector  $\mathbf{q}$  from the contour, and (d) shows the way to obtain the vector  $\mathbf{x}$  from  $\mathbf{q}$ . In (e), the classification stage is shown with the parameters that interact in each step. The matrices and vectors show their dimensions in a subscript (in order to check the coherence of the operations involved). In the classification stage, three parts, the Z-score normalization, an LSFN network of 10 hidden nodes, and the decision rule, are employed.

#### 4. Conclusions

In this work, a method was presented that reproduced the procedure that experts perform manually to separate species of fish that exhibit a very similar shape. The technique consists of classifying a fragment from the profile of fish contours. Thus, a way to parameterise open contours was developed based on both the DFT and the DCT transformations, which already allow the discrimination of contour fragments with the use of only five parameters.

The presented approach for the *Mullus* binary classification case can be compared to the seminal work of Reference [10], which based on handmade measurements of the angle of head contour obtained an 88% correct distinction between both species. The mean accuracy values obtained (89.4%) in this work slightly improved those results. In addition to the difficulty of discriminating between very similar forms, we also had the limitation of having small, unreliably labeled image data sets.

For the second case considered in this work, the multi-class classification problem, it is important to note that there have been no handmade studies carried out that would allow for the comparison of results. The specimens of the database were cataloged at the ICM (Institut de Ciències del Mar) in Barcelona while different types of studies were being conducted. Given the complexity of the problem, the average classification results were close to 80%, which is in line with what trained professionals obtain when they take on the identification and classification task from photographs. If the classification is considered by classes, we observed that the class of *Lepidotrigla dieuzeidei* was very difficult to classify and thus caused the global accuracy to decrease; however, for the rest of the classes, high correct classification rates were reached.

Our goal was to prove that it is possible to automatically achieve results similar to those obtained by specialists when it comes to classifying very similarly-shaped species of fish. The goal achieved was to automatically classify the images of the specimens in the database into their species. Given the physical similarity between species and the size of the database, this is a difficult objective, which (as we mentioned before) we have overcome with success comparable to that obtained by the expert. The results suggest that we should expect even better results when working with species with more differentiated forms.

For future work, a far more ambitious goal involving government administrations, which must be addressed incrementally, is to computerize the analysis of fish boxes sold at fish markets. In these boxes, fish are stacked in a non-orderly fashion, with the body parts partially hidden but usually with their heads being visible. Thus, the classification of specimens based only on their head contour seems to be indeed an important improvement to be made.

These studies are important as fish markets are beginning to develop projects with the aim of photographing all fish boxes sold, followed by these images being then registered containing information on the labeling, sale prices, and so forth. These large amounts of images can only be processed automatically. As long as the techniques for extracting information from images continues improving, we will be able to obtain a gradually deeper understanding on the exploitation of marine resources. This work is a pioneer in the classification of biological forms through contour segments.

In this paper, the identification of the contour fragments was made with the help of marks. In this way, we were sure to deal with the fragment used by the experts. In the type of images used, the marks were located in easily identifiable positions and, even though this was not the purpose of this work, it was even shown in a prior subsection that these are not entirely necessary. Although this technique has been developed for a certain type of image, the capacity of representing contour segments between landmarks with very few parameters could be interesting and further explored in different contexts and applications.

The human visual system is particularly sensitive to the contour of a shape, and curves constitute an essential information element when it comes to image analysis, classification and pattern recognition, which are vital tasks that humans carry out on a daily basis. The necessity to quantify this visual information continues in present challenges, as was the case for this work, in which the relevant information that led to a good classification rate was concentrated only on a specific fragment of the

silhouette instead of the whole contour. In this work, we have the additional restriction of having little data available to use for training purposes, which imposes a classic machine learning work-flow, typically seen when dealing with small datasets.

In spite of that, this study suggested that the combination of machine learning (ML) and deep learning (DL) techniques can provide compelling solutions. On the one hand, through ML, we can (1) handle small datasets and (2) maintain the understanding of the problem. At the same time, DL reached milestones in image processing that were previously difficult to obtain, such as the identification of the ROI where the contour fragment is located. Finally, the ELM-based classifiers provided excellent results at low computational and time costs, which allowed for complete statistical tests.

**Author Contributions:** Conceptualization, P.M.-P. and A.L.; methodology, P.M.-P., A.M., and A.L.; software, P.M.-P. and A.M.; validation, P.M.-P., A.M., and A.L.; formal analysis, P.M.-P., and A.L.; investigation, P.M.-P. and A.L.; resources, A.L.; writing, original draft preparation, P.M.-P. and A.L.; writing, review and editing, P.M.-P., A.M., and A.L.; supervision, P.M.-P. and A.L.; funding acquisition, A.L. All authors read and agreed to the published version of the manuscript.

**Funding:** This work has been partially supported by the Spanish Government projects PHENOFISH with references: CTM2015-69126-C2-2-R and Catalan Government project PESCAT, with reference ARP029/18/00003.

**Acknowledgments:** We thank Marta Blanco and Marc Balcells from SAP-ICATMAR for images of commercial fish that were made available to us.

**Conflicts of Interest:** The authors declare no conflict of interest.

## References

1. Azzurro, E.; Tuset, V.M.; Lombarte, A.; Maynou, F.; Simberloff, D.; Rodríguez-Pérez, A.; Solé, R.V. External morphology explains the success of biological invasions. *Ecol. Lett.* **2014**, *17*, 1455–1463.
2. Stergiou, K.; Petrakis, G.; Papaconstantinou, C. The Mullidae (*Mullus barbatus*, *M. surmuletus*) fishery in Greek waters, 1964–1986. In *FAO Fisheries Report (FAO)*; FAO: Rome, Italy, 1992.
3. Renones, O.; Massuti, E.; Morales-Nin, B. Life history of the red mullet *Mullus surmuletus* from the bottom-trawl fishery off the Island of Majorca (north-west Mediterranean). *Mar. Biol.* **1995**, *123*, 411–419.
4. Bautista-Vega, A.; Letourneur, Y.; Harmelin-Vivien, M.; Salen-Picard, C. Difference in diet and size-related trophic level in two sympatric fish species, the red mullets *Mullus barbatus* and *Mullus surmuletus*, in the Gulf of Lions (north-west Mediterranean Sea). *J. Fish Biol.* **2008**, *73*, 2402–2420.
5. Cresson, P.; Bouchoucha, M.; Miralles, F.; Elleboode, R.; Mahe, K.; Maruszcak, N.; Thebault, H.; Cossa, D. Are red mullet efficient as bio-indicators of mercury contamination? A case study from the French Mediterranean. *Mar. Pollut. Bull.* **2015**, *91*, 191–199.
6. Golani, D.; Galil, B. Trophic relationships of colonizing and indigenous goatfishes (Mullidae) in the eastern Mediterranean with special emphasis on decapod crustaceans. *Hydrobiologia* **1991**, *218*, 27–33.
7. Labropoulou, M.; Eleftheriou, A. The foraging ecology of two pairs of congeneric demersal fish species: importance of morphological characteristics in prey selection. *J. Fish Biol.* **1997**, *50*, 324–340.
8. Lombarte, A.; Recasens, L.; González, M.; de Sola, L.G. Spatial segregation of two species of Mullidae (*Mullus surmuletus* and *M. barbatus*) in relation to habitat. *Mar. Ecol. Prog. Ser.* **2000**, *206*, 239–249.
9. Maravelias, C.D.; Tsitsika, E.V.; Papaconstantinou, C. Environmental influences on the spatial distribution of European hake (*Merluccius merluccius*) and red mullet (*Mullus barbatus*) in the Mediterranean. *Ecol. Res.* **2007**, *22*, 678–685.
10. Bougis, P. *Recherches Biométriques sur les Rougets, 'Mullus barbatus' L. et 'Mullus surmuletus' L...*; Centre National de la Recherche Scientifique: Paris, France, 1952.
11. Tortonese, E.; di Entomologia, A.N.I.; Italiana, U.Z. *Osteichthyes (Pesci Ossei): Parte Seconda*; Edizioni Calderini Bologna: Bologna, France, 1975.
12. Lombarte, A.; Aguirre, H. Quantitative differences in the chemoreceptor systems in the barbels of two species of Mullidae (*Mullus surmuletus* and *M. barbatus*) with different bottom habitats. *Mar. Ecol. Prog. Ser.* **1997**, *150*, 57–64.

13. Aguirre, H. Presence of dentition in the premaxilla of juvenile *Mullus barbatus* and *M. surmuletus*. *J. Fish Biol.* **1997**, *51*, 1186–1191.
14. Aguirre, H.; Lombarte, A. Ecomorphological comparisons of sagittae in *Mullus barbatus* and *M. surmuletus*. *J. Fish Biol.* **1999**, *55*, 105–114.
15. Hureau, J.; Bauchot, M.; Nielsen, J.; Tortonese, E. *Fishes of the North-Eastern Atlantic and the Mediterranean*; Unesco: Paris, France, 1986; Volume 3.
16. Kuhl, F.P.; Giardina, C.R. Elliptic Fourier features of a closed contour. *Comput. Graph. Image Process.* **1982**, *18*, 236–258.
17. El-ghazal, A.; Basir, O.; Belkasim, S. Farthest point distance: A new shape signature for Fourier descriptors. *Signal Process. Image Commun.* **2009**, *24*, 572–586.
18. Persoon, E.; Fu, K.S. Shape discrimination using Fourier descriptors. *IEEE Trans. Syst. Man Cybern.* **1977**, *7*, 170–179.
19. Tracey, S.R.; Lyle, J.M.; Duhamel, G. Application of elliptical Fourier analysis of otolith form as a tool for stock identification. *Fish. Res.* **2006**, *77*, 138–147.
20. Lestrel, P.E. *Fourier Descriptors and Their Applications in Biology*; Cambridge University Press: Cambridge, UK, 2008.
21. Mokhtarian, F.; Abbasi, S.; Kittler, J. Robust and Efficient Shape Indexing through Curvature Scale Space. In Proceedings of the 1996 British Machine and Vision Conference BMVC, Edinburgh, UK, 9–12 September 1996; Volume 96.
22. Dudek, G.; Tsotsos, J.K. Shape representation and recognition from multiscale curvature. *Comput. Vis. Image Underst.* **1997**, *68*, 170–189.
23. Mokhtarian, F.; Bober, M. *Curvature Scale Space Representation: Theory, Applications, and MPEG-7 Standardization*; Springer Science & Business Media: Berlin/Heidelberg, Germany, 2013; Volume 25.
24. Parisi-Baradad, V.; Lombarte, A.; García-Ladona, E.; Cabestany, J.; Piera, J.; Chic, O. Otolith shape contour analysis using affine transformation invariant wavelet transforms and curvature scale space representation. *Mar. Freshw. Res.* **2005**, *56*, 795–804.
25. Gibbs, J.W. Fourier's series. *Nature* **1899**, *59*, 606.
26. Toubin, M.; Dumont, C.; Verrecchia, E.P.; Lalignant, O.; Diou, A.; Truchetet, F.; Abidi, M. Multi-scale analysis of shell growth increments using wavelet transform. *Comput. Geosci.* **1999**, *25*, 877–885.
27. Allen, E.G. New approaches to Fourier analysis of ammonoid sutures and other complex, open curves. *Paleobiology* **2006**, *32*, 299–315.
28. Yang, C.; Tiebe, O.; Shirahama, K.; Łukasik, E.; Grzegorzec, M. Evaluating contour segment descriptors. *Mach. Vis. Appl.* **2017**, *28*, 373–391.
29. Dommergues, C.H.; Dommergues, J.L.; Verrecchia, E.P. The discrete cosine transform, a Fourier-related method for morphometric analysis of open contours. *Math. Geol.* **2007**, *39*, 749–763.
30. Wilczek, J.; Monna, F.; Barral, P.; Burlet, L.; Chateau, C.; Navarro, N. Morphometrics of Second Iron Age ceramics—strengths, weaknesses, and comparison with traditional typology. *J. Archaeol. Sci.* **2014**, *50*, 39–50.
31. Stefanini, M.I.; Carmona, P.M.; Iglesias, P.P.; Soto, E.M.; Soto, I.M. Differential Rates of Male Genital Evolution in Sibling Species of *Drosophila*. *Evol. Biol.* **2018**, *45*, 211–222.
32. Ahmed, N.; Natarajan, T.; Rao, K.R. Discrete cosine transform. *IEEE Trans. Comput.* **1974**, *100*, 90–93.
33. Kou, W.; Mark, J.W. A new look at DCT-type transforms. *IEEE Trans. Acoust. Speech Signal Process.* **1989**, *37*, 1899–1908.
34. Martucci, S. Symmetric convolution and the discrete sine and cosine transforms. *IEEE Trans. Signal Process.* **1994**, *42*, 1038–1051.
35. Strang, G. The discrete cosine transform. *SIAM Rev.* **1999**, *41*, 135–147.
36. Suresh, K.; Sreenivas, T. Linear filtering in DCT IV/DST IV and MDCT/MDST domain. *Signal Process.* **2009**, *89*, 1081–1089.
37. Britanak, V.; Yip, P.C.; Rao, K.R. *Discrete Cosine and Sine Transforms: General Properties, Fast Algorithms and Integer Approximations*; Academic Press: Cambridge, MA, USA, 2010.
38. Tsitsas, N.L. On block matrices associated with discrete trigonometric transforms and their use in the theory of wave propagation. *J. Comput. Math.* **2010**, *28*, 864–878.
39. Ito, I.; Kiya, H. A computing method for linear convolution in the DCT domain. In Proceedings of the 2011 19th European Signal Processing Conference, Barcelona, Spain, 29 August–2 September 2011; pp. 323–327.



40. Rao, K.R.; Yip, P. *Discrete Cosine Transform: Algorithms, Advantages, Applications*; Academic Press: Cambridge, MA, USA, 2014.
41. Wang, Z. Fast algorithms for the discrete W transform and for the discrete Fourier transform. *IEEE Trans. Acoust. Speech Signal Process.* **1984**, *32*, 803–816.
42. Roma, N.; Sousa, L. A tutorial overview on the properties of the discrete cosine transform for encoded image and video processing. *Signal Process.* **2011**, *91*, 2443–2464.
43. Massutí, E.; Reñones, O. Demersal resource assemblages in the trawl fishing grounds off the Balearic Islands (western Mediterranean). *Sci. Mar.* **2005**, *69*, 167–181.
44. Balcells, M.; Fernández-Arcaya, U.; Lombarte, A.; Ramon, M.; Abelló, P.; Mecho, A.; Company, J.; Recasens, L. Effect of a small-scale fishing closure area on the demersal community in the NW Mediterranean Sea. In *Rapports et Procès-Verbaux des Réunions de la Commission Internationale pour l'Exploration Scientifique de la Mer Méditerranée*; Kiel, Germany, 2016; pp. 41–517.
45. Otsu, N. A threshold selection method from gray-level histograms. *IEEE Trans. Syst. Man Cybern.* **1979**, *9*, 62–66.
46. Huang, G.B.; Chen, L.; Siew, C.K. Universal approximation using incremental constructive feedforward networks with random hidden nodes. *IEEE Trans. Neural Netw.* **2006**, *17*, 879–892.
47. Huang, G.B.; Zhu, Q.Y.; Siew, C.K. Extreme learning machine: Theory and applications. *Neurocomputing* **2006**, *70*, 489–501.
48. Vapnik, V. *The Nature of Statistical Learning Theory*; Springer Science & Business Media: Berlin/Heidelberg, Germany, 2013.
49. Ren, S.; He, K.; Girshick, R.; Sun, J. Faster r-cnn: Towards real-time object detection with region proposal networks. In Proceedings of the Advances in Neural Information Processing Systems, Montreal, QC, Canada, 7–12 December 2015; pp. 91–99.
50. Ren, S.; He, K.; Girshick, R.; Sun, J. Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks. *IEEE Trans. Pattern Anal. Mach. Intell.* **2017**, *39*, 1137.
51. He, K.; Gkioxari, G.; Dollár, P.; Girshick, R. Mask r-cnn. In Proceedings of the IEEE International Conference on Computer Vision, Venice, Italy, 22–29 October 2017; pp. 2961–2969.
52. Álvarez-Ellacuría, A.; Palmer, M.; Catalán, I.A.; Lisani, J.L. Image-based, unsupervised estimation of fish size from commercial landings using deep learning. *ICES J. Mar. Sci.* **2019**, doi:10.1093/icesjms/fsz216.



© 2020 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<http://creativecommons.org/licenses/by/4.0/>).





Article

# Applying Knowledge Inference on Event-Conjunction for Automatic Control in Smart Building

Hangli Ge <sup>1,\*</sup> , Xiaohui Peng <sup>2</sup> and Noboru Koshizuka <sup>1</sup>

<sup>1</sup> Interfaculty Initiative in Information Studies, The University of Tokyo, Tokyo 1130033, Japan; noboru@koshizuka-lab.org

<sup>2</sup> Chinese Academy of Sciences, Institute of Computing Technology; Beijing 100190, China; pengxiaohui@ict.ac.cn

\* Correspondence: hangli@g.ecc.u-tokyo.ac.jp; Tel.: +81-03-38155411

**Abstract:** Smart building, one of IoT-based emerging applications is where energy-efficiency, human comfort, automation, security could be managed even better. However, at the current stage, a unified and practical framework for knowledge inference inside the smart building is still lacking. In this paper, we present a practical proposal of knowledge extraction on event-conjunction for automatic control in smart buildings. The proposal consists of a unified API design, ontology model, inference engine for knowledge extraction. Two types of models: finite state machine(FSMs) and bayesian network (BN) have been used for capturing the state transition and sensor data fusion. In particular, to solve the problem that the size of time interval observations between two correlated events was too small to be approximated for estimation, we utilized the Markov Chain Monte Carlo (MCMC) sampling method to optimize the sampling on time intervals. The proposal has been put into use in a real smart building environment. 78-days data collection of the light states and elevator states has been conducted for evaluation. Several events have been inferred in the evaluation, such as room occupancy, elevator moving, as well as the event conjunction of both. The inference on the users' waiting time of elevator-using revealed the potentials and effectiveness of the automatic control on the elevator.



**Citation:** Ge, H.; Peng, X.; Koshizuka, N. Applying Knowledge Inference on Event-Conjunction for Automatic Control in Smart Building. *Appl. Sci.* **2021**, *11*, 935. <https://doi.org/10.3390/app11030935>

Received: 23 December 2020

Accepted: 14 January 2021

Published: 20 January 2021

**Publisher's Note:** MDPI stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.



**Copyright:** © 2021 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

**Keywords:** smart building; Internet of Things (IoT); Markov chain Monte Carlo (MCMC); ontology; graph model

## 1. Introduction

Internet of Things (IoT) technologies [1] have enabled a variety of sensors and devices inside building, such as light, HVAC (heating, ventilating, and air conditioning), alarm system, surveillance camera, power meters, occupancy sensor, etc. being real-time monitored or controlled. Furthermore, artificial intelligence (AI) provides opportunities for innovative application development, for instance, supervisory automation, occupancy comfort optimization, energy efficiency improvement, indoor health management, security management, thus empower the building to be smart.

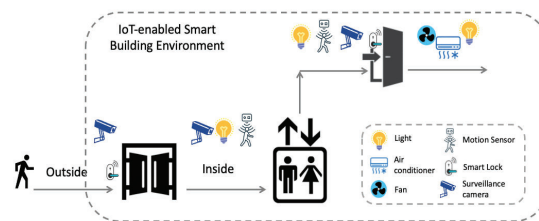
The most appealing benefit of smart building technologies is this revolution in building management systems (BMSs) [2], where the data collected from various sensors is processed and analyzed for enabling energy optimization, automation, and so on. In terms of revenues, researchers estimated that connected devices into the global BIoT market generated revenues of more than \$1.2 billion in 2018. While building automation market will grow at a compound annual growth rate (CAGR) of 44 percent to reach 19.4 million in 2022. This trend will grow at a CAGR of 21 percent to almost \$2.7 billion in 2022.

However, studies show that dynamic automation solutions are still insufficient [2]. Deploying automation in smart buildings requires a large amount of manual effort and building specific domain expertise. Yet, this vision is far from realization. It is still a challenge for modeling the context including users, sensors, actuators (so-called smart

device), spaces, etc, in an effective way for knowledge computation. Various sensory data collected from sensors need to be analyzed by algorithms, transformed into information, and minted to extract knowledge so that machines can have a better understanding of humans [3]. So far, most existing studies mainly focus on human activity recognition in a small-size space with limited numbers of devices or sensors. These machine learning-based approaches usually treat the building as a black-box. They ignore the building's physical structure, do not capture the global relations among the deployed sensors, spaces, as well as the observation of both the sensor value and the timestamp in a holistic view.

We consider in most applications of smart building, there must be tight relationships among the users, sensors, and the physical structure of the space. Especially, regarding human motion trace, there are strong spatial dependencies among the sensor observations. Therefore, a holistic and conditional probabilistic approach that considers human activity contexts and human-machine interactions (e.g., elevator motion, door-open, light-on, etc.) could be suggested. As shown in Figure 1, while a user entered a building and was heading to his/her sit, he/she activates multiple sensors/devices along the path.

Though, camera-based image processing approaches could achieve relatively high accuracy for human trajectory tracking. Contrary to the outdoor environment or other open/public spaces (roads, streets, etc.), privacy preservation is generally required well indoors (e.g., offices, meeting rooms, residential spaces, etc.). Therefore, non-invasive sensing technologies (sensor data of devices/appliances, etc.) are more appropriate than the use of cameras. In this study, we focus on such non-invasive sensor nodes.



**Figure 1.** Scenario of user walking outside to inside the smart building with the sensors being triggered.

However, while deploying machine learning (ML) approaches for detection, a large scale of label data was required if pursuing high level of accuracy. Especially, regarding human motion trace related event, the label data could be collected is small or sometimes incomplete. For an example, the room occupancy event happens several times per day, or rarely happens if the functionalities of the room are restricted or the space is not publicly open. That means it is quite difficult and time-costly to collect such room occupancy label data at a large scale. Moreover, it is still an ad-hoc process of defining which sensor node should be used and how to combine them for ML computation. Finally, it hinders the development of machine learning methods to extract knowledge of event in an automatic manner, for realizing such as room occupancy detection, human motion tracking, and so on.

In this paper, we present our proposal: a practical proposal of knowledge inference on event in IoT-enabled smart building environment. The proposal leverages the Building Topology Ontology (BOT) for constructing spatial graphs among sensors and spaces for further enabling conditional reasoning. In particular, considering the collected data is small, we utilized the Markov Chain Monte Carlo (MCMC) sampling method to approximate the time interval values of two correlated events. The proposal has been put into use in a real smart building environment. Several inference scenarios have been conducted. Moreover, the inference on users' waiting time revealed the effectiveness of automatic control on elevator for pursuing zero-waiting time. Hence, the primary contributions and novelties of this work can be summarized as follows:

- Unified API development of IoT sensor network for event inference based on sensor-data collaboration in the smart building;
- Ontology-based graph model for constructing the spatial relations among sensors and space for further enabling automatic event extraction based on conditional reasoning;
- Inference engine, in which two types of models: FSMs (finite state machine) and BN (Bayesian Networks) have been proposed for event-mapping. Further, Markov Chain Monte Carlo (MCMC) model has been utilized for enhancing the accuracy of sampling the small-size dataset of time-interval values.
- Usage scenario of automatic control on elevator control has been conducted for evaluation. The numerical results demonstrate the potential and effectiveness of automatic control application based on knowledge inference in the smart building.

## 2. Related Work

### 2.1. Machine Learning Approaches

Most existing solutions that use machine learning (ML) for smart building applications focused on the occupant, including occupancy detection [2,4], activity recognition [5], and estimating users' preferences and identification [2]. Khan et al. [4] dealt with occupancy of premise range from binary occupancy (occupied or out of occupied), categorical and exact numbers by integrating several types of sensors, including PIR, acoustic noise, humidity, and light, and so on. Hossain et al. [6] proposed using an active learning approach for activity recognition in residential buildings. The proposal was motivated by the variety of human activities thus based on K-means algorithm. It requires the provision of vast amounts of labeled data for ensuring the supervised learning approaches be effective, which is not always possible. Most of these existing ML-based works focused on solving the detection problem which has been taken in a limited space. The physical sensor deployment has been ignored.

To reduce the complexity of knowledge transfer across different domains, Chiang et al. [7] focused on exploring the differences caused by the ambient sensors and the target domain, proposed a framework that knowledge transfer that uses standard SVM (support vector machine) and RBF (radial basis function). However, in their proposal, only single-resident scenario was considered. Similarly, Hong et al. [8] proposed automatic inference on the type of sensors in a building. They focused on the classification of sensor types without manual labeling. However, these related works focused on the inference of sensor types, parameters and so on. They did not capture sensor observations for event detection.

### 2.2. Modeling Tool of Spatial Graph

In regard to the modeling tools of sensor deployment, standard practical solutions are still lacking. Building Information Modeling (BIM) is a framework to support the planning and construction of buildings. Industry Foundation Classes (IFC) standard [9], which is a well-known representation of BIM, considers the elements inside a building as objects that are defined by a 3D geometry and normalized semantics. The Green BuildingXML (gbXML) [10] emerged to allow sharing information between BIM and energetic analysis software. However, the main intentions of these tools are the modeling of the physical structure and used materials, which is static. Their main focus is on the physical environment setup. That means they are often used for structural analysis or 2D/3D modeling by using CAD tools. The functional aspects of knowledge extraction of building systems are not covered by these approaches.

Many research projects are actually elaborating semantic models for facilitating building management, such as rule-based methods for supervision [11], definition or classification of metadata schema for facilitating building management [8,12,13]. An ontology is a vocabulary based method for defining the concepts and relationships used to describe an area of concern based on RDF (Resource Description Framework) [14]. A few of specific ontologies have been proposed for the domain of smart homes and

buildings [11,15–19]. Most of these ontologies focus on realizing specific applications like energy management [18–20], or automated design and operation [11,16–19].

The SOSA (sensor, observation, sample, and actuator) ontology [21] and semantic sensor network (SSN) ontology are W3C recommendations, providing an approach to describe hardware, observation of physical entities and actuation, etc. The BOnSAI [15] ontology was proposed for describing the functionality of sensors, actuators, and appliances. However, it does not provide sufficient information on spatial relationships among the sensors and other building assets. The Building Topology Ontology (BOT) [17] defines the relationships between the sub-components of a building. It also follows general W3C principles and was suggested as an extensible baseline for reuse.

To summarize these ontology-based studies, whether merely for modeling the resource description or knowledge extraction, ontology that constructs the physical relations in the smart building is considered as suitable to present graph concepts. Moreover, aiming for removing ambiguity, and pursuing application portability, unifying the sensor data format is one of the most important considerations. In addition, the W3C endorsed ontologies demonstrate high potentials of upper-layer application development. Thus, rather than proposing a new ontology, our approach preferred to reuse the existing ontologies, and extend them by adding other necessary specific information.

### 3. Problem Definition

Before knowledge inference in the smart building being ubiquitous, there are still several technical challenges to confront:

- **Integration and interoperability of heterogeneous data set**  
The most fundamental problem is the compatibility of heterogeneous sensors / devices, providing different networking features, protocols, and interfaces from different vendors. The transition and integration between the heterogeneous sensors are costly and make smart building implementation slow down. High level of manual effort is currently required to integrate the sensor or device nodes, which are often decentralized in both the cyber and physical dimensions, varying with their parameters. Such processes are both time-consuming and error-prone [12]. That leads to, while deploying inference framework for smart building, the developers need to map various data from heterogeneous sensors without a common format or unit.
- **Lack of semantic approach**  
Relevant description logics (DL) is required to deal with environmental data within smart buildings, such as the type, instance as well as the relevance, relation among the entities for knowledge extraction. However, the complicated indoor environment with various features including general, spatial, temporal, spatio-temporal leads to a standard description logics (DL) for the IoT sensor network in the smart building being lacked. Ontology could be used for constructing the relational graph among sensors and spaces, etc. However, the ontology needs to capture the dependencies accurately, while not being excessively complex to make inference hard.
- **The small and incomplete data features**  
Although environmental data could be collected over time, the human-motion related sensor data is somehow sparsely triggered in both the spatio and temporal dimensions. Most of the sensor data would be got rid of being labels for machine learning. Thus, it is difficult to collect the label data at a large scale in the real phenomenon. That means while developing knowledge inference on event conjunction by taking into account continuous changes inside the whole environment, the problem of handling small and incomplete data should not be ignored.

### 4. Proposal and Experiment

We introduce our proposal as below: a practical knowledge extraction platform in IoT-enabled smart building environment. The proposal fuses various IoT-enabled devices or sensors inside the building for knowledge extraction on events. It consists of three major

components: (1) unified IoT API for sensor data collaboration; (2) knowledge base in which ontology for constructing the physical relational graph was utilized; (3) inference engine. Figure 2 describes the overview of our proposal.

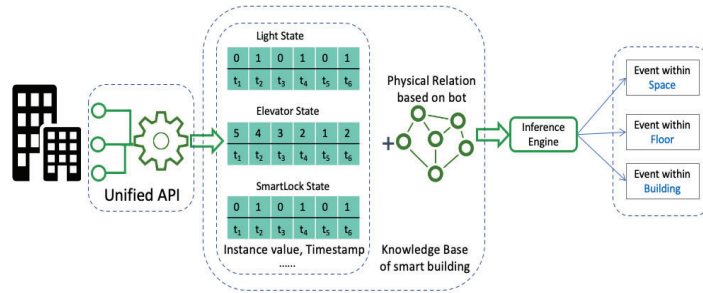


Figure 2. Overview of our proposal for knowledge extraction.

#### 4.1. Unified API Design

In order to solve the above-mentioned problems, the development of a unified API for interoperability among IoT devices becomes a fundamental aspect. Unifying the device API would ease and accelerate the new service development in the smart building, which brings innovation and productivity. However, unifying of device API is considered challenging, because the heterogeneous devices with different functionalities have different specifications and configurations. A unified API has to cover all IoT devices and simplify the properties of such devices.

Based on the exploration of device properties, we designed a unified API for receiving the monitored state information in real-time. The details of API is interpreted by the following Figure 3. The following set of properties has been contained: (1) ‘ucode’ [22] that used as the identifier of the sensor node; (2) ‘name’ that assigned with the description of the node; (3) ‘data’, that composed by the sub-properties of ‘instance’ and ‘time’, with ‘instance’ showing the sensed value and ‘time’ indicating the timestamp.

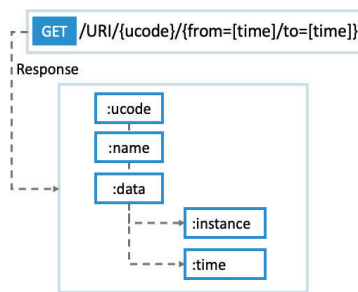


Figure 3. The unified API design.

The ‘ucode’ [22] is a 128-bit fixed-length identifier. It could be used as a unique identification for associating the objects in different databases. While accessing the API for retrieving data, the ‘ucode’ is required to be given whereas the time duration is not necessary. The system would capture the monitored sensing data during the time window for a response if the parameters of ‘from’ and ‘to’ were assigned. Meanwhile, if the timestamp was omitted, only the latest data value would be responded.

Table 1 lists several examples of sensor data that could be retrieved from the unified API. The timestamp value is automatically detected by the system, such as the example

of '2020-10-07 09:46:54'. In addition, as sensor features consists of both numerical values and logic state (on/off) causing computation complexity, the state data are converted to numerical data with logical meanings, where 1 represents on/active/open, etc; 0 represents off/inactive/close, etc.

**Table 1.** Data value examples of the unified API.

Sensor/device	The meaning of binary state value (1/0)
Human detector	human-detected; otherwise
Light	on; off
Fan	on; off
Air conditioner	on ; off
Elevator_potion	3F, 2F, 1F, B1F, B2F

#### 4.2. Knowledge Database

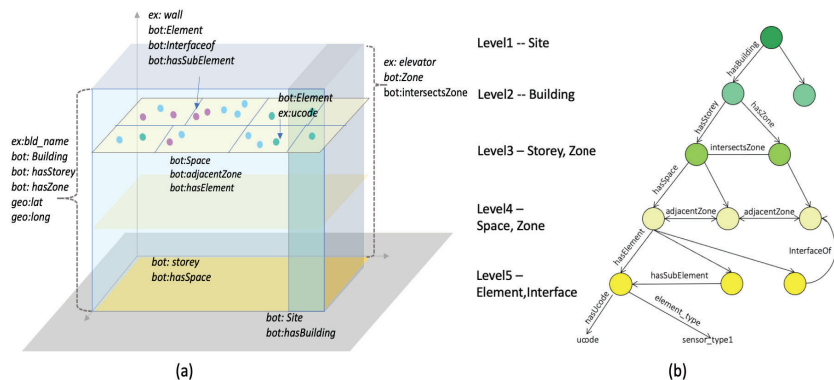
Spatial and functional relationships inside buildings could be considered as being graphical. In this research, our goal is not to develop a new ontology for modeling the building structure, but to show how IoT collaboration platforms in smart buildings leveraging semantic technologies could implement some global knowledge inference functions for automatic control.

Therefore, to comply with and take full advantage of existing standard works in this field, the Building Topology Ontology (BOT) was imported. The BOT is an OWL-DL ontology [23] for defining both the physical and semantic relationships of the sub-components inside a building. It was used to describe the semantic schema such as the physical relationships, functionality parameters inside the building, of whom the data set was static.

As shown in Figure 4a, high-level concepts of node hierarchies have been defined by BOT. It was composed of: classes (e.g., Building, Space, etc.) for representing a spatial entity; properties (e.g., has\_storey, has\_space, interface\_of, etc.) for representing the relations between entities; rules (e.g., wall could be modeled as an interface of two rooms; door within a wall could be modeled as a tuple of <wall, has\_element, door>). According to the BOT ontology, the corresponding triplets (representing the entities and relationships) of our building have been created. We referred the classes of BOT, i.g, site, building, zone, space, element, etc. The details have been described as follows.

- Site: An area contains one or more buildings, that further could be used in the field of city area computing.
- Building: Building node assembles all the sub-nodes within the building. Thus the whole context inside the building could be referenced through the building node.
- Storey: A level/floor part of the building.
- Zone: Elevator and stairwells are modeled as a Zone, used as links of multiple storeys through the BOT:intersectsZone relation.
- Space: A part of a storey, whose 3D spatial extent is bounded actually or theoretically. In general, it represents a common space, e.g., a room, an elevator hall, toilet, corridor, and so on. The space node assembles all the elements (i.g., sensor, actuator, device) located within it.
- Element: Sensor, actuator, device were defined as element nodes deployed in space. For each element, an identifier of 'ucode' has been assigned for associating the dynamic sensor observations through the unified API.





**Figure 4.** (a) Classes and relationships involved in the Building Topology Ontology; (b) The hierarchy of entities and relationships in BOT.

As shown in Figure 4b, aiming to interconnecting heterogeneous devices for further efficient reasoning and heuristics, the model was hierarchically structured along with the building architecture. We consider the graph-based hierarchical structure of the model fits the building structure for the reasoning strategy because the indoor users’ motion trace follows the features and layouts of the building. The model that is on representing hierarchy in the building and it fits bottom-up data collection and decomposition. Also, the model follows ordinary building designs that make it practicable for almost all the common buildings.

#### RDF Data Store and Sparql Query Language

We separated the static ontology and dynamic sensing data into RDF store and relational database store. Spatial relationships were constructed by a BOT graph. Meanwhile, sensor nodes (which were described as Element in the BOT) are continuously submitting data in real-time that causes the data store to be large and get updated frequently. We chose a relational database to store the sensor data. Those two data stores were associated with the unique identification of ‘uocode’. Combining the ontology and relational database enables to process of spatial-temporal data efficiently.

The choice of ontology informed the RDF data model [24], and SPARQL query language [25] being selected for representing and querying graphs, respectively. The RDF triple is a 3-tuple of <subject, predicate, object> that states a subject has a relationship predicate (directed edge) to an entity object. SPARQL [25] defines a set of patterns that constrains the set of RDF terms returned from the graph. Figure 5a shows a part of the triple examples of our ontology data store. We chose Apache Jena for storing the ontology data by RDF data structure. Apache Jena [26] is an open-source framework for managing and querying RDF data. It contains a web frontend (Fuseki) and a SPARQL backend (TDB) that supports all SPARQL 1.1 features. It also provides an API to extract data from and write to RDF graphs by sparql protocol and RDF query language (SPARQL).



moving speed ( $T_{moving}$ ) and several deterministic states. Therefore, finite state machines are suitable to describe the state logic of observations on elevator’s moving. As shown in Figure 6, a graph model of FSMs was used to implement the state transition of the elevator. The details of parameters description were listed in Table 2.

Table 2. Parameters of the FSM model.

No.	Symbol	Description
1	$\Delta t$	the time duration after the last sensed time
2	$T_{moving}$	the constant time of moving between two floors
3	$T_{threshold}$	the time threshold value to distinguish the states of “Boarding” and “Waiting”

In essence, the elevator changes its states according to passengers’ activities and requests. Thus, the simultaneous state-mapping on users and elevator has been clarified as shown in Figure 8. According to the predefined state transition of both the elevator and users, the following event patterns could be elaborated: (1) while the elevator was moving for picking-up, the user remains to wait at the departure floor; (2) while the elevator was transporting, the user was riding on the cabin for heading to the destination floor; (3) while the elevator was boarding for picking-off, that means the user had arrived at the destination floor.

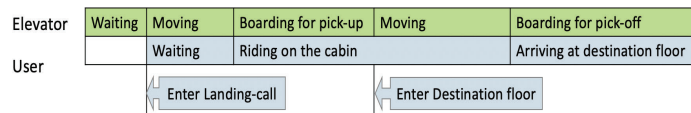


Figure 8. The state transitions of both the elevator and user.

**Byesian network (BN).** On the other hand, a few of events with high-level semantics were inferred based on the conjunction of multiple correlated events. Graphical dependencies among these event and sensor data observations in both the spatial and temporal dimensions could be observed. A probabilistic graphical model has been proposed based on the conditional inference on two correlated events. For example, there are bidirectional effects on both the use of elevator and room-occupancy event (namely that when the user leaves the room, he/she would call the elevator to move to another floor; when the elevator comes to the floor for picking-off users, an event that user enters the room might happen).

In this proposal, Bayesian network (BN) was utilized for modeling the probabilistic graph on these conditional events inside the building. As the probabilistic state transitions shown in Figure 7, let  $P_e(s_i, S)$  denotes the probability of estimated event based on the sensor observation of  $s_i$  on the overall observations of  $S$ . The sensor observation  $s_i$  was denoted as the tuple values  $\langle v_i, t_i \rangle$  and could be directly extracted from the unified API. For every sensor node in the BOT graph we defined the true state on the probability distribution. The sensor observations  $S = \{s_1, s_2..s_n\}$  was used for computing the probability distributions with the measurement of mean, median, minimum and maximum, etc.

Algorithm 1 shows the overall algorithm that was applied for inferring the continuous events based on the sensor observations. For every sensor node in the original dependency graph, we add an event based on the observed state and timestamp in the event graph. For every  $\Delta t$  value on edge  $i, j$  of two estimated events, if the conditional probability of  $\Delta t$  is greater than the threshold value, the event conjunction was estimated and event  $E_{i \rightarrow m}$  was set to be true and be added on the event graph.

### Markov Chain Monte Carlo (MCMC) Sampling Process

One of the main challenges in this inference engine is the probabilistic modeling on  $\Delta t$ . The graph needs to capture the probability based on the statistical analysis of time intervals. As shown in Equation (1), the Bayesian paradigm (so-called Bayes theorem)

---

**Algorithm 1:** Generic Knowledge Extraction Algorithm

---

**Input:** KG with the data triple of  $(e_i, s_i)$  where  $e_i$  is an element in the sensor knowledge graph;  $s_i$  denotes the sensor observation consisting of  $v_i$  and  $t_i$ , which could be queried from the unified API

**Output:**  $\{E\}$  is noted as the set of  $(E_i, t_i)$ , means the inferred event  $E_i$  at the time of  $t_i$  based on  $P_{(E_i|s_i, S_i)}$

```

1 For each sensor observation  $s_i$  on the space entity  $e_i$ :
2   Compute  $P_{(E_i|s_i, S_i)}$ 
3   while  $P_{(E_i|s_i, S_i)} > \check{\tau}$  do
4     Add  $(E_k, t_i)$  to  $\{E\}$ 
5     while exists the linked neighbor  $e_m$  of which  $P_{(E_m|s_m, S_m)} > \hat{\tau}$  do
6       Add  $(E_m, t_j)$  to  $\{E\}$ 
7        $\Delta t = t_j - t_i$ 
8       Compute  $P(E_m|E_i, \Delta t)$ 
9     end
10  end

```

---

was deployed to express the relation between three terms: a prior knowledge, a likelihood (the knowledge coming from the observation), and a “posterior” (the updated knowledge). Meanwhile, it can be noticed that one of the main difficulties faced when dealing with a Bayesian inference problem comes from while the size of posterior samples is not enough to converge.

In this proposal, we utilized the MCMC sampling method to overcome the mentioned above issue. MCMC algorithms are aimed at generating samples from a given probability distribution. It is useful for obtaining a sequence of random samples from a probability distribution in which direct sampling is difficult, or the sample data is small or incomplete. Thus, instead of trying to deal with intractable computations involving the posterior, we can get samples from using the existing samples and some definite prior value to compute various punctual statistics to approximate the distribution by kernel density estimation.

$$\overbrace{p(\mu | Data)}^{\text{posterior}} = \frac{\overbrace{p(Data | \mu)}^{\text{likelihood}} \cdot \overbrace{p(\mu)}^{\text{prior}}}{\underbrace{p(Data)}_{\text{marginal likelihood}}} \tag{1}$$

The Metropolis–Hastings algorithm, one of the most common methods of MCMC based sampling process, was utilized for drawing samples from probability distribution  $P(\Delta t)$ , provided that we know a function  $f(\Delta t)$  is proportional to the density of  $P(\Delta t)$  and the values of  $f(\Delta t)$  can be calculated. The requirement that  $f(\Delta t)$  must only be proportional to the density, rather than exactly equal to it, makes the Metropolis-Hastings algorithm particularly useful while the size of event-related sensor observations is relatively small.

**Local distance-based adjustment** In this sampling process, as shown in Equation (2), the prior distribution of  $\Delta t$  was adjusted according to the shortest distance  $Dist(e_i, e_j)$  between two space entities  $(e_i, e_j)$  extracted from the BOT graph, where the prior probability  $P(\mu)$  is the probability of the hypothesis  $\mu$  before the Data D, was modeled as a Gaussian distribution.

$$\mu \sim N(f_{Dist(e_i, e_j)}, \sigma) \tag{2}$$

where the value of  $f_{Dist(e_i, e_j)}$  showing the shortest distance is linearly-correlated with the count of hops between two triggered sensor nodes in the graph, as denoted in Equation (3).

$$f_{Dist(e_i, e_j)} = ax + b \tag{3}$$

Based on the shortest distance values extracted from the BOT graph and the  $\Delta t$  values collected in the evaluation, the values of  $a, b$  in the linear regression could be calculated by the Equations (4) and (5). Figure 9 shows the result of linear regression optimization of  $Dist(e_i, e_j)$  with several distance examples have been illustrated as well.

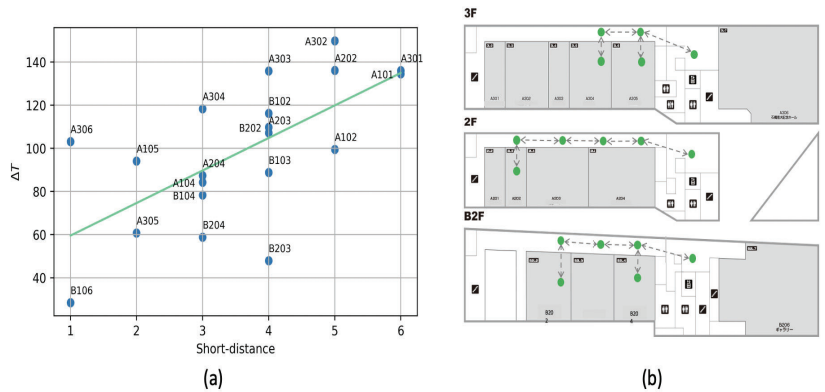
$$C = \sum_1^n (y - \hat{y})^2, \text{ where } \frac{\partial C}{\partial a} = 0, \frac{\partial C}{\partial b} = 0 \tag{4}$$

$$a = \frac{\sum_1^n xy - \frac{1}{n} \sum_1^n x \sum_1^n y}{\sum_1^n x^2 - \frac{1}{n} (\sum_1^n x)^2}, b = \frac{1}{n} \sum_1^n (y - ax) \tag{5}$$

Collected data of  $\Delta t_1, \dots, \Delta t_n$ , was used for computing the prior probability. With given the measured quantities  $\Delta t_1, \dots, \Delta t_n$ , the probability function has been modeled as normally distributed, shown in Equation (6).

$$\Delta t_i \sim N(\mu, \sigma), \text{ where } f(\Delta t_i | \mu, \sigma) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(\Delta t_i - \mu)^2}{2\sigma^2}} \tag{6}$$

In order to derive the approximated value  $\mu$  of  $\Delta t_1, \dots, \Delta t_n$ , PyMC3 [27] was utilized for performing Bayesian statistical sample processing focused on MCMC. PyMC3 is an open-source probabilistic programming framework written in Python. It is based on Theano.



**Figure 9.** (a) Result of linear regression of the shortest-distance in the graph and the  $\mu$  value of  $\Delta t$ , where the value of  $a, b$  are estimated to be 14.46, 26.7, respectively; (b) Several shortest distance examples between rooms and elevator in our building.

**5. Experiment and Evaluation**

The experiments have been conducted in a real smart building named “Daiwa ubiquitous computing research building” in the University of Tokyo. Figure 10 shows the IoT-enabled environment of our smart building. The building has 5 floors including B2F, B1F, 1F, 2F, 3F, and 43 space entities (i.e., room, hall, or corridor, etc.) and 1 elevator. At the meantime, 846 spatial relation triples have been restored in the BOT graph.

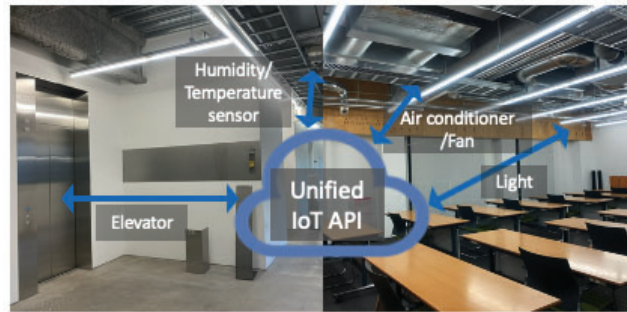


Figure 10. The IoT-enabled smart building environment.

In this experiment, we used the data collection of 114 lights and 1 elevator for knowledge inference on the events of room occupancy and elevator motion. Figure 11 shows the visualization of the light status in several rooms of the building, spanning 78-days (from 13 September 2020 to 30 November 2020). The sizes of the sensor observations have been listed in Table 3. It is worth to note that during the covid-19 pandemic, the collected occupancy-related or event-conjunction-related sensor observations were much less than normal periods.

Table 3. Sizes of the collected data.

Item	Raw Data	Validate Data
Light state	1,260,139	2669
Elevator state	14,377	14,377

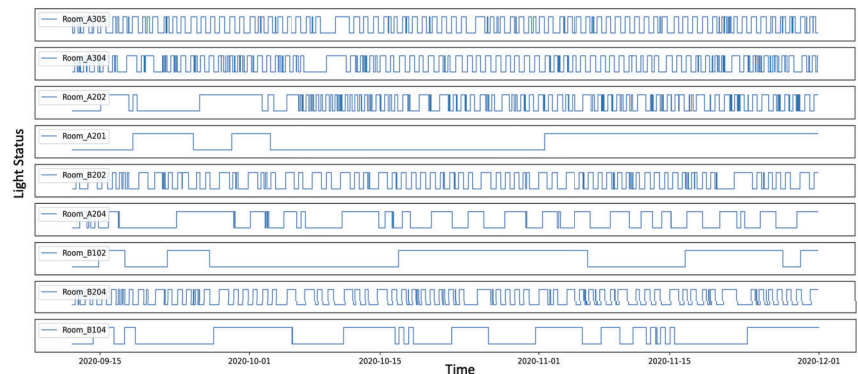
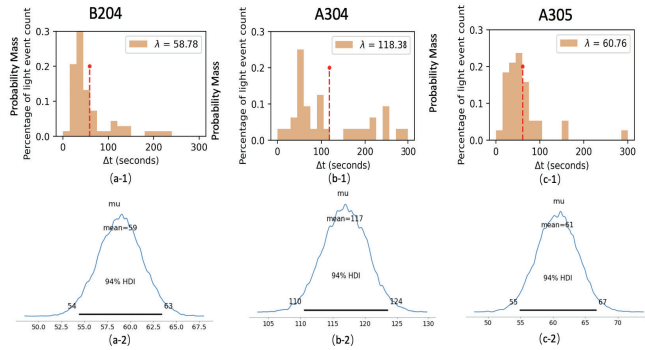


Figure 11. The real examples of the monitored light states in several rooms.

### 5.1. Trace on Event Conjunction

Based on the MCMC sampling process, the value of  $\Delta t$  between two corresponding events has been approximated. Figure 12 shows the results of approximated time intervals between the event of elevator-arriving and light-up of several representative spaces. Table 4 listed the detail results as well. Further, based on the approximated  $\Delta t$ , such as the event conjunction of 'light\_turn\_off -> elevator\_arriving' could be inferred. As a result, the count of inferred room occupancy ( $E_r$ ), event conjunction ( $E_r \rightarrow E_{el}$ ) and their conditional probabilities have been summarized (see the details listed in Table 5). Here, the room of which a total number of events over 80 have been picked on the list.



**Figure 12.** Examples of approximated  $\mu$  values of  $\Delta t$ , where (a-1,b-1,c-1) show the mean value result on the observed raw data, while (a-2,b-2,c-2) show the approximated  $\Delta t$  after MCMC posterior sampling process.

**Table 4.** MCMC sampling results  $\Delta t$  of several representative spaces.

	Short_Dist (Hops)	$\mu$ Based on by LR Optimization	$\bar{\mu}$ of Observed Sensor Data	MCMC Sampling Process		
				$\check{\mu}$	hdi_3%	hdi_97%
Room_A305	2	55.62	60.76	60.68	54.852	66.673
Room_A304	3	70.08	118.38	116.933	110.495	123.556
Room_A202	5	99.00	136.09	134.465	126.945	142.471
Room_B202	5	84.54	109.72	109.144	103.462	115.141
Room_B204	3	70.08	58.78	55.62	54.366	63.421

**Table 5.** Inferred results on event-conjunction of several representative spaces.

	$E_r$ (Count)	$E_r \rightarrow E_{el}$ (Count)	$P_{e_r \rightarrow e_{el}}$ Conditional Probability
<b>Sum</b>	<b>1331</b>	<b>471</b>	<b>0.35</b>
Room_A305	110	38	0.345
Room_A304	129	32	0.248
Room_A202	122	22	0.180
Room_B202	90	39	0.433
Room_B204	89	72	0.809

5.2. Assumption for Automatic Control on Elevator for Zero-Waiting Time

In order to evaluate the usability of the proposed framework, an automatic control scenario on elevator has been assumed. In general, when the user wanted to use the elevator, he/she has to first reach the elevator hall and press the upward or downward button to make a call on the elevator (given the timestamp of  $t_2$  shown in Figure 13). Then the elevator received the command and arrived at the floor (given the timestamp of  $t_3$  shown in Figure 13) to pick-up the user. The automatic control scenario was an assumption based on knowledge inference on event conjunction. Suppose:

1. Room occupancy has been monitored in real-time based on the sensor observations;
2. If the agent detected a user leaving the room and further predicted the user has an intention of using the elevator;
3. Then, the agent triggers the elevator in-advance to move to the user’s departure floor.



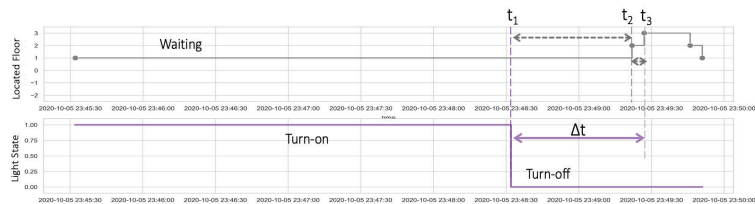


Figure 13. Example shows the real event trace of room-usage and elevator.

The above assumption means when the user arrives at the elevator hall at  $t_2$  time, the elevator has been ready for picking up the user. Thus, he/she can take a ride on the elevator immediately and no waiting time ( $t_3 - t_2$ ) was needed.

Therefore, the effectiveness of automatic control on elevator was evaluated by extracting the historical users' waiting time. The numbers of inferred event conjunction of room occupancy  $\rightarrow$  elevator arriving ( $E_r \rightarrow E_{el}$ ) has been listed in Table 6. Regarding the elevator usage, the count of non-waiting, waiting have been calculated as well. Furthermore, the percentages of the different waiting time also have been inferred, respectively.

Table 6. Statistics of the users' waiting event on elevator based on the mapping result of FSMs.

	$E_r \rightarrow E_{el}$	Without Waiting	With Waiting			
Sum	471	210	267			
Probability			5 s	10 s	15 s	20 s
			0.232	0.446	0.101	0.221

Figure 14 shows the statistical result of users' waiting time calculated from collected sensor data. The result demonstrated there were 267 waiting events happened among the total of 471 elevator-using events, with the probability of waiting was 56.7%. In addition, the total amount of waiting time was 3085 s and the average waiting time per user was 11.55 s (SD = 5.30 s).

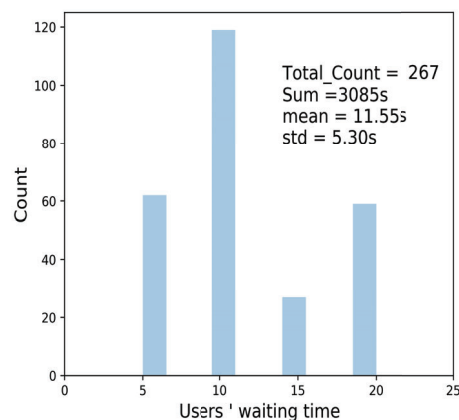


Figure 14. The statistics of users' waiting time on elevator mapped by FSMs.

These numerical results of users' waiting time demonstrated the great potentials of automatic control on elevator for pursuing the goal of zero-waiting of using the elevator. The quantitative results also showed the effectiveness of the knowledge inference on

event conjunction in smart building, for further improving the transport efficiency and productivity of indoor users.

## 6. Discussion

In the experiments, room occupancy was determined as binary-mapping from light state changing. The rule was set as: while the light is turning-on, the room is in occupancy; otherwise, the room is out of occupancy. However, in real usage, more complicated situations could be taken into consideration. Since people left the room without turning off the light sometimes happens, multi-modal sensor fusion should be considered if pursuing the inference accuracy. For example, the state of light and smart lock could be combined for improving the accuracy of room occupancy detection. Nevertheless, the BOT-based graph has provided opportunities for modeling other sensor observations in a structured hierarchical graph. There could be few challenges for modeling other different types of sensors to the existing ontology graph. Thus we consider that our approach could be adapted to other sensor resources in the smart building if available, and the methodology is practical for other smart buildings.

On the other hand, various sensors are currently installed in the smart building. In addition to the diversity of sensors, more benefits of our proposal could be quantified after a diverse range of automatic control application being implemented. For example, tracing on human motion in the smart building, to automatic control the appliance pursuing reducing the energy consumption, improving human comfort, health in the smart building. These mentioned-above application scenarios rely on knowledge inference in smart building. Each part of this proposal: unified API of sensor network, knowledge graph of the physical environment and the inference engine, was considered to be indispensable. In this experiment, we formed a graph with the physical relations. Semantic schema (e.g., users' identification, preferences, relations, as well as space affiliations, etc.) has not been modeled in the graph. However, adding to these attributes, the inference engine would be capable of analyzing user-related semantics.

## 7. Conclusions

In this paper, we presented a practical approach of event inference for automatic control in IoT-enabled smart building environment. The proposal consists of unified API development, knowledge base and inference engine. The event inference models based on sensor observations was separated into deterministic and probabilistic. Therefore, two types of models: finite state machine (FSMs) and bayesian network (BN) have been used for capturing the state transition and sensor data fusion. As opposed to earlier straightforward machine learning-based methods, our proposal focused on the conditional conjunction and transition of two correlated events, for which a graph model of the physical environment was considered necessary.

To tackle the problem of the sizes of time interval ( $\Delta t$ ) observations were too small to derive accurate results, MCMC sampling process has been utilized for approximating the time intervals ( $\Delta t$ ). Specifically, linear regression of local distances between two space entities on the ontology graph has been leveraged for the optimization of the sampling process. The proposal has been implemented in a real smart building environment and 78-days data collection of the state on light and elevator has been conducted for evaluation. Event conjunctions on the light and elevator have been utilized for further inferring room occupancy and indoor users' trajectories.

To show the usability of the proposal, we extracted the knowledge of users' waiting time on the elevator. The FSM mapping result of elevator-using demonstrated the probability of users' waiting event was 56.7%, with the total waiting time during the evaluation was 3085 s and average waiting time was 11.55 s. The numerical results demonstrated the potential of automatic control for zero-waiting on elevator based on knowledge inference on event conjunction in smart building.

**Author Contributions:** Writing—original draft preparation, H.G.; conceptualization, X.P.; supervision, N.K. All authors have read and agreed to the published version of the manuscript.

**Funding:** This research received no external funding.

**Institutional Review Board Statement:** Not applicable.

**Informed Consent Statement:** Not applicable.

**Data Availability Statement:** Not applicable.

**Conflicts of Interest:** The authors declare no conflict of interest.

## References

- Atzori, L.; Iera, A.; Morabito, G. The internet of things: A survey. *Comput. Netw.* **2010**, *54*, 2787–2805.
- Djenouri, D.; Laidi, R.; Djenouri, Y.; Balasingham, I. Machine learning for smart building applications: Review and taxonomy. *ACM Comput. Surv. (CSUR)* **2019**, *52*, 1–36.
- Qolomany, B.; Al-Fuqaha, A.; Gupta, A.; Benhaddou, D.; Alwajidi, S.; Qadir, J.; Fong, A.C. Leveraging machine learning and big data for smart buildings: A comprehensive survey. *IEEE Access* **2019**, *7*, 90316–90356.
- Khan, A.; Nicholson, J.; Mellor, S.; Jackson, D.; Ladha, K.; Ladha, C.; Hand, J.; Clarke, J.; Olivier, P.; Plötz, T. Occupancy monitoring using environmental & context sensors and a hierarchical analysis framework. In *Buildsys '14 - 1st ACM Conference on Embedded Systems for Energy-Efficient Buildings*, Memphis, USA, 3 November 2014:90–99.
- Ahmadi-Karvigh, S.; Ghahramani, A.; Becerik-Gerber, B.; Soibelman, L. One size does not fit all: Understanding user preferences for building automation systems. *Energy Build.* **2017**, *145*, 163–173.
- Hossain, H.S.; Khan, M.A.A.H.; Roy, N. Active learning enabled activity recognition. *Pervasive Mob. Comput.* **2017**, *38*, 312–330.
- Chiang, Y.T.; Lu, C.H.; Hsu, J.Y.J. A feature-based knowledge transfer framework for cross-environment activity recognition toward smart home applications. *IEEE Trans. Hum. Mach. Syst.* **2017**, *47*, 310–322.
- Hong, D.; Wang, H.; Ortiz, J.; Whitehouse, K. The building adapter: Towards quickly applying building analytics at scale. In *Proceedings of the 2nd ACM International Conference on Embedded Systems for Energy-Efficient Built Environments*, Seoul, Korea, 4, November 2015; pp. 123–132.
- Liebich, T. (2013). IFC4—The new buildingSMART standard. In *IC Meeting*. Helsinki, Finland: BSI Publications.
- Autodesk. gbXML. Available online: <https://www.gbxml.org/> (accessed on 23 August 2020).
- Tamani, N.; Ahvar, S.; Santos, G.; Istasse, B.; Praca, I.; Brun, P.E.; Ghamri, Y.; Crespi, N.; Becue, A. Rule-based model for smart building supervision and management. In *Proceedings of the 2018 IEEE International Conference on Services Computing (SCC)*, San Francisco, CA, USA, 2 July 2018; pp. 9–16.
- Gao, J.; Ploennigs, J.; Berges, M. A data-driven meta-data inference framework for building automation systems. In *Proceedings of the 2nd ACM International Conference on Embedded Systems for Energy-Efficient Built Environments*, Seoul, Korea, November 4–5 2015; pp. 23–32.
- Balaji, B.; Bhattacharya, A.; Fierro, G.; Gao, J.; Gluck, J.; Hong, D.; Johansen, A.; Koh, J.; Ploennigs, J.; Agarwal, Y.; et al. Brick: Metadata schema for portable smart building applications. *Appl. Energy* **2018**, *226*, 1273–1292.
- Miller, Eric. *Resource Description Framework (RDF) Model and Syntax Specification*; Bulletin of the American Society for Information Science and Technology 25.1 (1998): 15–19.
- Stavropoulos, T.G.; Vrakas, D.; Vlachava, D.; Bassiliades, N. BOnSAI: A smart building ontology for ambient intelligence. In *Proceedings of the 2nd International Conference on Web Intelligence, Mining and Semantics*, Craiova, Romania, June 6–8 2012; pp. 1–12.
- Ploennigs, J.; Hensel, B.; Dibowski, H.; Kabitzsch, K. BASont-A modular, adaptive building automation system ontology. In *Proceedings of the IECON 2012-38th Annual Conference on IEEE Industrial Electronics Society*, Montreal, QC, Canada, October 25 2012; pp. 4827–4833.
- Rasmussen, M.H.; Lefrançois, M.; Schneider, G.F.; Pauwels, P. BOT: The Building Topology Ontology of the W3C Linked Building Data Group. *Semantic Web, Volume 12, 1 Number 2021 (in press)*.
- Degha, H.E.; Laallam, F.Z.; Said, B. Intelligent context-awareness system for energy efficiency in smart building based on ontology. *Sustain. Comput. Inform. Syst.* **2019**, *21*, 212–233.
- Petrushevski, F.; Gaida, S.; Beigelboeck, B.; Sipetic, M.; Zucker, G.; Schiefer, C.; Schachinger, D.; Kastner, W. Semantic building systems modeling for advanced data analytics for energy efficiency. *Build. Simul.* *Proceeding of the 15th IBPSA Conference San Francisco, CA, USA, Aug. 7–9, 2017*
- Kofler, M.J.; Reinisch, C.; Kastner, W. A semantic representation of energy-related information in future smart homes. *Energy Build.* **2012**, *47*, 169–179.
- Janowicz, K.; Haller, A.; Cox, S.J.; Le Phuoc, D.; Lefrançois, M. SOSA: A lightweight ontology for sensors, observations, samples, and actuators. *J. Web Semant.* **2019**, *56*, 1–10.
- Koshizuka, N.; Sakamura, K. Ubiquitous ID: Standards for ubiquitous computing and the internet of things. *IEEE Pervasive Comput.* **2010**, *9*, 98–101.

23. World Wide Web Consortium. *OWL 2 Web Ontology Language Document Overview*; World Wide Web Consortium: Boston, MA, USA, 2012.
24. RDF. Available online: <https://www.w3.org/TR/1999/REC-rdf-syntax-19990222> (accessed on August 23 2020).
25. Seaborne, A.; Manjunath, G.; Bizer, C.; Breslin, J.; Das, S.; Davis, I.; Harris, S.; Idehen, K.; Corby, O.; Kjernsmo, K.; et al. SPARQL/Update: A language for updating RDF graphs. *W3c Memb. Submiss.* **Jul 2008**; 15
26. Apach. Available online: <https://jena.apache.org/> (accessed on 25 August 2020).
27. Salvatier, J.; Wiecki, T.V.; Fongesbeck, C. Probabilistic programming in Python using PyMC3. *PeerJ Comput. Sci.* **2016**, *2*, e55.



Article

# Training Set Enlargement Using Binary Weighted Interpolation Maps for the Single Sample per Person Problem in Face Recognition

Yonggeol Lee <sup>1</sup>  and Sang-II Choi <sup>2,\*</sup> <sup>1</sup> Police Science Institute, Korean National Police University, Asan 31539, Korea; pattern@police.go.kr<sup>2</sup> Department of Computer Science and Engineering, Dankook University, Yongin 16890, Korea

\* Correspondence: choisi@dankook.ac.kr; Tel.: +82-31-8005-3657

Received: 22 August 2020; Accepted: 21 September 2020; Published: 23 September 2020



**Abstract:** We propose a method of enlarging the training dataset for a single-sample-per-person (SSPP) face recognition problem. The appearance of the human face varies greatly, owing to various intrinsic and extrinsic factors. In order to build a face recognition system that can operate robustly in an uncontrolled, real environment, it is necessary for the algorithm to learn various images of the same person. However, owing to limitations in the collection of facial image data, only one sample can typically be obtained, causing difficulties in the performance and usability of the method. This paper proposes a method that analyzes the changes in pixels in face images associated with variations by extracting the binary weighted interpolation map (B-WIM) from neutral and variational images in the auxiliary set. Then, a new variational image for the query image is created by combining the given query (neutral) image and the variational image of the auxiliary set based on the B-WIM. As a result of performing facial recognition comparison experiments on SSPP training data for various facial-image databases, the proposed method shows superior performance compared with other methods.

**Keywords:** image generation; weighted interpolation map; binarization; single sample per person

## 1. Introduction

Face recognition technology is used to identify individuals from their captured facial images by leveraging a labeled database containing people's identities. Compared with other types of biometric recognition, face recognition is less invasive and does not require a subject to be in proximity to or in contact with a sensor, making the method widely applicable to user identification, e-commerce, access control, surveillance, and human–computer interaction. However, because variations caused by extrinsic factors (e.g., illumination and pose) and intrinsic factors (e.g., facial expression, age, and accessories) are very large, it is difficult to robustly recognize a face under uncontrolled conditions [1,2]. To deal with these variations, facial recognition methods have been studied under the assumption that several images can be made available for each person, and high-performance methods have been built using vast databases of this nature (e.g., VGGface2 [3], Tufts Face [4], UMDfaces [5]), MegaFace [6], and LFW [7,8] databases).

However, for many large-scale face recognition applications (e.g., passport authentication, drivers' license identification, and police investigations), the training data required to learn algorithms do not offer many samples per person. In many cases, there is only a single sample per person (SSPP) available [9,10]. For example, law enforcement agencies have constructed databases of facial images (i.e., mug-shots) for decades. The related datasets comprise frontal face images under steady illumination and blank expressions. However, owing to cost and privacy issues, these databases are rarely augmented with extra multi-conditional candid photos. Furthermore, it is known that

criminals usually attempt to disguise their identities when committing a crime [11,12]. Even if they do not, it remains very difficult for systems to match active faces with the collected neutral images. As such, the dearth of learnable data restricts the use of feature-extraction and various other supervised methods [13–16].

To solve the SSPP problem, several methods of enlarging training datasets have been proposed to generate new images from a given one. The theory of the evolution of technology suggests that such datasets can be expanded using existing means [17–20]. In  $E(PC)^2A2+$  [21], extended from  $(PC)^2A$  [22], an image and its corresponding half-, first-, and second-order projected images were used as the training set. In the  $(2D)^2PCA$  method [23], new images were generated by simultaneously applying two-directional principal component analysis (PCA) [24] in the row and column directions of 2D images. In the SPCA+ method [25], the training set was enlarged by combining the original image linearly with its derived image obtained by perturbing the image matrix's singular values. In [26], concatenated left- and right-side images obtained from the symmetry of the face were used as training samples. In [27], images were generated using a symmetry transform for the intraclass and a linear combination for the interclass. In the interclass relationship (ICR) [28], data were generated by a weighted combination of (at least) two images in the training set. The ICR rectified the underestimated intraclass and overestimated the interclass. In MVI [29], the training set was enlarged by generating multiple low-resolution virtual images using a single high-resolution image. In SRGES [30], images were generated by adding mean images of the difference between neutral and variational images for each variation in the auxiliary set to the query image. In [31], occluded images were generated by using a weighted interpolation map and an auxiliary set. The weighted interpolation map represented the degrees of changes in pixels at the same positions between an image and its occluded version. The degree of changes was measured using the standard deviation of the difference between neutral and variational images in the auxiliary set. When generating a new image, the pixels at positions of large differences were replaced with the pixel values of the average image of the occluded images in the auxiliary set.

In this paper, we propose binary weighted interpolation maps (B-WIM) to enlarge the training set for face recognition. Generally, the occurrence of variations leads the local pixels to change in the face image. Supposing it is possible to grasp the change in local pixels between the original and varied images, it then becomes possible to capture the characteristics of the image changes caused by the variation. By analyzing these characteristics, the proposed method can maintain most of the characteristics of the neutral image while replacing only the changed areas with the pixel values of the variational one. For this, we first construct an auxiliary set consisting of neutral images and their variational images. Then, the normalized weighted interpolation map is extracted by using the log-scaled standard deviation of the absolute difference between the neutral images and the corresponding variational images in the auxiliary set. Each element of the weighted interpolation map reflects the degree of change caused by the variation in individual pixels, and a B-WIM is obtained via binarization.

When generating a new image for a given query (neutral) image, the variational image corresponding to the neutral image having the highest correlation with the query image is selected from the auxiliary set. Then, a new image is generated by combining the query and selected variational images. The overall procedure of the proposed method is shown in Figure 1.

The idea for the proposed method was motivated by the ICR concept and the weighted interpolation map method, which are face-generating frameworks. However, unlike ICR, by simply increasing the number of images by the weight combinations of two images, the proposed method has the advantage of generating a natural image with a specific variation. Additionally, the proposed method creates an image of higher quality than the weighted interpolation map (WIM) method by selecting the neutral image and the variational image corresponding to the query image.

Face recognition experiments are evaluated using the following criteria. First, we measure the change in the face recognition rate according to the degree of variation of different databases. Second,



the face recognition performance is analyzed using unsupervised and supervised learning methods. Finally, the overall face recognition rates of all methods are assessed. We compare the proposed method with other methods dealing with the SSPP problem: WIM, ICR, E(PC<sup>2</sup>)A+, SPCA+, (2D)<sup>2</sup>PCA, SLC, MVI, and SRGES. The results of the experiment show that the proposed method exhibits high face recognition performance for all criteria.

The remainder of this paper is organized as follows. Section 2 explains the proposed method for generating data and describes each procedure in detail. The experimental face recognition results are described in Section 3, and the discussion and conclusion follow.

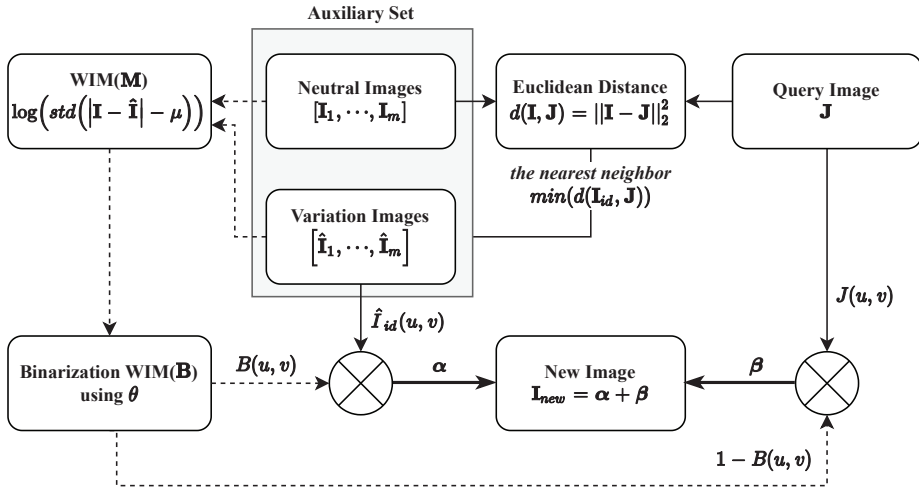


Figure 1. Overall procedure of the proposed method.

## 2. Proposed Method

Using ICR [28], a new image,  $I_{new}$ , is generated using a weighted combination of neutral images,  $I_i$  and  $I_j$ , as

$$I_{new} = (1 - \lambda) \cdot I_i + \lambda \cdot I_j, \quad \text{where } 0 \leq \lambda \leq 1 \quad (1)$$

In Equation (1), the weight,  $\lambda$ , decides the ratio of  $I_i$  and  $I_j$  to be reflected in  $I_{new}$ . If  $\lambda$  is 0.5, the two images are assumed to have been reflected equally. In the case of ICR, it is easy to generate images by applying a single parameter for all pixels. However, the changes in some areas within the image will not be reflected accurately, owing to variations. Extrinsic factors, such as occlusions, cause changes in the neutral image. For example, pixels around the eyes change significantly when wearing sunglasses. On the other hand, areas unrelated to these variations generally retain the pixel information of the neutral image. Therefore, it is necessary to obtain the weights for each pixel.

In this paper, we propose a method to enlarge the training set. A new image,  $I_{new}$ , with variations, is generated via a combination of the neutral image,  $I$ , and the variational image,  $\hat{I}$ , derived from  $I$  by referring to the ICR. The method of generating a new image is redefined as follows:

$$I_{new}(u, v) = (1 - B(u, v))I(u, v) + B(u, v)\hat{I}(u, v). \quad (2)$$

In Equation (2),  $B(u, v)$  is the pixel weight at a certain position in both  $I$  and  $\hat{I}$ . A new image,  $I_{new}$ , is generated, containing the variations only in some areas while maintaining the characteristics of the neutral image as much as possible.

2.1. Binary Weighted Interpolation Maps (B-WIM)

The occurrence of variations changes the pixel values of the neutral image. Figure 2 shows the difference between  $I$  and  $\hat{I}$ , caused by facial expression variations. In the aligned face images, the area around the mouth where the smile occurs significantly changes the pixel value of  $I$ .

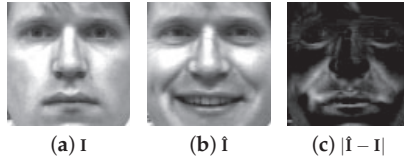


Figure 2. (a) Neutral image; (b) variational image; (c) absolute difference between neutral and variational image.

Absolute difference is the difference between  $(I \cup \hat{I})$  and  $(I \cap \hat{I})$ . The standard deviation represents the degree of statistical variation from the difference between the pixels of these images

$$\mathcal{M}(u, v) = \log \left( \frac{1}{m-1} \sqrt{\sum_{i=1}^m (|\hat{I}_i(u, v) - I_i(u, v)| - \mu(u, v))^2} \right), \tag{3}$$

$$\mu(u, v) = \frac{1}{m} \sum_{i=1}^m |\hat{I}_i(u, v) - I_i(u, v)|$$

where subscripts  $i(= 1, 2, \dots, m)$  denote the  $i$ th individual. In Equation (3), when the value is very large according to the degree of change, some pixels are saturated in the generated image. Therefore, the normalized  $M$  is calculated as follows:

$$M = \frac{M - \min(M)}{\max(M) - \min(M)}. \tag{4}$$

Figure 3a–f shows each  $M$  for facial expressions, such as angry, afraid, disgusted, sad, smiling, and surprised. Facial expressions are related to the activation of a distinct set of facial muscles [32,33]. When smiling, pixels around the mouth, which are related to the levator anguli oris muscle, change significantly when compared with the neutral image [34,35].

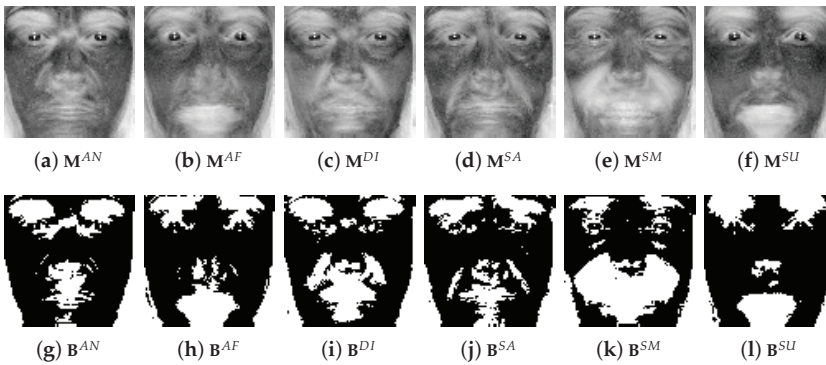


Figure 3. Weighted interpolation maps ( $M$ ) and binary weighted interpolation maps ( $B$ ) for facial expressions. (AN: angry, AF: afraid, DI: disgusted, SA: sad, SM: smiling, SU: surprised).

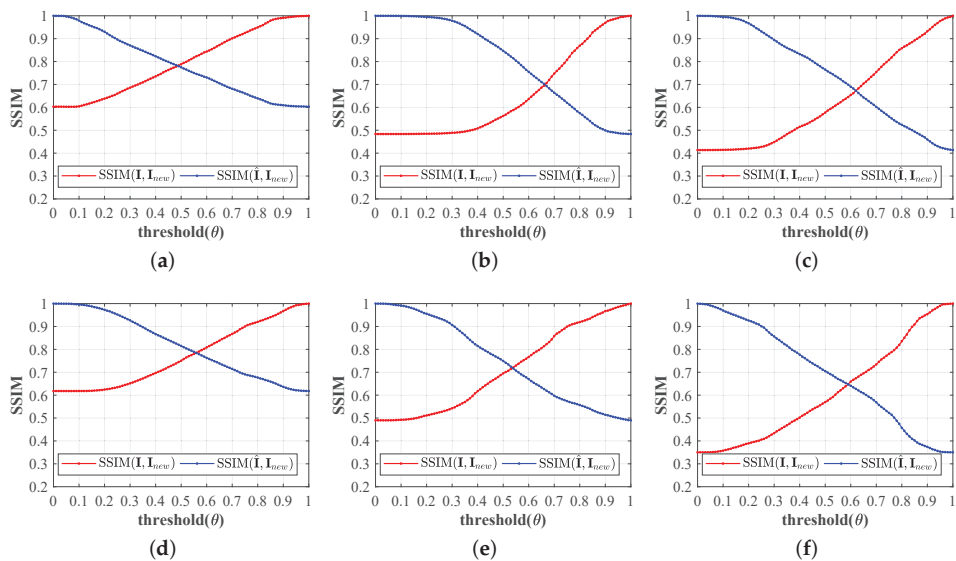
In a previous work [31], the WIM was extracted to measure the degree of change between neutral and variational images in pixels, and the query image and mean-variational image from the WIM were combined. However, the WIM has a problem in that the weight values of the locations associated with variations may be relatively lower in the normalization process when the maximum value obtained from Equation (3) is too large. If  $\mathcal{M}(u, v)$  is 0.5, the location where the variation has occurred is not replaced by the pixel value of  $\hat{I}(u, v)$ . In this case, the pixel values for  $I(u, v)$  and  $\hat{I}(u, v)$  will be mixed equally at  $I_{new}(u, v)$ . This is a type of noise. To overcome this, we define B-WIMs **B** as follows:

$$\mathcal{B}(u, v) = \begin{cases} 1 & \mathcal{M}(u, v) > \theta \\ 0 & \mathcal{M}(u, v) \leq \theta \end{cases} \quad (5)$$

In Equation (5), depending on the threshold, ( $\theta$ ),  $\mathcal{B}(u, v)$  has a logical value of 0 or 1 (Figure 3g–l). The pixel value of  $I(u, v)$  is fully reflected when the value of  $\mathcal{B}(u, v)$  is 0. Conversely, if  $\mathcal{B}(u, v)$  becomes 1, the pixel value of  $I_{new}(u, v)$  will completely replace the pixel value of  $\hat{I}(u, v)$ . Accordingly, the WIM problem can be solved.

We use the structural similarity (SSIM) index [36] to find the optimal  $\theta$ . The SSIM index evaluates a distorted image with respect to a reference image to quantify their structural similarity [37].

If  $\theta$  is 0, all elements of **B** have a value of 1. Thus,  $I_{new}$  is obtained from Equation (2), which becomes the same as the variational image ( $\hat{I}$ ) in the auxiliary set. However, if  $\theta$  is 1, the  $I_{new}$  is identical to the query image (**J**). To find the value of  $\theta$  to generate a new image in which the variation is reflected in a balanced manner while maintaining the unique identity of the query image, we investigate  $SSIM(I, I_{new})$ ,  $SSIM(\hat{I}, I_{new})$  of **I**, and  $\hat{I}$  in the auxiliary set by increasing  $\theta$  from 0 to 1, respectively. As shown in Figure 4, two SSIMs are balanced when  $\theta$  is between 0.5 and 0.7. Thus, we set  $\theta$  from 0.5 to 0.7, depending on the type of variation.



**Figure 4.** Structural similarity (SSIM) results for each variation (a) angry; (b) afraid; (c) disgusted; (d) sad; (e) smiling; (f) surprised.

Figure 5 shows the image samples generated by applying the B-WIM constructed from  $\theta$  for the “smiling” variation for images in the auxiliary set. It is visually confirmed that the variations in the generated image are included while  $\theta$  is less than 0.7.

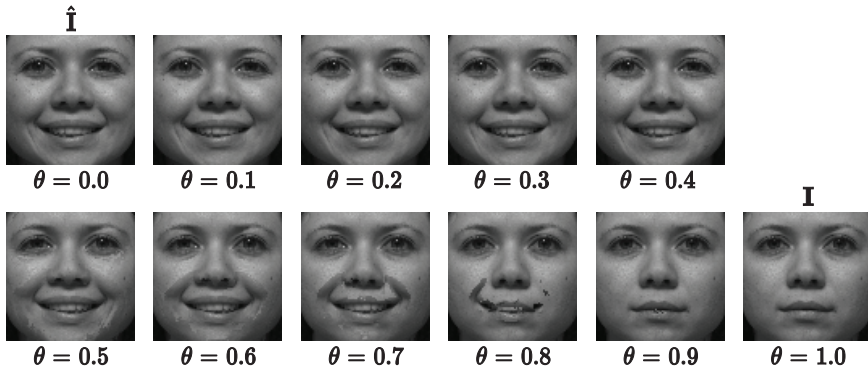


Figure 5. Generated images from  $I$  and  $\hat{I}$  for each threshold ( $\theta$ ).

2.2. Generation of New Images from a Query Image

The new image ( $I_{new}$ ) can be generated from a query image ( $J$ ) as follows:

$$I_{new}(u, v) = (1 - \mathcal{B}(u, v))J(u, v) + \mathcal{B}(u, v)\hat{J}(u, v). \tag{6}$$

Unlike during the phase of B-WIM extraction, a variational image ( $\hat{J}$ ) derived from  $J$  cannot be obtained in the phase of image generation. Therefore,  $\hat{J}$  must be replaced with another image in a separate auxiliary set. In WIM, the mean image for each variation in the auxiliary set is used as  $\hat{J}$ . Although it can be applied equally to all query images, morphological elements may be lost if the variation’s own changes are large. For example, mufflers can be worn in various ways depending on a person’s personality. Moreover, the designs of mufflers are also very diverse. Therefore, the mean image cannot preserve the form of all mufflers.

In this study, we select the neutral image (i.e., nearest neighbor) of the auxiliary set with a minimum Euclidean distance (L2-norm) from the query image based on the whole pixel [28].

$$d(I, J) = \|I - J\|_2^2 \tag{7}$$

Then,  $\hat{J}$  is replaced by  $\hat{I}_{id}$ , derived from  $I_{id}$ , where  $id$  is the index with a minimum distance from  $J(\min(d(I, J)))$ . Equation (6) is redefined as

$$I_{new}(u, v) \approx (1 - \mathcal{B}(u, v))J(u, v) + \mathcal{B}(u, v)\hat{I}_{id}(u, v). \tag{8}$$

Finally,  $I_{new}$  is generated from Equation (8).

The overall procedure of the proposed method is summarized as follows:

- **Step 1:** Extraction of the normalized WIM using log-scaled standard deviation of the absolute difference between  $I$  and  $\hat{I}$  in the auxiliary set;
- **Step 2:** Binarization of WIM from threshold ( $\theta$ ) ;
- **Step 3:** Selection of the index ( $id$ ) of the nearest neighbor in the auxiliary set based on Euclidean distance with the query image; ( $\min(\|I - J\|_2^2)$ )
- **Step 4:** Replacement of  $\hat{J}$  with  $\hat{I}_{id}$ , derived from  $I_{id}$ ;
- **Step 5:** Generation of the new image ( $I_{new}$ ).

### 3. Experiments

#### 3.1. Database

In the experiment, all images were aligned to  $80 \times 80$  pixels via affine transformation based on manually detected eye coordinates. Then, the image was compensated using histogram equalization [38].

We used the Bosphorus [32] and RaFD [39] databases in our face recognition experiments (Table 1). The Bosphorus database comprises images captured under seven different facial expression conditions from 58 subjects and includes various expressions, such as neutral, angry, disgusted, afraid, sad, smiling, and surprised. The neutral images (“indexed 5”) were selected to generate new images, and the remaining images were used for the face recognition test. The RaFD database contains 536 images captured from 67 subjects. Each subject provided images of eight facial expressions (i.e., neutral, angry, contemptuous, disgusted, afraid, sad, smiling, and surprised). We used a neutral image (“indexed 6”) to generate new images, and the remaining facial expression images were used for testing. Both databases applied practiced expressions using a Facial Action Coding System (FACS) [33] specialist. Furthermore, all subjects were tightly controlled through negative feedback to acquire the required activation of action units (AU).

**Table 1.** Characteristics of each facial expression database.

Expression	Bosphorus	RaFD
Neutral	○	○
Afraid	○	○
Angry	○	○
Disgusted	○	○
Sad	○	○
Smiling (smiling)	○	○
Surprise (scream)	○	○
Contemptuous		○
No. of subjects	58	67
No. of images per subject	7	8
Index of neutral face images	5	6
No. of testing images	348	469

#### 3.2. Face Recognition Results

We compared the proposed method with other methods dealing with the SSPP problem (i.e., WIM, ICR,  $E(PC^2)A+$ ,  $SPCA+$ ,  $(2D)^2PCA$ , SLC, MVI, and SRGES methods). The proposed method, WIM, and SRGES generated as many images as the number of variations contained in the auxiliary set. With the ICR, the number of generated images depended on the  $k$  neighbors in the training set and the feature extraction method used by each database. In  $E(PC^2)A+$ , the half-, first-, and second-order projected images of each neutral image were used as the training set. In  $SPCA+$ , seven images were enlarged from different  $n$ -order singular values for each neutral image in the training set. In  $(2D)^2PCA$ , an image was generated using a two-directional PCA in the row and column directions of the 2D images. In SLC, 11 images from the neutral image were added to the training set, which included symmetric images and linear combinations of virtual images. In MVI, four low-resolution images (size  $40 \times 40$ ,  $26 \times 26$ ,  $20 \times 20$ , and  $16 \times 16$ ) are generated from the neutral image and various scaling factors (i.e., 2, 3, 4, 5, respectively).

In this study, the face recognition performance of all methods was evaluated based on the given criteria. First, we measured the change in the face recognition rate according to the degree of variation in each database. Both databases contained similar facial expression variations, but they differed in the intensity of the facial expressions. In the Bosphorus database, the AUs were captured at their given peak intensity levels. In the RaFD database, there were large deviations in the intensities of expressions according to subject. Thus, the RaFD database was closer to the real

world than the Bosphorus one. Second, the face recognition performance was analyzed according to unsupervised and supervised learning methods. An unsupervised learning-based PCA [24] and supervised learning-based discriminant common vector (DCV) [40] were used to extract the features for face recognition. PCA extracted  $(\mathcal{N} + \mathcal{N}' - 1)$  features, including the number of images of training data ( $\mathcal{N}$ ) and the number of enlarged images ( $\mathcal{N}'$ ) from itself. DCV extracted  $(c - 1)$  features, where  $c$  was the total number of classes, regardless of the number of images. In the face recognition experiment, the recognition rates were measured using the maximum number of features extracted from each method. If a given set had been modeled properly, it could be expected to show high performance, regardless of the two methods. When evaluating the face recognition performance, the one nearest-neighbor rule was used with the  $l_2$  norm as a classifier.

In this study, two protocols for face recognition were used [41]: “Closed Set” and “Open Set”, according to the auxiliary set. These are described as follows:

- *Closed set*: In this case, all images were collected under similar conditions. Thus, all images belonged to the same database. In the experiment, the database was divided into a face recognition set and an auxiliary set. The face recognition set consisted of training and test sets. Neutral images for each class used to generate images were included in the training set, and the remaining images containing only variations were used as the test set. This method had the same variations (“expression”) in both face recognition and auxiliary sets;
- *Open set*: This case used a separate auxiliary set from a given database to demonstrate the superiority of the proposed method. The training and test sets were collected under similar conditions. However, the auxiliary set was taken in environments different from those. The face recognition set was constructed in the same way as in the “Closed Set” case, and neutral images were used to enlarge the others. Both face recognition and auxiliary sets included “expression” variations. However, the types of detailed variations could be different.

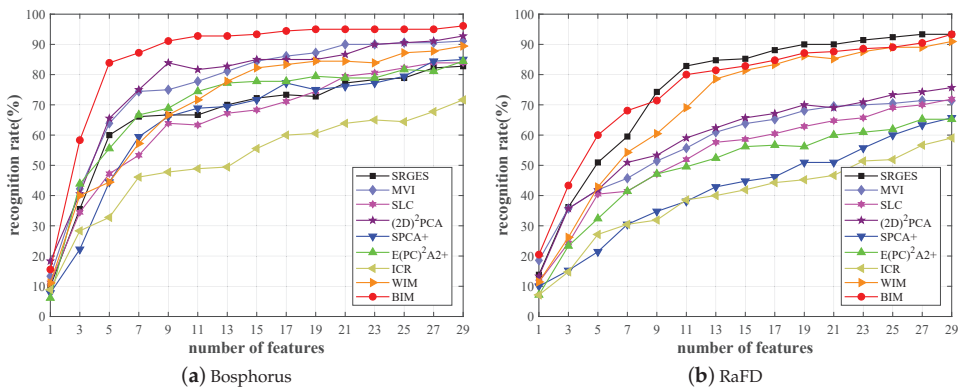
First, we divided the given databases into face recognition and auxiliary sets. A total of 30 subjects from all subjects in each database were used for the face recognition set, and the others were used for the auxiliary set. Among the 210 and 240 images for 30 subjects in the Bosphorus and RaFD databases, respectively, neutral images were used as training data to construct PCA and DCV feature spaces for face recognition, and variational images were generated using the proposed methods from these images to enlarge the training set.

Table 2 shows the face recognition results for the “Closed Set” protocol. In the experimental results, the proposed method, WIM, and  $(2D)^2$ PCA presented similar facial-recognition results within the same database, regardless of the feature-extraction manner. The other methods had a face recognition rate difference of up to 21.11% between each manner. Additionally, the proposed method and WIM showed high performance in the face recognition results according to the degree of variation by database. For the rest of the methods, the face recognition performance decreased by more than 15.87~33.17% as the degree of variation increased. Figure 6 shows the recognition rates for a different number of DCV features. The proposed method gives a recognition rate of 96.11% and 93.33%, with 29 features for the Bosphorus and RaFD databases, respectively. As can be seen from Figure 6, the proposed method shows a comparable or better recognition performance to the other methods, regardless of the number of features. Finally, we confirmed that the proposed method was excellent in the absolute comparison of face recognition rates for both databases. Because the proposed method consistently showed high face recognition performance regardless of the various criteria, it could be inferred that a new image was generated by reflecting the various variations from the given neutral images.

**Table 2.** Face recognition results in *Closed Set* protocol.

Method	Bosphorus		RaFD	
	PCA	DCV	PCA	DCV
B-WIM	95.56%	96.11%	90.48%	93.33%
WIM [31]	91.11%	89.44%	84.29%	90.95%
ICR [28]	93.89%	71.67%	76.67%	59.05%
E(PC) <sup>2</sup> A2+ [21]	92.78%	84.44%	65.71%	65.24%
SPCA+ [25]	88.89%	85.00%	55.71%	65.71%
(2D) <sup>2</sup> PCA [26]	91.11%	92.78%	76.19%	75.71%
SLC [27]	91.11%	83.89%	75.24%	71.90%
MVI [29]	92.78%	90.56%	74.76%	75.24%
SRGES [30]	92.22%	82.78%	77.62%	93.33%

Generally, the basic emotion group consists of angry, disgusted, afraid, sad, smiling, and surprised [42]. For the “*Open Set*” protocol, the auxiliary set containing six defined facial expressions consists of the AR [43], CK+ [44], Jaffe [45], PF07 [46], and Yale [47] databases. Additionally, it included various races and genders. To measure the degree of change from the variations, only subjects without glasses (occlusion) were used to construct the auxiliary set. The selected subjects had images of both neutral and facial expressions.



**Figure 6.** Face recognition results for a various number of DCV features in “*Closed Set*” protocol.

The AR database contained images from 85 subjects (37 males and 48 females) of different races [43]. We selected four facial expressions: neutral, angry, smiling, and screaming. The CK+ database contained 84 subjects from many different races. Image sequences contained changes in facial expressions over time. The neutral image at the start time and the facial expression image at the end time comprised the auxiliary set. We used seven facial expressions (i.e., neutral, angry, afraid, disgusted, sad, smiling, and surprised), except for “contemptuous.” The Jaffe database included 10 subjects (only females) of Asian ethnic groups. Seven facial expressions of each subject were taken (e.g., neutral, angry, afraid, disgusted, sad, smiling, and surprised). The PF07 database contained the images of 200 subjects (100 males and 100 females) of Asian ethnic groups, all of whom provided four images with different facial expression conditions (i.e., neutral, angry, smiling, and surprised). The Yale database included 15 subjects (14 males and a female) of many different races. We used four facial expressions (i.e., neutral, sad, smiling, and surprised), excluding those with eyes closed or winking (Table 3).



**Table 3.** Facial expressions in each database in the auxiliary set.

Expressions	AR	CK+	Jaffe	PF07	Yale
Neutral	○	○	○	○	○
Afraid		○	○		
Angry	○	○	○	○	
Disgusted		○	○		
Sad		○	○		○
Smiling (smiling)	○	○	○	○	○
Surprised (screaming)	○	○	○	○	○

Table 4 shows the face recognition results with a separate auxiliary set. Because the auxiliary set was constructed from separate databases, the face recognition experiment used images of all the subjects contained in each database.

**Table 4.** Face recognition results in *Open Set* protocol.

Method	Bosphorus		RaFD	
	PCA	DCV	PCA	DCV
B-WIM	<b>87.64%</b>	<b>88.51%</b>	<b>81.88%</b>	<b>87.21%</b>
WIM [31]	86.49%	75.29%	76.12%	82.30%
ICR [28]	83.91%	66.67%	66.52%	49.89%
E(PC) <sup>2</sup> A2+ [21]	83.62%	72.70%	58.85%	61.19%
SPCA+ [25]	78.16%	76.72%	47.33%	55.86%
(2D) <sup>2</sup> PCA [26]	81.32%	82.76%	67.59%	67.38%
SLC [27]	82.18%	76.15%	71.86%	65.46%
MVI [29]	82.76%	83.05%	65.88%	69.08%
SRGES [30]	81.90%	83.91%	73.13%	82.73%

Depending on the degree of variation, the differences in face recognition rates were measured in the order of the proposed method: (5.77%), SRGES (8.76%), WIM (10.37%), SLC (10.69%), (2D)<sup>2</sup>PCA (15.38%), MVI (16.87%), ICR (17.38%), E(PC)<sup>2</sup>A2+ (24.77%), and SPCA+ (30.83%). For the criteria, the proposed method and SLC maintained high performance, whereas the remaining methods showed differences in face recognition performance. Generally, the proposed method showed the highest face recognition rates. Figure 7 shows the recognition rates for a different number of DCV features. The proposed method gives a recognition rate of 88.51% with 57 features and 87.21% with 66 features for the Bosphorus and RaFD databases, respectively. As can be seen from Figure 7, the proposed method shows the best recognition performance compared to the other methods for all number of features. This experiment also confirmed the superiority of the proposed method for each criterion.

On the other hand, from the results of Tables 2 and 4, it can be seen that the recognition rate in the “*Closed Set*” protocol was about 10% higher than that in the “*Open Set*” protocol. It is generally known that face recognition performance decreases as the number of subjects to be recognized increases [48]. In our experiment, however, we think the main reason for the difference between the results of Tables 2 and 4 is that, in the “*Closed Set*” protocol, the images included in the auxiliaries set had homogeneous characteristics, because they were taken under similar conditions of resolution, camera type, lighting conditions, etc. However, in the “*Open Set*” protocol, the auxiliary set comprised images from various kinds of databases, which differ from the query images.

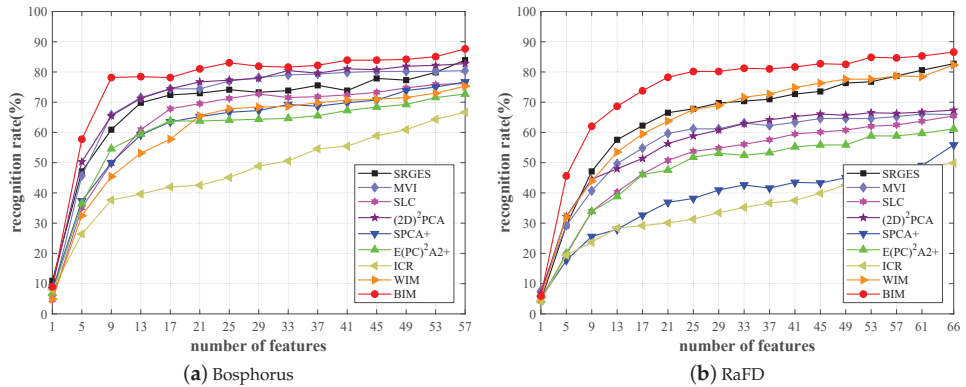


Figure 7. Face recognition results for a various number of DCV features in “Open Set” protocol.

#### 4. Discussion and Conclusions

Building a face recognition system that works robustly in various environments involves difficulties in securing the data needed to learn recognition algorithms. Moreover, large-scale face recognition applications typically use databases that contain SSPPs. A single image is not sufficiently representative for face recognition. The SSPP problem makes using the feature extraction method in a supervised manner quite difficult, because the interclass variations are unknown. To overcome this, several methods have been proposed. However, there have been limitations in that these methods did not reflect facial characteristics that could have various variations.

We proposed an image generation method that uses a B-WIM that leverages the fact that the pixels of specific parts of the neutral face image vary significantly compared with other areas when there is an environmental variation in face recognition. The B-WIM statistically reflects the change in individual pixel values caused by the variation from the neutral and variational images included in the auxiliary set. For a given query image (neutral image), the proposed method creates a new variational image that reflects the characteristics of the variation while maintaining the unique characteristics of the face in the query image based on B-WIM. Through this, a training dataset containing only one sample per person can be made into a richer set that includes variational images for each person, further improving the performance of the face recognition system.

The proposed method has the following advantages. The proposed method does not require a large amount of computation or a large dataset for creating new images. When the number of pixels in an image is  $n$ , while SPCA+ has the complexity of  $O(n^2)$ , the complexity of the proposed method is  $O(n)$ . Some methods, such as ICR, E(PC²)A+, (2D)²PCA, SLC, and MVI, require similar computations as the proposed method but do not address specific variations. In contrast, the proposed method generates high-quality variational images for query images in real-time, effectively improving the performance of existing face recognition systems at a low cost. Face recognition experiments using Bosphorus and RaFD databases showed that the proposed method outperformed the existing methods for solving the SSPP problem. In addition to general facial recognition algorithms, images generated using the proposed method can be utilized in the study of various facial images, including the fake image detection algorithms [49,50].

On the other hand, by comparing the recognition rates for two protocols of face recognition, “Closed Set” and “Open Set,” we found that the quality of the image created using the proposed method was affected by the images included in the auxiliary set. Although the proposed method can effectively generate new images for a specific variation, it does not control the degree of variation or handle more than two variations simultaneously. It is expected that the small sample-size problems, including the SSPP problem, can be solved more effectively by subdividing the degree of variation

within the proposed method's algorithmic structure and applying the interpolation maps for two or more types of variations together. We leave these problems to future works.

**Author Contributions:** Y.L. and S.-I.C. designed the experiments and drafted the manuscript. All authors have read and agreed to the published version of the manuscript.

**Funding:** The present research was supported by a National Research Foundation of Korea(NRF) grant funded by the Korean government (MSIT) (No. 2018R1A2B6001400) and the MSIT(Ministry of Science and ICT), Korea, under the ICAN(ICT Challenge and Advanced Network of HRD) program(IITP-2020-2020-0-01824) supervised by the IITP(Institute of Information & Communications Technology Planning & Evaluation).

**Conflicts of Interest:** The authors declare no conflict of interest.

## Abbreviations

AUs	action units
B-WIM	binary weighted interpolation maps
DCV	discriminant common vector
FACS	facial action coding system
ICR	interclass relationship
PCA	principal component analysis
SSIM	structural similarity
SSPP	single sample per person
WIM	weighted interpolation maps

## References

1. Kortli, Y.; Jridi, M.; Al Falou, A.; Atri, M. Face recognition systems: A Survey. *Sensors* **2020**, *20*, 342. [CrossRef]
2. Choi, S.I.; Lee, Y.; Lee, M. Face Recognition in SSPP Problem Using Face Relighting Based on Coupled Bilinear Model. *Sensors* **2019**, *19*, 43. [CrossRef]
3. Cao, Q.; Shen, L.; Xie, W.; Parkhi, O.M.; Zisserman, A. Vggface2: A dataset for recognising faces across pose and age. In Proceedings of the 2018 13th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018), Xi'an, China, 15–19 May 2018; pp. 67–74.
4. Panetta, K.; Wan, Q.; Aghaian, S.; Rajeev, S.; Kamath, S.; Rajendran, R.; Rao, S.; Kaszowska, A.; Taylor, H.; Samani, A.; et al. A comprehensive database for benchmarking imaging systems. *IEEE Trans. Pattern Anal. Mach. Intell.* **2018**, *42*, 509–520. [CrossRef]
5. Bansal, A.; Nanduri, A.; Castillo, C.D.; Ranjan, R.; Chellappa, R. Umdfaces: An annotated face dataset for training deep networks. In Proceedings of the 2017 IEEE International Joint Conference on Biometrics (IJCB), Denver, CO, USA, 1–4 October 2017; pp. 464–473.
6. Kemelmacher-Shlizerman, I.; Seitz, S.M.; Miller, D.; Brossard, E. The megaface benchmark: 1 million faces for recognition at scale. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 4873–4882.
7. Huang, G.B.; Mattar, M.; Berg, T.; Learned-Miller, E. Labeled Faces in the Wild: A Database for Studying Face Recognition in Unconstrained Environments. 2008. Available online: <http://vis-www.cs.umass.edu/lfw> (accessed on 1 September 2020).
8. Huang, G.B.; Learned-Miller, E. *Labeled Faces in the Wild: Updates and New Reporting Procedures*; Technical Report UM-CS-2014-003; Department of Computer Science, University of Massachusetts Amherst: Amherst, MA, USA, 2014.
9. Tan, X.; Chen, S.; Zhou, Z.H.; Zhang, F. Face recognition from a single image per person: A survey. *Pattern Recognit.* **2006**, *39*, 1725–1745. [CrossRef]
10. Ríos-Sánchez, B.; Costa-da Silva, D.; Martín-Yuste, N.; Sánchez-Ávila, C. Deep Learning for Facial Recognition on Single Sample per Person Scenarios with Varied Capturing Conditions. *Appl. Sci.* **2019**, *9*, 5474. [CrossRef]
11. Noyes, E.; Jenkins, R. Deliberate disguise in face identification. *J. Exp. Psychol. Appl.* **2019**, *25*, 280. [CrossRef]
12. Demleitner, N.V. Witness Protection in Criminal Cases: Anonymity, Disguise or Other Options? *Am. J. Comp. Law* **1998**, *46*, 641–664. [CrossRef]

13. Wang, F.; Cheng, J.; Liu, W.; Liu, H. Additive margin softmax for face verification. *IEEE Signal Process. Lett.* **2018**, *25*, 926–930. [CrossRef]
14. Deng, J.; Guo, J.; Xue, N.; Zafeiriou, S. Arcface: Additive angular margin loss for deep face recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15–20 June 2019; pp. 4690–4699.
15. Wang, H.; Wang, Y.; Zhou, Z.; Ji, X.; Gong, D.; Zhou, J.; Li, Z.; Liu, W. Cosface: Large margin cosine loss for deep face recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–22 June 2018; pp. 5265–5274.
16. Zheng, Y.; Pal, D.K.; Savvides, M. Ring loss: Convex feature normalization for face recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–22 June 2018; pp. 5089–5097.
17. Coccia, M.; Watts, J. A theory of the evolution of technology: Technological parasitism and the implications for innovation management. *J. Eng. Technol. Manag.* **2020**, *55*, 101552. [CrossRef]
18. Coccia, M. Sources of technological innovation: Radical and incremental innovation problem-driven to support competitive advantage of firms. *Technol. Anal. Strateg. Manag.* **2017**, *29*, 1048–1061. [CrossRef]
19. Arthur, W.B. *The Nature of Technology: What It Is and How It Evolves*; Simon and Schuster: New York City, NY, USA, 2009.
20. Arthur, W.B.; Polak, W. The evolution of technology within a simple computer model. *Complexity* **2006**, *11*, 23–31. [CrossRef]
21. Chen, S.; Zhang, D.; Zhou, Z.H. Enhanced  $(PC)^2A$  for face recognition with one training image per person. *Pattern Recognit. Lett.* **2004**, *25*, 1173–1181. [CrossRef]
22. Wu, J.; Zhou, Z.H. Face recognition with one training image per person. *Pattern Recognit. Lett.* **2002**, *23*, 1711–1719. [CrossRef]
23. Zhang, D.; Zhou, Z.H.  $(2D)^2PCA$ : Two-directional two-dimensional PCA for efficient face representation and recognition. *Neurocomputing* **2005**, *69*, 224–231. [CrossRef]
24. Turk, M.; Pentland, A. Eigenfaces for recognition. *J. Cogn. Neurosci.* **1991**, *3*, 71–86. [CrossRef]
25. Zhang, D.; Chen, S.; Zhou, Z.H. A new face recognition method based on SVD perturbation for single example image per person. *Appl. Math. Comput.* **2005**, *163*, 895–907. [CrossRef]
26. Xu, Y.; Zhu, X.; Li, Z.; Liu, G.; Lu, Y.; Liu, H. Using the original and ‘symmetrical face’ training samples to perform representation based two-step face recognition. *Pattern Recognit.* **2013**, *46*, 1151–1158. [CrossRef]
27. Zhang, T.; Li, X.; Guo, R.Z. Producing virtual face images for single sample face recognition. *Opt.-Int. J. Light Electron Opt.* **2014**, *125*, 5017–5024. [CrossRef]
28. Li, Q.; Wang, H.J.; You, J.; Li, Z.M.; Li, J.X. Enlarge the training set based on inter-class relationship for face recognition from one image per person. *PLoS ONE* **2013**, *8*, e68539. [CrossRef]
29. Moon, H.M.; Kim, M.G.; Shin, J.H.; Pan, S.B. Multiresolution face recognition through virtual faces generation using a single image for one person. *Wirel. Commun. Mob. Comput.* **2018**, *2018*. [CrossRef]
30. Ding, Y.; Qi, L.; Tie, Y.; Liang, C.; Wang, Z. Single sample per person face recognition based on sparse representation with extended generic set. In Proceedings of the 2018 International Conference on Cyber-Enabled Distributed Computing and Knowledge Discovery (CyberC), Zhengzhou, China, 18–20 October 2018; pp. 37–375.
31. Lee, Y.; Kang, J. Occlusion Images Generation from Occlusion-Free Images for Criminals Identification based on Artificial Intelligence Using Image. *Int. J. Eng. Technol.* **2018**, *7*, 161–164.
32. Savran, A.; Alyüz, N.; Dibeklioglu, H.; Çeliktutan, O.; Gökberk, B.; Sankur, B.; Akarun, L. Bosphorus database for 3D face analysis. In *European Workshop on Biometrics and Identity Management*; Springer: Berlin/Heidelberg, Germany, 2008; pp. 47–56.
33. Friesen, E.; Ekman, P. *Facial Action Coding System: A Technique for the Measurement of Facial Movement*; Consulting Psychologists Press: City of Palo Alto, CA, USA, 1978.
34. Scheve, T. How Many Muscles Does It Take to Smile? How Stuff Works Science. June 2009; Volume 2. Available online: <https://science.howstuffworks.com/life/inside-the-mind/emotions/muscles-smile.htm> (accessed on 1 September 2020).
35. Waller, B.M.; Cray, J.J., Jr.; Burrows, A.M. Selection for universal facial emotion. *Emotion* **2008**, *8*, 435. [CrossRef]

36. Wang, Z.; Bovik, A.C.; Sheikh, H.R.; Simoncelli, E.P. Image quality assessment: From error visibility to structural similarity. *IEEE Trans. Image Process.* **2004**, *13*, 600–612. [CrossRef]
37. Renieblas, G.P.; Nogués, A.T.; González, A.M.; León, N.G.; Del Castillo, E.G. Structural similarity index family for image quality assessment in radiological images. *J. Med. Imaging* **2017**, *4*, 035501. [CrossRef]
38. Gonzalez, R.C.; Woods, R.E. *Digital Image Processing*; Prentice Hall: Upper Saddle River, NJ, USA, 2002.
39. Langner, O.; Dotsch, R.; Bijlstra, G.; Wigboldus, D.H.; Hawk, S.T.; Van Knippenberg, A. Presentation and validation of the Radboud Faces Database. *Cogn. Emot.* **2010**, *24*, 1377–1388. [CrossRef]
40. Cevikalp, H.; Neamtu, M.; Wilkes, M.; Barkana, A. Discriminative common vectors for face recognition. *IEEE Trans. Pattern Anal. Mach. Intell.* **2005**, *27*, 4–13. [CrossRef]
41. Liu, W.; Wen, Y.; Yu, Z.; Li, M.; Raj, B.; Song, L. Sphereface: Deep hypersphere embedding for face recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 21–26 July 2017; Volume 1, p. 1.
42. Du, S.; Tao, Y.; Martinez, A.M. Compound facial expressions of emotion. *Proc. Natl. Acad. Sci. USA* **2014**, *111*, E1454–E1462.
43. Martinez, A.; Benavente, R. The AR face database. *Rapp. Tech.* **1998**, *24*. Available online: <http://www2.ece.ohio-state.edu/~aleix/ARdatabase> (accessed on 1 September 2020).
44. Lucey, P.; Cohn, J.F.; Kanade, T.; Saragih, J.; Ambadar, Z.; Matthews, I. The extended cohn-kanade dataset (ck+): A complete dataset for action unit and emotion-specified expression. In Proceedings of the 2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), San Francisco, CA, USA, 13–18 June 2010; pp. 94–101.
45. Lyons, M.; Akamatsu, S.; Kamachi, M.; Gyoba, J. Coding facial expressions with gabor wavelets. In Proceedings of the 1998 Third IEEE International Conference on Automatic Face and Gesture Recognition, Nara, Japan, 14–16 April 1998; pp. 200–205.
46. Lee, H.S.; Park, S.; Kang, B.N.; Shin, J.; Lee, J.Y.; Je, H.; Jun, B.; Kim, D. The POSTECH face database (PF07) and performance evaluation. In Proceedings of the 8th IEEE International Conference on Automatic Face & Gesture Recognition (2008 FG'08), Amsterdam, The Netherlands, 17–19 September 2008; pp. 1–6.
47. Georghiades, A. Yale Face Database. Center for Computational Vision and Control at Yale University. 1997. Available online: <http://cvc.cs.yale.edu/cvc/projects/yalefaces/yalefaces.html> (accessed on 1 September 2020).
48. Shamir, L. Evaluation of face datasets as tools for assessing the performance of face recognition methods. *Int. J. Comput. Vis.* **2008**, *79*, 225.
49. Dang, L.M.; Hassan, S.I.; Im, S.; Moon, H. Face image manipulation detection based on a convolutional neural network. *Expert Syst. Appl.* **2019**, *129*, 156–168. [CrossRef]
50. He, M. Distinguish computer generated and digital images: A CNN solution. *Concurr. Comput. Pract. Exp.* **2019**, *31*, e4788. [CrossRef]



© 2020 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<http://creativecommons.org/licenses/by/4.0/>).

Article

# Innovatively Fused Deep Learning with Limited Noisy Data for Evaluating Translations from Poor into Rich Morphology

Despoina Mouratidis <sup>1,\*</sup>, Katia Lida Kermanidis <sup>1</sup> and Vilelmini Sosoni <sup>2</sup><sup>1</sup> Department of Informatics, Ionian University, 491 00 Corfu, Greece; kerman@ionio.gr<sup>2</sup> Department of Foreign Languages, Translation and Interpreting, Ionian University, 491 00 Corfu, Greece; sosoni@ionio.gr

\* Correspondence: c12mour@ionio.gr; Tel.: +30-266-1087756

**Abstract:** Evaluation of machine translation (MT) into morphologically rich languages has not been well studied despite its importance. This paper proposes a classifier, that is, a deep learning (DL) schema for MT evaluation, based on different categories of information (linguistic features, natural language processing (NLP) metrics and embeddings), by using a model for machine learning based on noisy and small datasets. The linguistic features are string based for the language pairs English (EN)–Greek (EL) and EN–Italian (IT). The paper also explores the linguistic differences that affect evaluation accuracy between different kinds of corpora. A comparative study between using a simple embedding layer (mathematically calculated) and pre-trained embeddings is conducted. Moreover, an analysis of the impact of feature selection and dimensionality reduction on classification accuracy has been conducted. Results show that using a neural network (NN) model with different input representations produces results that clearly outperform the state-of-the-art for MT evaluation for EN–EL and EN–IT, by an increase of almost 0.40 points in correlation with human judgments on pairwise MT evaluation. It is observed that the proposed algorithm achieved better results on noisy and small datasets. In addition, for a more integrated analysis of the accuracy results, a qualitative linguistic analysis has been carried out in order to address complex linguistic phenomena.



**Citation:** Mouratidis, D.; Kermanidis, K.L.; Sosoni, V. Innovatively Fused Deep Learning with Limited Noisy Data for Evaluating Translations from Poor into Rich Morphology. *Appl. Sci.* **2021**, *11*, 639. <https://doi.org/10.3390/app11020639>

Received: 18 December 2020

Accepted: 4 January 2021

Published: 11 January 2021

**Publisher's Note:** MDPI stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.



**Copyright:** © 2021 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

**Keywords:** machine learning; deep learning; machine translation; pairwise evaluation; educational data; small datasets; noisy datasets

## 1. Introduction

Machine translation (MT) applications have nowadays infiltrated almost every aspect of everyday activities. For the development of efficient MT solutions, reliable automated evaluation schemata are required. Over the past few years, neural network (NN) models have improved the state-of-the-art of different natural language processing (NLP) applications [1], such as language modeling [2,3], improving answer ranking in community question answering [4], improving translation modeling [5–7], as well as evaluating machine translation output [4,8,9]. Embeddings are a powerful way of representing text, provided that they are able to capture the linguistic identity (morphosyntactic and semantic profile) of a sentence/word. In 2013, Mikolov et al. [3] released the word2vec library. Word2vec became quickly the dominant approach for vectorizing textual data. The NLP models that were already well studied based on traditional approaches, such as latent semantic indexing (LSI) and vector representations using term frequency–inverse document frequency (TF-IDF) weighting, have been tested against word embeddings and, in most cases, word embeddings have come out on top. Since then, the research focus has shifted towards embedding approaches.

The present study aims to find out how embeddings, obtained through various means, in combination with different kinds of information fuse, affect classification accuracy small and noisy dataset, when used to train a model to choose the best translation output. The target languages (in contrast to the source language) are rich in morphology, as the

proposed schema is applied to the English–Greek (EN–EL) and English–Italian (EN–IT) language pairs. Greek and Italian languages have a rich inflectional morphology, as the nouns have different grammatical morphemes for the genders and the verbs have different grammatical morphemes for the two numbers and for the first, second and third person as well. In particular, the proposed NN learning schema is set up to test:

- two different forms of text structure (an informal (noisy) corpus (*C1*), and a formal, well-structured corpus (*C2*)) will be experimented with;
- a comparative analysis of two different ways of calculating embeddings (the straightforwardly mathematically calculated layer embeddings and the use of pre-trained embeddings) will be conducted
- the application of the SMOTE [10] oversampling technique during training will be investigated in order to overcome data imbalance phenomena;
- the use of two string-based linguistic features (hand-crafted), that capture the similarity between the MT outputs and the reference translation (*Sr*).

Further innovative aspects of the present work include:

- a novel deep learning architecture with innovative feeding structure that involves features of various linguistic levels and sources;
- a qualitative linguistic analysis that aims to reveal linguistic phenomena linked to poor/rich morphology, that impact on translation performance;
- the exploration of two different validation options (k-fold cross validation (*CV*) and Percentage split);
- the application of feature selection and dimensionality reduction methods;
- the application of the proposed multi-input, multi-level learning schema on text data from very different genres.

The rest of the paper is organized as follows—Section 2 presents the related work in the addressed scientific area. Section 3 describes the data sets (corpora), the feature set used, the learning framework and the network settings. Section 4 describes more experimental details and the results of the classification process. Finally, Section 5 presents the paper’s conclusions and directions for future research.

## 2. Related Work

Some of the most popular methods in automatic MT evaluation rely on score based metrics. These metrics include (i) metrics based on n-gram counts, such as Bilingual Evaluation Understudy (BLEU) [11] and National Institute of Standards and Technology (NIST) [12], or on the edit distance, like Word Error Rate (WER) [13], (ii) metrics using external resources, like WordNet and paraphrase databases—METEOR [14] and Translation Error Rate (TER) [15], (iii) metrics based on lexical similarity or syntactic similarity (involving higher level information, such as part of speech tags (POS)) between the MT outputs and the reference, and iv) neural metrics such as ReVal [8] and Regressor Using Sentence Embeddings (RUSE) [16], which directly learn embeddings for the entire translation and reference sentences using long short-term memory (LSTM) networks and pre-trained sentence representations.

Several research approaches on text classification, system ranking and selection techniques have been proposed using machine learning schemata. Guzmán et al. [4] focus on a ranking approach based on predicting BLEU scores. Duh [17] decomposes rankings into parallel decisions, with the best translation for each candidate pair predicted, using a ranking-specific feature set and BLEU score information. The framework involves a Support Vector Machine (SVM) classifier. A similar pairwise ranking approach was proposed by Mouratidis and Kermanidis [9], using a random forest (RF) classifier.

Neural networks are also used in the literature frameworks. Recurrent neural networks (RNN) and long short term memory (LSTM) networks [18], which are widely popular for learning sentence representations, have been taken up widely in a variety of NLP tasks [6,7]. Cho et al. [7] proposed a score-based scheme to learn the translation proba-



bility of a source phrase to a target phrase (MT output) with an RNN encoder-decoder. They showed that this learning scheme has improved the translation performance. The scheme proposed by Sutskever et al. [19] is similar to Cho et al. [7] work, but Sutskever et al. [19] chose the top 1000 best candidate translations produced by a Statistical Machine Translation (SMT) system with a 4-layer LSTM sequence-to-sequence model. LSTM networks are also widely adopted in MT evaluation [8]. LSTM memory units incorporate gates to control the information flow and they can preserve information for long periods of time. Wu et al. [20] trained a deep LSTM network to optimize BLEU scores when translating from English to German and English to French, but they found that the improvement in BLEU scores did not reflect the human evaluation of translation quality. Mouratidis et al. [21] used LSTM layers in a learning framework for evaluating pairwise MT outputs using vector representations, in order to show that the linguistic features of the source text can affect MT evaluation. Convolutional neural networks (CNN) are less common for sequence to sequence modeling, despite several advantages [22]. Compared to RNN, CNN create representations for fixed size contexts and do not depend on the computations of the previous time step because they do not maintain a hidden state. Gehring et al. [23] proposed an architecture for sequence to sequence modeling based on CNN. The model is equipped with linear units [24] and residual connections [25]. They also used attention in every decoder layer and demonstrated that each attention layer only adds a very small amount of overhead. Vaswani et al. [26] proposed a self-attention-based model and dispensed convolutions and recurrences entirely. Bradbury et al. [27] introduced recurrent pooling between a succession of convolutional layers, while Kalchbrenner et al. [28] studied neural translation without attention.

However, little attention has been paid to their direct applicability to languages with rich morphology. The present work focuses on the automatic evaluation of translation into morphologically rich languages, (Greek and Italian). The aim of this work is to identify the input information that is more effective for feeding a learning schema. Input information is investigated according to certain criteria, that is, the different means of calculating embeddings, the features of varying levels of linguistic information, the different dataset genres.

### 3. Materials and Methods

This section describes the dataset, the linguistic features and the NN architecture used in the experiments.

#### 3.1. Dataset

In these experiments, two different types of parallel corpora in the two language pairs (EN-EL and EN-IT) are used. The first dataset (*C1*) consists of the test sets developed in the TraMOOC project [29]. It is a small and noisy dataset as it is comprised of educational video lecture subtitles, lecture presentation slides and assignments, while it contains mathematical expressions, spoken language features, fillers, repetitions, and many special characters, such as /, @. The second formal dataset (*C2*) consists of parallel corpora from European Union legal documents, found on EUR-Lex, the online gateway to European Union Law, under the category "Consolidated texts". The chosen sentences are from Directives, Decisions, Implementing Decisions, Regulations and Implementing Regulations of the European Council and the European Commission, on the following issues: general, financial and institutional matters, competition and development policy, energy, agriculture, economic, monetary and commercial policy, taxation, social policy and transport policy. As pointed out, *C1*, is not a well-structured corpus as it contains linguistic phenomena which are unorthodox and ungrammatical, like misspellings, repetitions, fillers, disfluencies, spoken language features and so forth. On the other hand, *C2* is formal language text. For the *C1* corpus it was necessary to perform data pre-processing, that is, removal of special symbols (@, /), and alignment corrections. For the *C2* corpus no pre-processing was required. Two MT outputs were used - one generated by SMT models, that is, the Moses toolkit [30] for

C1 and Google Translate [31] for C2, and the second was generated by Neural Machine Translation (NMT) models, that is, the Nematus toolkit [32] for C1 and Google Translate for C2. The Moses and Nematus prototypes are trained in both in- and out- of domain data. The Nematus is trained on additional in-domain data provided via crowdsourcing, and also includes layer normalization and improved domain adaptation. In-domain data included data from TED, Coursera, and so forth [33]. Out-of-domain data included data from Europal, OPUS, WMT News corpora and so forth. The Google Translate prototype was trained on over 25 billion examples. More details about the corpora are presented in Table 1.

**Table 1.** Corpora details on the two machine translation (MT) outputs (*S1* for the Statistical Machine Translation (SMT) output and *S2* for the Neural Machine Translation (NMT) output) *SSE* for the source sentences and the *Sr*.

Corpus	Number of Sentences	Average of Sentences Length <i>SSE/S1/S2/Sr</i>	Number of Total Words <i>SSE/S1/S2/Sr</i>	Unique Words <i>SSE/S1/S2/Sr</i>
EL_C1	2687	15.8/15.9/15.7/16.2	42518/42953/42216/43562	5167/7331/7424/7830
EL_C2	2022	31.5/33.9/33.0/33.7	66425/68457/66773/68119	6022/8729/9022/9797
IT_C1	2687	15.8/15./15.6/16.0	42894/43152/42001/42357	5167/6280/6059/6440
IT_C2	2022	31.5/32.0/30.1/31.8	66425/67521/66982/68521	6022/6728/6556/7374

### 3.2. Features

The employed feature set is divided into two categories: one consisting of hand-crafted string-based features from the MT outputs, *SSE* and *Sr*, and the other consisting of commonly used NLP Metrics. The first category contains (i) simple features (e.g., distances like Levenshtein [34], longest word for *S1*, *S2*, *Sr*, *SSE*, features using the Length Factor (LF) [35]), (ii) features identifying the noise in the corpus (e.g., repeated words/characters, unusually long words in number of characters), and (iii) features providing linguistic information from the *SSE* in EN (e.g., the length of the *SSE* in number of words and number of characters). The feature set was inspired by the work of References [36,37]. The second category contains the NLP metrics, that is, the BLEU score, METEOR, TER and WER for (*S1*, *S2*), (*S1*, *Sr*), (*S2*, *Sr*). To calculate the BLEU score, an implementation of the BLEU score from the Python Natural Language Toolkit library [38] is adopted. For the calculation of the other three metrics, the code from GitHub [39] is used. The total number of features is 82. A detailed description of the feature set can be found in Reference [21].

In the present work, the employed feature set is extended and two additional novel linguistic feature pairs, which belong to the first category, have been used (increasing thereby the feature dimensions from 82 to 86). These features are similarity-based. The first feature *cmt* shows the percentage of identical words between the MT outputs and *Sr*, without taking into account the word order. The second feature *rmt* shows the percentage of identical parts of MT output included in the *Sr*. More specifically, this feature shows whether the MT output is a contiguous subsequence of *Sr*. The features are defined in Equations (1) and (2) respectively:

$$c_{mt} = \frac{|S_{mt} \cap S_r|}{|S_{mt} \cup S_r|} \tag{1}$$

$$r_{mt} = \frac{|S_{mt} \cap S_r|}{|(S_{mt} \cap S_r)'|} \text{ with } |(S_{mt} \cap S_r)'| \neq 0. \tag{2}$$

where *Smt* is one of the *S1*, *S2*.

As an example, if

*Sr* = {η (the), υπηρεσία (department), προσδιορίζει (specify), το (the), διάστημα (period)},

*S1* = {το (the), χρονικό (time), διάστημα (period), η (the), υπηρεσία (department), καθορίζει (determines)},

$S2 = \{\eta$  (*the*), υπηρεσία (*department*), προσδιορίζει (*specify*), την (*the*), περίοδο (*period*), then  $cmt = 0.57$ ,  $rmt = 1.3$  for  $S1$  MT output and  $cmt = 0.43$ ,  $rmt = 0.75$  for  $S2$  MT output.

All feature values were calculated using MATLAB, and their values have been normalized and vary between 0 and 1.

### 3.3. Embedding Layers

Firstly, an embedding layer (mathematically-calculated embeddings) is used for the two MT outputs and the Sr. The encoding function applied is the one-hot function. The embedding layer size, in number of nodes, is 16. The input dimensions of the embedding layer is in agreement with the vocabulary of each language, taking into account the most frequent words (500 for EN-EL/700 for EN-IT). The embedding layer used is the one provided by Keras [40]. Secondly, a Greek version of WordSim353 [41] is adopted for pre-trained embeddings. More specifically WordSim353 contains the 300-dimensional Greek embeddings of 350 K words, trained on 20 M of URLs with Greek language content and they computed in 2018. More details about the number of unique sentences, unigrams, bigrams, trigrams and so forth can be found in Outsios et al. [41]. In this case, the embedding layer utilized the embedding matrix produced by the embedding\_index dictionary and the word\_index. The Embedding layer should be fed with padded sequences of integers. For this purpose, the `keras.preprocessing.text.Tokenizer` and the `keras.preprocessing.sequence.pad_sequences` [40] were run. For the pre-trained Italian embeddings, the Wikipedia2Vec tool is used [42]. The size, in number of nodes, of the embedding layer is 300, as is the dimension of pre-trained embeddings for both datasets.

### 3.4. NN Architecture

This study aims to identify the best MT output out of the two provided. Two linguists annotated the sentences with 1 if the NMT output is better than the SMT one and with 0 if the SMT output is better than the NMT. A low annotation percentage is observed for the SMT class (EL: 37% for  $C1$ , 48% for  $C2$ , IT: 43% for  $C1$ , 48% for  $C2$ ) compared with the NMT class (EL: 63% for  $C1$ , 52% for  $C2$ , IT: 57% for  $C1$ , 52% for  $C2$ ). A low annotation agreement rate is observed ( $C1$ : 5% for EN-EL/6% for EN-IT,  $C2$ : 3% for EN-EL/5% for EN-IT). For the few different answers, the annotators had a discussion and finally agreed on one common label. The NN model takes as input the tuple ( $S1$ ,  $S2$ ,  $Sr$ ). These sentences are passed to the embedding layer. Two ways for extracting embeddings are applied (described in Section 3.3) producing  $EmbS1$ ,  $EmbS2$ ,  $EmbSr$ . The  $EmbS1$ ,  $EmbS2$ ,  $EmbSr$  vectors are concatenated in a pairwise fashion as ( $EmbS1$ ,  $EmbS2$ ), ( $EmbS1$ ,  $EmbSr$ ), ( $EmbS2$ ,  $EmbSr$ ), and they form the input to the similarity-based hidden layers  $h_{12}$ ,  $h_{1r}$ ,  $h_{2r}$ . As extra inputs, the hidden layers are fed with the matrices  $H_{12}[i,j]$ ,  $H_{1r}[i,j]$ ,  $H_{2r}[i,j]$  (where  $i$  is the number of sentences and  $j$  the number of features), containing the second category features (NLP set). The hidden layer outputs form the input to the output layer. Moreover, an extra input to the output layer is used: the matrix  $A[i,j]$ , containing the first category features (described in Section 3.2). The DL NN schema is shown in Figure 1.

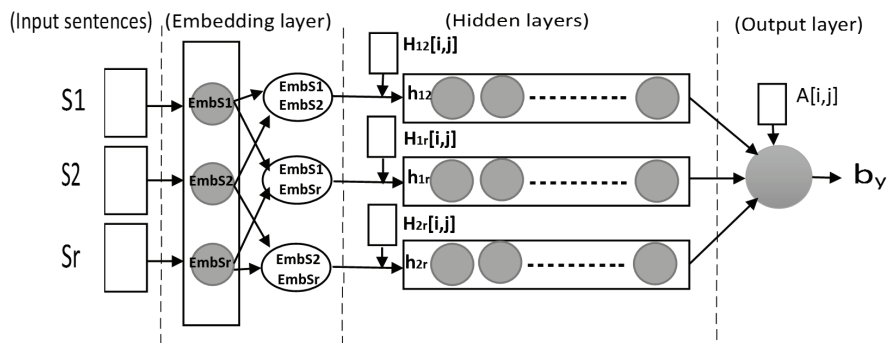


Figure 1. Neural network (NN) architecture.

The binary classification problem is modeled as a Bernoulli distribution (Equation (3)):

$$Y \sim \text{Bernoulli}(Y \setminus by), \tag{3}$$

where  $by$  is the sigmoid function  $\sigma(w^T x + b)$ ,  $w^T$  and  $b$  are the network's parameters.

### 3.5. Network Settings

The network model architecture for the experiments is a classic architecture of RNN networks (2 LSTM layers with 400 hidden units) and feedforward layers (4 Dense layers, that is, 3 layers with 50 hidden units and 1 layer with 400 hidden units). The network is trained using the Adam optimizer [43] to optimize parameters. To avoid over-fitting, dropout is applied with a rate of 0.05, using the loss function of binary cross entropy and the regularization parameter  $\lambda$  is set equal to “ $10^{-3}$ ”. 10-fold CV and 70% percentage split were employed for testing.

## 4. Results

### 4.1. Performance Evaluation

In this experiment (a) we investigate whether the predicted classifications have any correlation with human annotation, (b) we compare the proposed classification mechanism against the baseline classification models for small noisy and formal datasets respectively, (c) we compare two different ways of generating the embedding layer, and (d) we test two different options of validation methods. Table 2 presents the classification results (Precision and Recall) for the different MT outputs over the two different datasets. The  $C1$  corpus presents a classification increase, for both language pairs (accuracy: 72% EN-EL/70% for EN-IT), in contrast to the  $C2$  corpus (accuracy: 68% for EN-EL/65% for EN-IT), even though the  $C1$  corpus contains a lot of noise. This is probably due to the fact that the  $C1$  corpus contains more sentences, and, also, because the  $C2$  corpus has richer vocabulary and more formal structure. It is more difficult for the classifier to choose the best MT output, because the SMT output is more similar to the NMT output in this corpus ( $C2$ ). It is also observed that both evaluation metrics chose the NMT model over the SMT one, which is in accordance to the annotators' results. In addition, the aforementioned accuracy results are obtained when the NN uses the simple embedding layer. However, when the pre-trained embeddings are used, the model does not lead to better results (average accuracy of  $C1$  and  $C2$ : 66% for EN-EL/65% for EN-IT), since the embeddings are trained on the general-purpose corpus, which is not representative of the input corpora used therein. At this point, it is worth mentioning that the pre-trained embeddings seem to be more effective for the EN-IT pair than for the EN-EL language pair. As far as the different types of corpora are concerned, pre-trained embeddings are more efficient for the  $C2$  corpus (average accuracy of EN-EL and EN-IT: 66%) than the  $C1$  corpus (average accuracy of EN-EL and EN-IT:

64%). This is probably due to the fact that the C2 corpus has richer vocabulary than the C1 corpus.

An approach to improve the classification accuracy of a small and noisy dataset is to apply the SMOTE oversampling technique on the training data. Using SMOTE, the sentences of the minority class (SMT) doubled in number, and the total number of sentences reached 3024 for C1 and 2276 for C2. It is important to compare the performance between the 82 and the 86 feature dimensions, with and without the SMOTE filter. When SMOTE is applied, a small accuracy increase is observed on the 82 features (average accuracy of C1 and C2: 68% for EN-EL/67% For EN-IT), and an even higher increase on the 86 features (average accuracy of C1 and C2: 70% for EN-EL/68% for EN-IT). It is interesting that the EN-EL corpora outperformed EN-IT in all the experiments. The results with the use of the two new suggested features are generally better for both corpora and language pairs.

**Table 2.** Accuracy performance for two embeddings layer types for the two corpus English–Greek (EN–EL)/English–Italian (EN–IT).

MT Model	Simple Embedding Layer				Pre-Trained			
	82 Features		86 Features		82 Features		86 Features	
	Precision	Recall	Precision	Recall	Precision	Recall	Precision	Recall
<b>Language pair: EN-EL</b>								
<b>NN model with 2687 segments for C1 and 2022 segments for C2</b>								
SMT C1	70%	<b>89%</b>	70%	<b>92%</b>	70%	<b>77%</b>	68%	<b>77%</b>
NMT C1	67%	37%	<b>69%</b>	31%	54%	40%	50%	45%
SMT C2	62%	59%	63%	60%	60%	<b>58%</b>	62%	<b>64%</b>
NMT C2	58%	60%	55%	<b>59%</b>	56%	<b>59%</b>	57%	<b>62%</b>
<b>NN model_SMOTE with 3024 segments for C1 and 2276 segments for C2</b>								
SMT C1	68%	72%	68%	<b>78%</b>	65%	67%	65%	75%
NMT C1	48%	41%	48%	42%	40%	37%	<b>54%</b>	46%
SMT C2	58%	49%	60%	52%	66%	45%	65%	<b>73%</b>
NMT C2	60%	59%	60%	64%	54%	<b>65%</b>	55%	45%
<b>Language pair: EN-IT</b>								
<b>NN model with 2687 segments for C1 and 2022 segments for C2</b>								
SMT C1	62%	44%	65%	44%	68%	52%	70%	<b>80%</b>
NMT C1	<b>70%</b>	<b>87%</b>	<b>60%</b>	<b>80%</b>	65%	75%	<b>82%</b>	60%
SMT C2	55%	31%	57%	37%	56%	40%	59%	45%
NMT C2	54%	76%	55%	76%	60%	81%	62%	80%
<b>NN model_SMOTE with 3024 segments for C1 and 2276 segments for C2</b>								
SMT C1	50%	63%	58%	38%	<b>70%</b>	55%	68%	77%
NMT C1	56%	43%	61%	77%	65%	69%	70%	55%
SMT C2	51%	51%	57%	45%	56%	40%	58%	40%
NMT C2	52%	56%	60%	56%	62%	68%	70%	65%

Firstly, k-fold CV was used, which is a reliable method for testing the models, and a value of  $k = 10$  is very common in the field of machine learning [44] (Table 2). Secondly, part of the data (70%) is kept for training, and part (30%) is applied for testing (Table 3). Given that both classes are of interest, the symmetric Matthews correlation coefficient (MCC) metric [45] (a special case of the  $\phi$  phi coefficient [46]) is used, as it constitutes a good way to describe the relation of TP (true positive), FP (false positive) and FN (false negative) values by a single number. It is defined as follows:

$$MCC = \frac{TP \times TN + FP \times FN}{(TP + FP) \times (TP + FN) \times (TN + FP) \times (TN + FN)}. \tag{4}$$

When using 10-fold CV, C1 outperforms C2 for both language pairs. When the percentage split method (70% training–30% testing) is used, a small performance improvement is observed for the C2 corpus. Moreover, MCC achieves higher value for the C2 corpus, when pre-trained embeddings are used.

**Table 3.** Accuracy performance (MMC) in different cross validation options.

MT Model MCC/Corpus	10 Fold CV		70% Per. Split	
	C1	C2	C1	C2
EN-EL_simple emb layer	<b>0.32</b>	0.17	0.29	0.22
EN-IT_simple emb layer	0.10	0.12	0.10	<b>0.15</b>
EN-EL_pre-trained emb	<b>0.20</b>	0.15	0.18	0.17
EN-IT_pre-trained emb	0.11	0.13	0.10	<b>0.14</b>

Figure 2 shows the accuracy performance according to training speed and batch size. Increasing the batch size can increase the model’s accuracy. As seen above, the training speed decays more quickly for the simple embedding layer compared to the pre-trained embedding layer model. Moreover, the accuracy of the pre-trained embeddings is consistently higher for corpus C2. The best performance has been consistently obtained for batch size 512.

It is important to analyze the correlation with human-performed evaluations [47]. In this work, the correlation of the predicted scores with human judgments is reported using Kendall  $\tau$ . Kendall  $\tau$ , is a coefficient that measures the agreement between rankings produced by human judgments, and rankings produced by the classifier. The WMT’12 (Workshop of Machine Translation) definition of Kendall’s  $\tau$  is used, and it is calculated as follows:

$$\tau = \frac{(\text{concordant pairs} - \text{discordant pairs})}{\text{total pairs}} \quad (5)$$

where ‘concordant pairs’ is the number of times the human judgment and the predicted judgment agree in the ranking of any two translations that belong to the same SSE, and ‘discordant pairs’ is the opposite.

#### 4.1.1. Comparison to Related Work

As mentioned earlier, there is limited work on pairwise evaluation based on the small and noisy dataset. In order to compare our results with other methods, additional experiments were reproduced in order to imitate as closely as possible earlier work settings, that were (i) based on different classifiers such as SVM [17] and RF [37] and (ii) based on other evaluation methods, that is, the use of the BLEU score [4,17].

Figure 3 shows the overall Kendall  $\tau$  for the different approaches. The proposed DL schema has achieved comparable performance to the models proposed in earlier works. The SVM classifier succeeds in a strong positive relationship between the two classes for C1\_EN-EL: 0.7, and moderate positive relationship for C2\_EN-EL: 0.4, C1\_EN-IT: 0.4 and C2\_EN-IT: 0.6, while the RF classifier reached a moderate positive relationship for the C1 corpus (0.4 for EN-EL/0.6 for EN-IT) and for the C2 corpus (0.4 for EN-EL/0.6 for EN-IT). When the BLEU score information is used, the model achieved a moderate positive relationship. Kendall  $\tau$  reached its highest value when the proposed schema uses the simple embedding layer, the feature set of 86 dimensions, and the NLP set for both language pairs (EN-EL: 0.7 for C1/0.6 for C2 and EN-IT: 0.6 for C1/0.5 for C2).

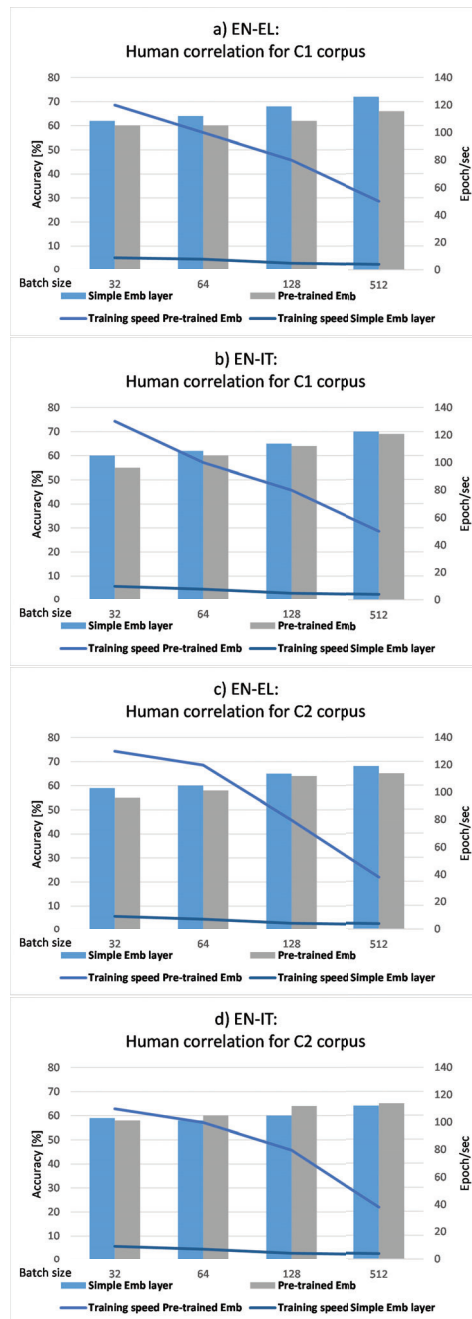


Figure 2. Human correlation. Simple embedding layer vs Pre-trained embeddings.



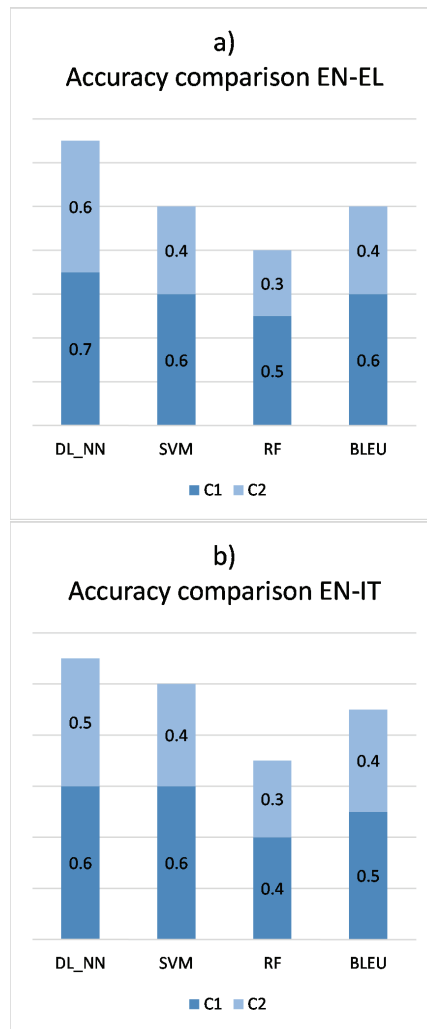


Figure 3. Accuracy performance (Kendall  $\tau$ ) compared with related work

#### 4.1.2. Feature Selection and Dimensionality Reduction

There are many techniques for improving the classifier’s performance. Feature selection (FS) and Dimensionality reduction (DR) are two commonly used techniques that improve classification accuracy [48]. The main idea behind FS is to remove redundant or irrelevant features that are not useful for the classifier [49]. The advantage of FS is that no information about the importance of single features is lost. With DR the size of the feature space is decreased, but without losing vital information [50].

FS methods are usually categorized in two basic methods: wrappers and filters [51]. Wrapper FS methods evaluate multiple models with different subsets of input features and select those features that result in the best performing model according to a performance metric. The number of possible results will increase geometrically as the number of features increases. Filter FS methods use statistical techniques to evaluate the relationship between each input variable and the target variable, and these scores are used as the basis to choose (filter) those input variables that will be used in the model. Filters are either global or

local. Global methods assign a single score to a feature regardless of the number of classes while local methods assign several scores, as every feature in every class has a score [52]. Global methods typically calculate the score for every feature and then choose the top- $N$  features as the feature set, where  $N$  is usually determined empirically. Local methods are similar but require converting a feature's single score before choosing the top- $N$  features. Wrappers require much more computation time than filters, and may work only with a specific classifier [51]. Filters are the most common *FS* method for text classification. Some commonly used *FS* methods are a. Recursive Feature Elimination Cross Validation (RFECV) that belongs to the Wrappers methods, b. the information gain (IG) [53] that belongs to filter global *FS* methods, and c. the Chi-square (CHI) [54], that belongs to the filter local methods. All these *FS* methods are language-independent feature selection methods that produce better accuracy.

In these experiments *RFECV* is tested using Support Vector Machines (SVM) with linear kernel and the number of cross validation folds is set to 10. Information gain is often applied to find out how well each single feature  $A$  separates the given feature data set  $S$  and it is calculated as follows:

$$IG(S, A) = I(S) - \sum_{n \in A} \frac{|S_n|}{|S|} I(S_n), \quad (6)$$

where  $n$  is the value of every feature ( $A$ ) and  $S_n$  is the set of instances where  $A$  has value  $n$ .

*CHI* is a supervised *FS* method that calculates the correlation of a feature value  $n$  with the class  $m$ , and it is calculated as follows:

$$x^2 = \sum_{i=1}^n i \sum_{j=1}^m j = \frac{(O_{ij} - E_{ij})^2}{E_{ij}}, \quad (7)$$

where  $O_{ij}$  is the observed frequency and  $E_{ij}$  is the expected frequency.

*DR* refers to algorithms and techniques that create new features which are combinations of the old features [54]. The most important *DR* technique is principal component analysis (PCA) [55]. *PCA* is an unsupervised dimensional reduction technique. *PCA* produces new features from the original features by converting the high dimensional space of the original features to a low dimensional space while keeping linear structure. Dimensionality reduction is accomplished by choosing enough eigenvectors to account for some percentage of the variance in the original data (a default value is 0.95). Attribute noise was filtered by transforming the original into the *PC* space, eliminating some of the worst eigenvectors, and then transforming back to the original space. The maximum number of attributes to include in the transformed space was set to 5.

Better accuracy results are observed, in general, when a feature selection method is used, in contrast to the whole feature set model (Table 4). The accuracy performance increased 4% for the *C1* corpus for EN-EL and 3% for EN-IT. It seems that the application of these methods is more efficient for the SMT for the informal *C1* corpus and NMT for the formal (well-structured) *C2* corpus. More specifically, there is an increase up to 4% for the SMT class for *C1* and 2% for *C2*, while, for the NMT class, there is 2% for *C1* and *C2*. In addition, the feature selection methods work better for *C1* (an increase up to 3.5% in average for both language pairs) rather than the *C2* (an increase up to 2.5% in average for both language pairs). We conclude that feature selection methods help more the noisy corpus. This is in accordance with the accuracy results of the previous model.

Table 4. Feature selection accuracy performance for the two corpus EN-EL/EN-IT.

Method	RFECV		IG		CHI2		PCA	
No of Features	23/86		49/86		70/86		54 new	
	Precision	Recall	Precision	Recall	Precision	Recall	Precision	Recall
<b>Language pair: EN-EL_2687 segments for C1 and 2022 segments for C2</b>								
SMT C1	66%	87%	67%	91%	70%	93%	67%	90%
NMT C1	54%	26%	63%	26%	68%	31%	61%	25%
SMT C2	63%	85%	66%	70%	67%	73%	61%	80%
NMT C2	74%	46%	62%	60%	68%	61%	68%	45%
<b>Language pair: EN-IT_2687 segments for C1 and 2022 segments for C2</b>								
SMT C1	59%	40%	62%	40%	65%	39%	58%	32%
NMT C1	52%	60%	59%	82%	60%	87%	56%	79%
SMT C2	56%	30%	56%	30%	57%	30%	53%	25%
NMT C2	54%	70%	55%	79%	55%	79%	53%	75%

Concerning the features, it is verified that, for the proposed model, the more effective features are those containing ratios, features identifying the presence of noise in a segment (for example the occurrence of repeated characters) and features used linguistic information from the *SSE*. They all seem to be useful for prediction. Also, the new string-based features added in this paper are presumed to enclose valuable information for the model as they capture the similarity between the MT outputs and the reference translation. The new string-based features were selected almost from every method. Regarding the *FS* method, it seems that better accuracy results were produced with *CHI* square and *IG*. Additionally, it is observed that the feature reduction space method (*PCA*) does not help the accuracy performance regardless of the corpora structure-type, since in all experiments the performance was less than or equal to the classifier performance using the whole feature set.

#### 4.2. Linguistic Analysis

In order to have a more comprehensive analysis of the accuracy results, we have carried out a qualitative linguistic analysis as well. In this context, problems have been identified regarding some complex linguistic phenomena for both language pairs (Table 5). For the first sentence (ID1): (Both NN and the Annotator's choice was S2)

- The verb *to deploy* means: *to develop, to emplace, to set up*. Both S1 and S2 erroneously translated that verb as *to use*. Nevertheless, the verb *to use* is one of the secondary meanings of the verb *to deploy*.
- The most common meaning of the word *bug* is *insect, virus*, but it also means: *error*. The word *fix* means *repair, determine, nominate*. In this sentence, *bug fix* is used as a noun phrase, where the first word functions as a subjective genitive, and the phrase means: *error correction*. S1 commits two errors when translating "*fix*" (*φτιάξουμε*), i. *Fix* is erroneously considered to have a verb function. ii. It is difficult to explain why the same verb is translated in the first person of plural of the simple past-subjunctive. As a consequence, S1, S2's translations for the verbal phrase (*deploys a bug fix*) are both nonsensical: S1: "*χρησιμοποιεί ένα έντομο φτιάξουμε*" ("*uses an insect*" + simple past-subjunctive of "*repair*"), S2: *χρησιμοποιεί "ένα σφάλμα για τα έντομα"* (*uses an error for the insects*).
- In addition, it is important to notice that S2 has translated the same phrase (*bug fix*) at the end of the sentence in a different way. S2 tried to improve the translation and it certainly succeeded, but only for the word *fix* (*διόρθωση*). S2 also "*spotted*" that *bug* is a subjective genitive (the correction of the *error*), but it still identified *bug* as an *insect* and it has erroneously translated it: *ζουζιού*. In Greek, this is a nonexistent word, but it is strongly reminiscent of the word *ζουζούνη* (*insect*), which is an onomatopoeic

word (*the buzz of a bug*) and especially of its genitive case: ζουζουινιού, with some letters missing.

- *S1* has correctly “identified” the meaning of the verb *to list* (*to enumerate*), but not in the correct grammatical number—third-person plural, instead of third-person singular. *S2* chose the correct grammatical inflectional morphemes for the number and the person, but not the correct meaning, for this context: *referred*, instead of *enumerated*, *indexed* or *set out*. So, the proposed NN model has correctly chosen *S2*, as *S2* “recognized” the correct grammatical morphemes of number and person features.
- Regarding the passive future verb: *will be updated*: In *S1*, the preceding particle of future tense in Greek θα (According to the Cambridge Dictionary, a particle is a word or a part of a word that has a grammatical purpose but often has little or no meaning. <https://dictionary.cambridge.org/dictionary/english/particle>) (*will*) is separated from the subjunctive (επικαιροποιηθεί), which is wrong.
- Both *S1* and *S2* have erroneously translated the noun phrase: *cache manifest*. As they failed to identify the multi-word expression, they have translated them separately. The word *cache* means: *crypt*, *hideout*, *cache memory*, and, in this sentence, it has the last meaning (κρυφή μνήμη). However, *S1* “chose” the first meaning (κρύπτη) (*crypt*), whereas *S2* left the word untranslated. *Manifest* means obvious, apparent. Both *S1* and *S2* “chose” from these synonyms. Nevertheless, the *cache manifest* in HTML5 is a software storage feature which provides the ability to access a web application even without a network connection ([https://en.wikipedia.org/wiki/Cache\\_manifest\\_in\\_HTML5](https://en.wikipedia.org/wiki/Cache_manifest_in_HTML5)). So, the best translation would be: κρυφή μνήμη ιστότοπου (*website cache memory*), a translation that was not even produced in the reference.

For the second sentence (ID2): (NN chose *S1* / Annotator’s choice was *S2*)

- *Sit*: Both *S1* and *S2* have erroneously translated this verb (κάτσετε, καθήσετε). In this sentence, the verb *to sit* is transitive and means: *to place*, *to put*, requiring an inanimate object, whereas the very common meaning of this verb, that is *to have a seat*, presupposes that the verb is intransitive (+animate subject) or transitive (but: +animate object: I make someone sit down). Both *S1* and *S2* have erroneously adopted the second meaning, without “noticing” that its object (*spheroids*) is an inanimate noun. Even more, the form chosen by *S1* belongs to oral speech (κάτσετε) (*to sit*), while *S2*’s form is misspelled (καθήσετε *to sit*), instead of the correct: καθίσετε).
- Kind of is an informal expression modifying and especially attenuating (It is the opposite of really. In the UK, it is considered quite informal. <https://english.stackexchange.com/questions/296634/kind-of-like-is-a-verb>) the meaning of the verb *plonk*. *S1* has erroneously “identified” that word as a noun and so mistranslated it as: *form*, *genre*, *species* (είδους). Nevertheless, *S1* “identified” the inflectional grammatical morpheme of the genitive case: -ους for *of*.
- *Plonk down*: This phrasal verb has a lot of meanings: *drop heavily*, *place down*, *impale*, *attract* and so forth (<https://glosbe.com/en/el/plonk%20down>) *S1* has erroneously translated this verb in the meaning of *impale*, which is not the case. *S1* has separately translated the whole sentence (*they kind of plonk down*: είδους παλουκώσει τους κάτω), which is completely nonsensical in Greek. In addition, the verb object them has been erroneously placed after the verb (in Greek, the clitic form of the personal pronoun is placed before the verb) and has been translated by a wrong grammatical morpheme (masculine plural (τους) instead of neutral plural (τα)). On the other hand, *S2* has correctly “found” the connection of those words (*kind of plonk down*), but it translated them in a wrong and, at first sight, non-understandable way: συναρπάζουν (*fascinate*).

For the third sentence (ID3): (NN chose *S1* / Annotator’s choice was *S2*)

- *S1* incorrectly translated the phrase: *will get us accustomed to*, considering that the two verbs are independent of each other (θα δώσει (*will give*), συνηθίσει (*will get used*)), without taking into account that the verb *get* has a metaphorical meaning: *cause something to happen*, and not the literal one: *take*. The verb *get*, in this sentence, forms a

multi-word expression with the verb *accustomed* and the preposition *to*, which, as a past participle, depends on the first. *S2* correctly translated the phrase as: θα μας κάνει συνηθισμένους (*it will make us get used*), left the word untranslated.

- *S1* incorrectly translated the last link of the sentence: (να τις ιδιαιτερότητες (*here the particularities*)(!)), translating the preposition *to* as if it were before an infinitive, without taking into account that this is the second part of: *accustomed to... and to*. Related to the latter is that *S1* incorrectly translated the word after *to*, that is, the possessive adjective *their*, as a definite article in plural: τις (*the*).

For the third sentence (ID4):(Both NN and the Annotator's choice was *S2*)

- *Fee*: the word has a lot of translations in Italian language: *tassa*, *retribuzione* (*salary*), *compenso* (*compensation*), *pagamento* (*payment*), *contributo* (*contribution*) and so forth. Both *S1* and *S2* chose the most common meaning (*tassa*), but not the right one for this context: *spesa* (*expenditure or charges*).
- Both *S1* and *S2* erroneously put question marks for the accented morphemes: ? instead of è and *attivit?* instead of *attività* (*activities*).
- *Atteggiamento*: both *S1* and *S2* correctly translated the word (*attitude*), but they both did not put it in the right position, as in Italian sentence structure (in contrast with the English language) the quotation, functioning as a title, follows the word *atteggiamento* (*attitude*), characterising and explaining it.
- *Assets*: Both *S1* and *S2* translated this word as *attività*. The most common meanings of the word *attività* are: *activity, practice, action, operation* etc, but it also means: *business, assets, resources, occupation* etc), whereas *assets* meanings are: *property, benefit, resource, investment* and so forth. Both *S1* and *S2* chose the closer meaning, but not the right one (*risorse*). The reason for this relatively successful choice may be the first word of the concordance (*underused assets*), in opposite meaning with the most of the other translations.
- *Save*: Both *S1* and *S2* erroneously translated the word as *salvare* and *salvano*, respectively, instead of *risparmiano*. Even though the English verb to save derives from the same Latin verb (*salvare*), in Italian the main meanings of *salvare* are *rescue, salvage or safeguard*.

In conclusion, the NN model has chosen *S2* in the first sentence, since *S1* faces difficulties with some linguistic phenomena, like homonymy (e.g., the homographs of *bug*), synonymy (e.g., the similar meanings of *fix*) and polysemy as well. In addition, *S1* often fails to address certain grammatical and syntactic phenomena: subject-verb agreement, phrase structure rules, phrasal verb schemata, and so forth. However, the NN model has mainly chosen *S1* in the second sentence, because *S1* "recognizes difficult" grammatical morphemes (like "kind of"). *S2* addresses effectively the aforementioned linguistic phenomena, and generally "recognizes" the rich morphology of the Greek and Italian language (e.g., grammatical agreements, different grammatical genders, structure rules), and, in certain cases, it misses multi-word expressions and phrasal meanings as well. Nevertheless, *S1* seems to employ richer vocabulary (e.g., *απαριθμούνται* (*enumerate*), *κρυπτή* (*crypt*), *πρόδηλο* (*obvious*)) than *S2*. Indeed, *S1* supports different and not so common senses for each word and it often chooses the one closer to the correct translation, whereas *S2*, without this extended vocabulary, sometimes fails to translate the less common word, or translates it with a nonexistent word (e.g., *cache*, *ζουζιού* respectively).

**Table 5.** Linguistic Analysis for EN-EL and EN-IT.

ID	SSE	S1	S2	Sr
1	If an ARSnova developer deploys a bug fix which will modify a single file listed in the cache manifest, will the local file concerning the bug fix be updated in your browser?	Εάν ένας προγραμματιστής ARSnova χρησιμοποιεί ένα έντομο φτιάξουμε το οποίο θα τροποποιήσουν ένα ενιαίο αρχείο που απαρτιούνται στην κρύπτη πρόδηλο, θα το τοπικό αρχείο σχετικά με το μικρόβιο φτιάξουμε επικαιροποιηθεί στον περιηγητή σας;	Αν ένας προγραμματιστής ARSnova χρησιμοποιεί ένα σφάλμα για τα έντομα, το οποίο θα τροποποιήσει ένα μόνο αρχείο που αναφέρεται στο δηλωτικό του cache, θα ενημερωθεί το τοπικό αρχείο σχετικά με την διόρθωση του ζουζιού στο πρόγραμμα περιήγησης;	Αν ένας προγραμματιστής ARSnova αναπτύξει μια διόρθωση για ένα σφάλμα του προγράμματος που θα τροποποιεί ένα μοναδικό αρχείο που εμφανίζεται στην κρυφή μνήμη, θα ενημερωθεί το τοπικό αρχείο σχετικά με τη διόρθωση του σφάλματος στον περιηγητή σας;
2	Then he's made a structure where you can sit these spheroids, I think they kind of plonk them down on these metal pyramids.	Στη συνέχεια έκανε μια δομή όπου μπορείτε να κάτσετε αυτά τα σφαιρίδια, νομίζω ότι είδους παλουκώσει τους κάτω από αυτά τα μεταλλικά πυραμίδες.	Μετά έφτιαξε μια δομή όπου μπορείτε να καθίσετε αυτά τα σφαιρικά, νομίζω ότι τους συναρπάζουν σε αυτές τις μεταλλικές πυραμίδες.	Έπειτα αυτός έχει δημιουργήσει μια δομή όπου μπορείς να τοποθετήσεις αυτά τα σφαιροειδή, νομίζω ότι αυτοί κατά κάποιον τρόπο τα ρίχνουν σε αυτές τις μεταλλικές πυραμίδες.
3	Deductive vs Inductive, or Definitely vs Probably, will get us accustomed to the two main breeds of arguments and to their particularities.	Επαγωγικό έναντι επαγωγικά, ή Σίγουρα έναντι Πιθανόν, θα μας δώσει συνηθίσει τα δύο κύρια φυλών επιχειρήματα και να τις ιδιαιτερότητες.	Επαγωγικό εναντίον του Inductive, ή σίγουρα εναντίον πιθανόν, θα μας κάνει συνηθισμένους στις δύο κύριες φυλές των επιχειρημάτων και στις ιδιαιτερότητες τους.	Παραγωγική έναντι Επαγωγικής σκέψης ή Πιθανότητα έναντι Βεβαιότητας, θα μας εξοικειώσει με τα δυο βασικά είδη επιχειρημάτων και τις ιδιαιτερότητες τους.
4	"The what's mine is yours, for a small fee" attitude helps owners make some money from underused assets and at the same time the collaborators save a huge percentage of their resources.	"Quello che ? mio ? tuo, per una piccola tassa" atteggiamento proprietari aiuta a fare dei soldi da attiviti? sottoutilizzato e allo stesso tempo i collaboratori salvano una grande percentuale delle loro risorse.	"La mia ? la tua, per una piccola tassa" aiuta i proprietari a fare un po 'di soldi da attiviti? sottoutilizzate e allo stesso tempo i collaboratori salvano un'enorme percentuale delle loro risorse.	L'atteggiamento "Quello che è mio è tuo con una piccola spesa" aiuta i proprietari a guadagnare qualcosa dalle risorse sottoutilizzate e allo stesso tempo i collaboratori risparmiano una percentuale enorme delle loro risorse.

## 5. Conclusions and Future Work

This paper presented an innovative DL NN architecture for MT evaluation into morphologically rich languages. The architecture is tested on two different types of small corpora, one noisy and one formal and two different language pairs (EN-EL and EN-IT). The proposed DL schema used linguistic information from two MT outputs, SSE as well as the NLP set. Experiments revealed that when the DL schema utilizes the simple embedding layer and not the pre-trained embeddings, the results are better. In addition, the results using the two new suggested features and the SMOTE filter are generally better. Based on the linguistic analysis, when the MT output "recognized" the grammatical morphemes, the proposed NN model chose it as the best translation. According to the validation method, percentage split gave more balanced results for both corpora, but the 10-CV method gave higher accuracy results. The DL schema used many features, so it is important to thoroughly investigate the importance of these features for assigning them with proper weights during the NN model training. In this paper, feature selection and dimensionality reduction methods were employed and they showed that feature selection methods help more the noisy corpus. It is noticed that the proposed algorithm

conducted better results on the noisy and small dataset. For further experimentation, it is quite interesting to explore why all the classifiers led to worse results in terms of the evaluation accuracy in EN-IT than in the EN-EL language pair, taking into account that the linguistic features employed are language independent. Another idea to explore would be the pre-trained embeddings utilization, as an initialization for the embedding layer. Finally, we plan to verify another morphological schema that could improve classification performance.

**Author Contributions:** D.M., K.L.K., conceived of the idea, D.M., designed and performed the experiments, analyzed the results, drafted the initial manuscript and K.L.K., V.S., revised the final manuscript, supervision. All authors read and agreed to the published version of the manuscript.

**Funding:** This research received no external funding.

**Institutional Review Board Statement:** Not applicable.

**Informed Consent Statement:** Not applicable.

**Data Availability Statement:** Some or all data generated or used during the study are available from the corresponding author by request.

**Acknowledgments:** The authors would like to thank the two Greek and Italian and language experts for the annotation.

**Conflicts of Interest:** The authors declare no conflict of interest.

## References

- Goldberg, Y. A primer on neural network models for natural language processing. *J. Artif. Intell. Res.* **2016**, *57*, 345–420. [[CrossRef](#)]
- Bengio, Y.; Ducharme, R.; Vincent, P.; Jauvin, C. A neural probabilistic language model. *J. Mach. Learn. Res.* **2003**, *3*, 1137–1155.
- Mikolov, T.; Sutskever, I.; Chen, K.; Corrado, G.S.; Dean, J. Distributed representations of words and phrases and their compositionality. In Proceedings of the 27th Annual Conference on Neural Information Processing Systems, Lake Tahoe, NV, USA, 5–8 December 2013; pp. 3111–3119.
- Guzmán, F.; Joty, S.; Márquez, L.; Nakov, P. Pairwise neural machine translation evaluation. *arXiv* **2019**, arXiv:1912.03135.
- Devlin, J.; Zbib, R.; Huang, Z.; Lamar, T.; Schwartz, R.; Makhoul, J. Fast and robust neural network joint models for statistical machine translation. In Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics, Baltimore, MD, USA, 22–27 June 2014; Volume 1, pp. 1370–1380.
- Bahdanau, D.; Cho, K.; Bengio, Y. Neural machine translation by jointly learning to align and translate. *arXiv* **2014**, arXiv:1409.0473.
- Cho, K.; Van Merriënboer, B.; Gulcehre, C.; Bahdanau, D.; Bougares, F.; Schwenk, H.; Bengio, Y. Learning phrase representations using RNN encoder-decoder for statistical machine translation. *arXiv* **2014**, arXiv:1406.1078.
- Gupta, S.; Agrawal, A.; Gopalakrishnan, K.; Narayanan, P. Deep learning with limited numerical precision. In Proceedings of the International Conference on Machine Learning 2015, Lille, France, 6–11 July 2015; pp. 1737–1746.
- Mouratidis, D.; Keramidis, K.L. Automatic selection of parallel data for machine translation. In Proceedings of the IFIP International Conference on Artificial Intelligence Applications and Innovations, Rhodes, Greece, 25–27 May 2018; pp. 146–156.
- Chawla, N.V.; Bowyer, K.W.; Hall, L.O.; Kegelmeyer, W.P. SMOTE: Synthetic minority over-sampling technique. *J. Artif. Intell. Res.* **2002**, *16*, 321–357. [[CrossRef](#)]
- Papineni, K.; Roukos, S.; Ward, T.; Zhu, W.J. BLEU: A method for automatic evaluation of machine translation. In Proceedings of the 40th annual meeting of the Association for Computational Linguistics, Philadelphia, PA, USA, 7–12 July 2002; pp. 311–318.
- Doddington, G. Automatic evaluation of machine translation quality using n-gram co-occurrence statistics. In Proceedings of the Second International Conference on Human Language Technology Research, San Diego, CA, USA, 24–27 March 2002; pp. 138–145.
- Su, K.Y.; Wu, M.W.; Chang, J.S. A new quantitative quality measure for machine translation systems. In Proceedings of the 15th International Conference on Computational Linguistics, Nantes, France, 23–28 August 1992.
- Denkowski, M.; Lavie, A. Meteor universal: Language specific translation evaluation for any target language. In Proceedings of the Ninth Workshop on Statistical Machine Translation, Baltimore, MD, USA, 26–27 June 2014; pp. 376–380.
- Snover, M.; Dorr, B.; Schwartz, R. Language and translation model adaptation using comparable corpora. In Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing, Honolulu, HI, USA, 25–27 October 2008; pp. 857–866.
- Shimnaka, H.; Kajiwara, T.; Komachi, M. Ruse: Regressor using sentence embeddings for automatic machine translation evaluation. In Proceedings of the Third Conference on Machine Translation: Shared Task Papers, Belgium, Brussels, 31 October–1 November 2018; pp. 751–758.



17. Duh, K. Ranking vs. regression in machine translation evaluation. In Proceedings of the Third Workshop on Statistical Machine Translation, Columbus, OH, USA, 19 June 2008; pp. 191–194.
18. Hochreiter, S.; Schmidhuber, J. Long short-term memory. *Neural Comput.* **1997**, *9*, 1735–1780. [CrossRef]
19. Sutskever, I.; Vinyals, O.; Le, Q.V. Sequence to sequence learning with neural networks. In Proceedings of the NIPS 2014, Montreal, QC, Canada, 8–13 December 2014; pp. 3104–3112.
20. Wu, Y.; Schuster, M.; Chen, Z.; Le, Q.V.; Norouzi, M.; Macherey, W.; Krikun, M.; Cao, Y.; Gao, Q.; Macherey, K.; et al. Google's neural machine translation system: Bridging the gap between human and machine translation. *arXiv* **2016**, arXiv:1609.08144.
21. Mouratidis, D.; Kermanidis, K.L.; Sosoni, V. Innovative Deep Neural Network Fusion for Pairwise Translation Evaluation. In Proceedings of the IFIP International Conference on Artificial Intelligence Applications and Innovations, Neos Marmaras, Greece, 5–7 June 2020; pp. 76–87.
22. LeCun, Y.; Bengio, Y.; Hinton, G. Deep learning. *Nature* **2015**, *521*, 436–444. [CrossRef]
23. Gehring, J.; Auli, M.; Grangier, D.; Yarats, D.; Dauphin, Y.N. Convolutional sequence to sequence learning. *arXiv* **2017**, arXiv:1705.03122.
24. Dauphin, Y.N.; Fan, A.; Auli, M.; Grangier, D. Language modeling with gated convolutional networks. In Proceedings of the International Conference on Machine Learning, Sydney, Australia, 6–11 August 2017; pp. 933–941.
25. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep residual learning for image recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 770–778.
26. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, L.; Polosukhin, I. Attention is all you need. *Adv. Neural Inf. Process. Syst.* **2017**, *30*, 5998–6008.
27. Bradbury, J.; Merity, S.; Xiong, C.; Socher, R. Quasi-recurrent neural networks. *arXiv* **2016**, arXiv:1611.01576.
28. Kalchbrenner, N.; Espeholt, L.; Simonyan, K.; Oord, A.V.d.; Graves, A.; Kavukcuoglu, K. Neural machine translation in linear time. *arXiv* **2016**, arXiv:1610.10099.
29. Kordoni, V.; Birch, L.; Buliga, I.; Cholakov, K.; Egg, M.; Gaspari, F.; Georgakopoulou, Y.; Gialama, M.; Hendrickx, I.; Jermol, M.; et al. TraMOOC (translation for massive open online courses): Providing reliable MT for MOOCs. In Proceedings of the Annual conference of the European Association for Machine Translation 2016, Riga, Latvia, 30 May–1 June 2016; p. 396.
30. Koehn, P.; Hoang, H.; Birch, A.; Callison-Burch, C.; Federico, M.; Bertoldi, N.; Cowan, B.; Shen, W.; Moran, C.; Zens, R.; et al. Moses: Open source toolkit for statistical machine translation. In Proceedings of the 45th Annual Meeting of the ACL on Interactive Poster and Demonstration Sessions, Stroudsburg, PA, USA, 25–27 June 2007; pp. 177–180.
31. Palmquist, R. Translation System. U.S. Patent 10/234,015, 4 March 2004.
32. Sennrich, R.; Firat, O.; Cho, K.; Birch, A.; Haddow, B.; Hitschler, J.; Junczys-Dowmunt, M.; Läubli, S.; Barone, A.V.M.; Mokry, J.; et al. Nematius: A toolkit for neural machine translation. *arXiv* **2017**, arXiv:1703.04357.
33. Barone, A.V.M.; Haddow, B.; Germann, U.; Sennrich, R. Regularization techniques for fine-tuning in neural machine translation. *arXiv* **2017**, arXiv:1707.09920.
34. Rama, T.; Borin, L.; Mikros, G.; Macutek, J. Comparative evaluation of string similarity measures for automatic language classification. In *Sequences in Language and Text*; De Gruyter Mouton: Berlin, Germany, 2015.
35. Poulliquen, B.; Steinberger, R.; Ignat, C. Automatic identification of document translations in large multilingual document collections. *arXiv* **2006**, arXiv:cs/0609060.
36. Barrón-Cedeño, A.; Márquez Villodre, L.; Henríquez Quintana, C.A.; Formiga Fanals, L.; Romero Merino, E.; May, J. Identifying useful human correction feedback from an on-line machine translation service. In Proceedings of the 23rd International Joint Conference on Artificial Intelligence, Beijing, China, 5–9 August 2013; pp. 2057–2063.
37. Mouratidis, D.; Kermanidis, K.L. Ensemble and deep learning for language-independent automatic selection of parallel data. *Algorithms* **2019**, *12*, 26. [CrossRef]
38. Loper, E.; Bird, S. NLTK: The natural language toolkit. *arXiv* **2002**, arXiv:cs/0205028.
39. Sergio, G.C. gcunhase/NLPMetrics: The Natural Language Processing Metrics Python Repository. *Zenodo* **2019**. [CrossRef]
40. Keras. Deep Learning Library for Theano and Tensorflow 2015. Available online: <https://keras.io/k> (accessed on 11 January 2021).
41. Outsios, S.; Karatsalos, C.; Skianis, K.; Vazirgiannis, M. Evaluation of Greek Word Embeddings. *arXiv* **2019**, arXiv:1904.04032.
42. Yamada, I.; Asai, A.; Shindo, H.; Takeda, H.; Takefuji, Y. Wikipedia2Vec: An optimized tool for learning embeddings of words and entities from Wikipedia. *arXiv* **2018**, arXiv:1812.06280.
43. Kingma, D.P.; Ba, J. Adam: A method for stochastic optimization. *arXiv* **2014**, arXiv:1412.6980.
44. Kuhn, M.; Johnson, K. *Applied Predictive Modeling*; Springer: Berlin/Heidelberg, Germany, 2013; Volume 26.
45. Powers, D.M. Evaluation: From precision, recall and F-measure to ROC, informedness, markedness and correlation. *Int. J. Mach. Learn. Technol.* **2011**, *2*, 37–63.
46. Guilford, J.P. *Psychometric Methods*; McGraw-Hill: New York, NY, USA, 1954.
47. Soricut, R.; Brill, E. Automatic question answering: Beyond the factoid. In Proceedings of the Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics: HLT-NAACL 2004, Boston, MA, USA, 2–7 May 2004; pp. 57–64.
48. Smialowski, P.; Frishman, D.; Kramer, S. Pitfalls of supervised feature selection. *Bioinformatics* **2010**, *26*, 440–443. [CrossRef]
49. Chandrashekar, G.; Sahin, F. A survey on feature selection methods. *Comput. Electr. Eng.* **2014**, *40*, 16–28. [CrossRef]

50. Van Der Maaten, L.; Postma, E.; Van den Herik, J. Dimensionality reduction: A comparative. *J. Mach. Learn. Res.* **2009**, *10*, 13.
51. Breiman, L. Bagging predictors. *Mach. Learn.* **1996**, *24*, 123–140. [[CrossRef](#)]
52. Koller, D.; Sahami, M. *Toward Optimal Feature Selection*; Technical Report; Stanford InfoLab: Stanford, CA, USA, 1996.
53. Molina, L.C.; Belanche, L.; Nebot, A. Feature selection algorithms: A survey and experimental evaluation. In Proceedings of the 2002 IEEE International Conference on Data Mining, Maebashi, Japan, 9–12 December 2002; pp. 306–313.
54. Liu, H.; Motoda, H. *Feature Selection for Knowledge Discovery and Data Mining*; Springer Science & Business Media: Berlin/Heidelberg, Germany, 2012; Volume 454.
55. Wold, S.; Esbensen, K.; Geladi, P. Principal component analysis. *Chemom. Intell. Lab. Syst.* **1987**, *2*, 37–52. [[CrossRef](#)]

MDPI  
St. Alban-Anlage 66  
4052 Basel  
Switzerland  
Tel. +41 61 683 77 34  
Fax +41 61 302 89 18  
[www.mdpi.com](http://www.mdpi.com)

*Applied Sciences* Editorial Office  
E-mail: [appls@mdpi.com](mailto:appls@mdpi.com)  
[www.mdpi.com/journal/appls](http://www.mdpi.com/journal/appls)





MDPI  
St. Alban-Anlage 66  
4052 Basel  
Switzerland

Tel: +41 61 683 77 34  
Fax: +41 61 302 89 18

[www.mdpi.com](http://www.mdpi.com)



ISBN 978-3-0365-1287-7