

Special Issue Reprint

Machine Learning and Medicine

The Interface of Medicine, Engineering
and Artificial Intelligence

Edited by
Simon W. Rabkin

mdpi.com/journal/bioengineering

Machine Learning and Medicine: The Interface of Medicine, Engineering and Artificial Intelligence

Machine Learning and Medicine: The Interface of Medicine, Engineering and Artificial Intelligence

Guest Editor

Simon W. Rabkin



Basel • Beijing • Wuhan • Barcelona • Belgrade • Novi Sad • Cluj • Manchester

Guest Editor

Simon W. Rabkin

Department of Medicine

University of British Columbia

Vancouver

Canada

Editorial Office

MDPI AG

Grosspeteranlage 5

4052 Basel, Switzerland

This is a reprint of the Special Issue, published open access by the journal *Bioengineering* (ISSN 2306-5354), freely accessible at: https://www.mdpi.com/journal/bioengineering/special_issues/136V9755TU.

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

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

ISBN 978-3-7258-4563-7 (Hbk)

ISBN 978-3-7258-4564-4 (PDF)

<https://doi.org/10.3390/books978-3-7258-4564-4>

© 2025 by the authors. Articles in this book are Open Access and distributed under the Creative Commons Attribution (CC BY) license. The book as a whole is distributed by MDPI under the terms and conditions of the Creative Commons Attribution-NonCommercial-NoDerivs (CC BY-NC-ND) license (<https://creativecommons.org/licenses/by-nc-nd/4.0/>).

Contents

About the Editor	vii
Preface	ix
Iciar Usategui, Yoel Arroyo, Ana María Torres, Julia Barbado and Jorge Mateo	
Systemic Lupus Erythematosus: How Machine Learning Can Help Distinguish between Infections and Flares	
Reprinted from: <i>Bioengineering</i> 2024, 11, 90, https://doi.org/10.3390/bioengineering11010090 .	1
Simon W Rabkin	
Searching for the Best Machine Learning Algorithm for the Detection of Left Ventricular Hypertrophy from the ECG: A Review	
Reprinted from: <i>Bioengineering</i> 2024, 11, 489, https://doi.org/10.3390/bioengineering11050489 .	17
Roya Zandi, Joseph D. Fahey, Michael Drakopoulos, John M. Bryan, Siyuan Dong, Paul J. Bryar, Ann E. Bidwell, et al.	
Exploring Diagnostic Precision and Triage Proficiency: A Comparative Study of GPT-4 and Bard in Addressing Common Ophthalmic Complaints	
Reprinted from: <i>Bioengineering</i> 2024, 11, 120, https://doi.org/10.3390/bioengineering11020120 .	33
Borislava Toleva, Ivan Atanasov, Ivan Ivanov and Vincent Hooper	
An Effective Methodology for Diabetes Prediction in the Case of Class Imbalance	
Reprinted from: <i>Bioengineering</i> 2025, 12, 35, https://doi.org/10.3390/bioengineering12010035 .	42
Carlo Metta, Andrea Beretta, Roberto Pellungrini, Salvatore Rinzivillo and Fosca Giannotti	
Towards Transparent Healthcare: Advancing Local Explanation Methods in Explainable Artificial Intelligence	
Reprinted from: <i>Bioengineering</i> 2024, 11, 369, https://doi.org/10.3390/bioengineering11040369 .	59
Akmalbek Abdusalomov, Sanjar Mirzakhililov, Zaripova Dilnoza, Kudratjon Zohirov, Rashid Nasimov, Sabina Umirzakova and Young-Im Cho	
Lightweight Super-Resolution Techniques in Medical Imaging: Bridging Quality and Computational Efficiency	
Reprinted from: <i>Bioengineering</i> 2024, 11, 1179, https://doi.org/10.3390/bioengineering11121179 .	76
Daniel Parres, Alberto Albiol and Roberto Paredes	
Improving Radiology Report Generation Quality and Diversity through Reinforcement Learning and Text Augmentation	
Reprinted from: <i>Bioengineering</i> 2024, 11, 351, https://doi.org/10.3390/bioengineering11040351 .	91
Erum Yousef Abbasi, Zhongliang Deng, Arif Hussain Magsi, Qasim Ali, Kamlesh Kumar and Asma Zubedi	
Optimizing Skin Cancer Survival Prediction with Ensemble Techniques	
Reprinted from: <i>Bioengineering</i> 2024, 11, 43, https://doi.org/10.3390/bioengineering11010043 .	106
Rama Krishna Thelagathoti, Dinesh S. Chandel, Wesley A. Tom, Chao Jiang, Gary Krzyzanowski, Appolinaire Olou and M. Rohan Fernando	
Machine Learning-Based Ensemble Feature Selection and Nested Cross-Validation for miRNA Biomarker Discovery in Usher Syndrome	
Reprinted from: <i>Bioengineering</i> 2025, 12, 497, https://doi.org/10.3390/bioengineering12050497 .	132

About the Editor

Simon W. Rabkin

Simon W. Rabkin is a Full Professor of Medicine in the Division of Cardiology at the University of British Columbia. Dr. Rabkin received his MD degree from the University of Manitoba and completed postgraduate training in cardiology at the University of Cincinnati and at Emory University. His research interests span the range from molecular/cellular pathways to cardiovascular disease, as well as clinical and population science. He has published over 300 scientific papers on a range of topics including brain–heart interconnection, heart failure, risk factors for coronary artery disease, and machine learning and AI.

He was the founder and first director of the program in experimental medicine which is now one of the largest postgraduate programs in the Faculty of Medicine at the University of British Columbia. He has served on various committees and is the former President of the Vancouver Hospital Medical staff and Canadian Hypertension Society. He now serves as an associate editor of several cardiovascular journals. Throughout his academic career he has received multiple awards.

Preface

This Reprint presents a range of studies that comprehensively explore the breadth of the techniques used in machine learning and highlighting the advances being made in the integration of these approaches to medicine. The contributions in this Reprint cover various aspects of application such as diagnosis, imaging, and assessing prognosis, collectively demonstrating the improvements in diagnostic accuracy and medical decision making that can be achieved by machine learning approaches.

In the field of diagnosis, Usategi and colleagues utilized the machine learning approach to diagnose an exacerbation of systemic lupus erythematosus and distinguish it from an infection. Their machine learning algorithm displayed superior sensitivity/specificity, as indicated by findings of area under the curve (AUC) analysis, as well as a high accuracy. This remarkable piece of research has already been viewed by almost 3,000 individuals, highlighting its groundbreaking nature. In addition, the diagnosis of cardiac (left ventricular) hypertrophy is of considerable importance, as this condition increases the risk of myocardial ischemia and fatal cardiac events. In relation to this, an evaluation of the different machine learning efforts to diagnose left ventricular hypertrophy using electrocardiogram has been presented, having already amassed over 2,000 views by individuals.

Zandi and colleagues developed a method to evaluate diagnostic precision and triage proficiency in order to investigate the capabilities of AI chatbots for ophthalmic diagnosis and triage. They found that chatbots were significantly better at ophthalmic triage than diagnosis and GPT-4 performed better than Bard for appropriate triage. They therefore concluded that these tools present potential utility in aiding patients or triage staff; however, they are not a replacement for professional ophthalmic evaluation or advice.

Toleva and colleagues present a novel methodology for predicting whether an individual would develop diabetes over time given a set of biological and social indicators and the proposed algorithms create effective classification models to predict the risk of diabetes. In an article that has already been viewed by almost 3,000 individuals, Metta and colleagues focus on the use of local Explainable Artificial Intelligence (XAI) methods, particularly the Local Rule-Based Explanations (LORE) technique, in healthcare. They concluded that XAI can significantly contribute to improved clinical decision making.

With regard to imaging, Akmalbeck and colleagues propose an enhanced Residual Feature Learning Network (RFLN) tailored for medical imaging. Their contributions include replacing the residual local feature blocks with standard residual blocks, increasing the model depth for improved feature extraction, and incorporating enhanced spatial attention mechanisms to refine the feature selection. They conclude that enhanced RFLN effectively mitigates noise and also preserves critical anatomical details, making it a promising solution for high-precision medical imaging applications.

Parres and colleagues set forth an improved radiology report by utilizing reinforcement learning and text augmentation to tackle issues. Their approach is shown to significantly improve report quality and variability, enhancing diagnostic precision and the quality of radiological interpretations.

In terms of assessing prognosis, Zubedi and colleagues used their algorithm to evaluate the survival of cutaneous melanoma cancers. Comparing their proposed approach with existing state-of-the-art techniques, they found significant improvements in several key aspects of accuracy and efficiency. In addition, Thelagathoti and colleagues employed a machine learning-based ensemble feature selection and nested cross-validation approach for miRNA biomarker discovery.

Together, these approaches are guideposts on the road to expanding the utilization of machine learning to improve disease diagnosis and patient care.

Simon W. Rabkin

Guest Editor

Article

Systemic Lupus Erythematosus: How Machine Learning Can Help Distinguish between Infections and Flares

Iciar Usategui ¹, Yoel Arroyo ², Ana María Torres ^{3,4}, Julia Barbado ⁵ and Jorge Mateo ^{3,4,*}

¹ Department of Internal Medicine, Hospital Clínico Universitario, 47005 Valladolid, Spain

² Department of Technologies and Information Systems, Faculty of Social Sciences and Information Technologies, Universidad de Castilla-La Mancha (UCLM), 45600 Talavera de la Reina, Spain

³ Medical Analysis Expert Group, Institute of Technology, Universidad de Castilla-La Mancha (UCLM), 16071 Cuenca, Spain

⁴ Medical Analysis Expert Group, Instituto de Investigación Sanitaria de Castilla-La Mancha (IDISCAM), 45071 Toledo, Spain

⁵ Department of Internal Medicine, Hospital Universitario Río Hortega, 47012 Valladolid, Spain

* Correspondence: jorge.mateo@uclm.es

Abstract: Systemic Lupus Erythematosus (SLE) is a multifaceted autoimmune ailment that impacts multiple bodily systems and manifests with varied clinical manifestations. Early detection is considered the most effective way to save patients' lives, but detecting severe SLE activity in its early stages is proving to be a formidable challenge. Consequently, this work advocates the use of Machine Learning (ML) algorithms for the diagnosis of SLE flares in the context of infections. In the pursuit of this research, the Random Forest (RF) method has been employed due to its performance attributes. With RF, our objective is to uncover patterns within the patient data. Multiple ML techniques have been scrutinized within this investigation. The proposed system exhibited around a 7.49% enhancement in accuracy when compared to k-Nearest Neighbors (KNN) algorithm. In contrast, the Support Vector Machine (SVM), Binary Linear Discriminant Analysis (BLDA), Decision Trees (DT) and Linear Regression (LR) methods demonstrated inferior performance, with respective values around 81%, 78%, 84% and 69%. It is noteworthy that the proposed method displayed a superior area under the curve (AUC) and balanced accuracy (both around 94%) in comparison to other ML approaches. These outcomes underscore the feasibility of crafting an automated diagnostic support method for SLE patients grounded in ML systems.

Keywords: Systemic Lupus Erythematosus; medical treatment; machine learning; artificial intelligence

1. Introduction

Systemic Lupus Erythematosus (SLE) is a chronic autoimmune affliction that affects various physiological systems. It serves as an exemplary autoimmune disorder, and its intricate nature poses significant challenges. The varied clinical presentations of SLE, coupled with distinct complexities in both diagnosis and treatment, present a formidable task for healthcare professionals. The emergence of multiple mechanisms results in the breakdown of self-tolerance and subsequent organ dysfunction. Progress in elucidating the molecular and cellular foundations of this condition, in conjunction with the identification of genetic variations, contributes to a more profound comprehension of its pathogenesis, offering promise for therapeutic advancements in the near future.

Commonly known as lupus, it varies in prevalence depending on geographic location, ethnicity, and research study design. In the United States, an estimated 241 cases per 100,000 adults have been reported, while in Spain, the updated figure is 210 cases per 100,000 inhabitants [1]. The Lupus Foundation of America estimates that approximately 161,000 to 322,000 individuals in the U.S. are affected by SLE, translating to a prevalence

of approximately 0.05% to 0.1% of the population. Predominantly, it affects young, fertile females and has resulted in increased mortality, although improved treatment modalities have positively impacted survival rates. Notably, the onset of the disease frequently occurs during the childbearing years. Certain demographic groups, including women, people of color (particularly African American, Hispanic, and Asian populations), and individuals of reproductive age, may experience higher prevalence rates. Simultaneously, several factors contribute to a state of relative immunodeficiency in individuals with SLE, including aging, the increasing use of targeted biologic therapies, and the chronic nature of the disease. Furthermore, the presence of other comorbidities such as malignancy, infections, malnutrition, and more further compounds the complexity of the disease. SLE is a complex and heterogeneous condition, manifesting symptoms across a spectrum from mild to severe. The precise etiology of SLE remains not fully understood, with its development believed to result from a combination of genetic and environmental factors. Moreover, the prevalence of SLE may undergo changes over time, influenced by factors such as improvements in diagnostic methods and increased awareness of the disease. Collectively, these multifaceted factors underscore the need for a comprehensive understanding of the diverse epidemiological and clinical aspects of SLE to inform effective management strategies and interventions.

Emerging evidence suggests that immunodeficiency and systemic autoimmunity are interconnected manifestations of a shared underlying process [2]. Immune disorders present as both susceptibility to infections and autoimmune symptoms, indicating a dual impact on the immune system—reduced ability to clear infections and a disruption of self-tolerance. On the other hand, infections are one of the most common causes of death and are often associated with high levels of activity in SLE. Early diagnosis of immunodeficiency in SLE is the first step to contribute to detect infections, which are likely to be associated with flares, allows prompt initiation of treatment, a better prognosis, and a reduction in organ dysfunction [3–7]. In the absence of specific criteria that can differentiate between a severe infection and an exacerbation in SLE, the development of clinical studies and guidelines becomes imperative to facilitate a more precise classification of these patients [8].

In pursuit of this objective, Machine Learning (ML) draws inspiration from biological nervous systems. Its fundamental principle revolves around presenting algorithms with input data, subjecting them to computer analysis to predict output values within an acceptable range of accuracy, recognizing data patterns and trends, and ultimately assimilating knowledge from prior experiences [9]. ML delves into intricate data distributions, establishes probabilistic relationships, and identifies the minimum set of features required to capture essential data patterns through repeated cross-validation, culminating in the formulation of predictive models. Numerous studies have leveraged ML methods to develop more precise diagnostic algorithms for stratifying autoimmune diseases, thereby preventing or mitigating observed morbidity [10]. ML methods consistently exhibit superior performance compared to traditional statistical models [9,11–13]. A variety of ML techniques, including Support Vector Machine (SVM), Binary Linear Discriminant Analysis (BLDA), k-Nearest Neighbors (KNN), and Decision Trees (DT) [14–17], have been employed for data analysis. These systems represent a selection of algorithms designed for classifying data and processing information, and they have been explored in the context of various autoimmune diseases, including SLE, rheumatoid arthritis, lupus tubulointerstitial inflammation, and neuropsychiatric SLE [18–23].

In this paper, we present a system that utilizes the Random Forest (RF) method for the analysis of immunodeficiency patterns in SLE patients. RF is an ML algorithm that operates by constructing a multitude of decision trees for classification and prediction. For its capacity to enhance accuracy and processing speed, and several notable advantages, including a low computational burden, flexibility in model tuning, high scalability, and algorithmic optimization, it serves as the cornerstone of this approach. Through the application of RF, we aim to predict the immunodeficiency status of our patients, with

the overarching goal of not only identifying optimal treatment options but also designing personalized preventive measures and tailoring patient-specific follow-up strategies.

The paper is structured as follows. The first section outlines the topic, purpose, and significance of this study. Second section introduces a detailed description of material and methods. Third section entails the main findings of the study, including data, analysis, and interpretation of the results obtained. Fourth section explores a discussion of these results. And finally, the paper concludes with a summary of the research and some concluding remarks.

2. Materials and Methods

2.1. Materials

The study cohort included 125 patients who met the American College of Rheumatology criteria for SLE in 2019 [23]. These individuals were enrolled from the Autoimmune Unit Registry at Valladolid Clinic Hospital (HCUV) between 2017 and 2019. The experimental protocol adhered to the principles outlined in the Declaration of Helsinki (2008) and received approval from the Clinical Research Ethics Committee of the HCUV. The study was conducted in compliance with Spanish data protection laws (LO 15/1999) and specifications (RD 1720/2007).

Consequently, a retrospective review of patients was systematically conducted, encompassing the collection of epidemiological, analytical, immunological, and clinical characteristics. Relevant immunological parameters for evaluating immune competence included leucocytes, neutrophils, CD3, CD4 and CD8 T-cell counts, CD19 B-cell and Natural Killer (NK) cell levels, serum immunoglobulin isotypes (IgG, IgA, IgM), IgG subclasses, and complement levels (C3, C4). Exclusion criteria involved patients with evidence of active disease (SLEDAI ≥ 4) or significant residual proteinuria (>500 mg). Following this selection strategy, 31 patients were excluded from the study.

Flow cytometry was performed to identify cell populations. Serum levels of immunoglobulin isotypes and IgG subclasses and complement were determined by nephelometry. Standardized reference ranges from the immunology laboratory of our institution were used to define control patients. Laboratory levels below the reference ranges were considered as possible immunodeficiency status: leucocytes < 4000 cL/ μ L, neutrophils < 1800 cL/ μ L, lymphocytes < 1500 cL/ μ L, CD3 T-cell < 700 cL/ μ L, CD19 B-cell < 100 cL/ μ L, CD4 T-cell < 300 cL/ μ L, CD8 T-cell < 200 cL/ μ L, NK cell < 90 cL/ μ L, IgG < 870 mg/dL, IgG1 < 383 mg/dL, IgG2 < 242 mg/dL, IgG3 < 22 mg/dL, IgG4 < 4 mg/dL, IgA < 117 mg/dL, IgM < 60 mg/dL, C3 < 90 mg/dL, C4 < 10 mg/dL; special data for patients between 14 and 18 years old were: IgG1 < 315 mg/dL, IgG2 < 242 mg/dL, IgG3 < 23 mg/dL, IgG4 < 11 mg/dL. Severe infection was defined as infection which required hospitalization of seriousness, treatment needed or recommended monitoring.

2.2. Method

This study introduces an ML method centered on the Random Forest (RF) algorithm. RF, a widely adopted ML algorithm within supervised learning, is applied for both classification and regression challenges in ML. Renowned for its simplicity, versatility, and robustness, RF embodies a potent ML algorithm with several noteworthy attributes: (1) operative as an ensemble learning approach, it combines decisions from multiple models to improve overall performance; (2) employing decision trees as base-level models; (3) mitigating overfitting by averaging results across several trees, thereby diminishing the risk of developing complex models performing well on training data but poorly on new data; (4) adeptly handling missing values by learning the optimal imputation value based on the reduction in the utilized criterion; (5) furnishing a reliable estimate of the importance of variables in the classification process; (6) demonstrating flexibility in its applicability to both regression and classification tasks; and (7) executing swiftly with minimal preprocessing requirements compared to alternative algorithms, capable of handling categorical variables

without necessitating the creation of dummy variables. Consequently, RF is the chosen algorithm for crafting the model aimed at detecting immunodeficiency patterns within the SLE population [24,25].

Given a dataset $S = \{x_j, y_j\}$, where x_j represents feature vectors and y_j corresponds to labels, the RF algorithm proceeds as follows:

For each of the n trees in the forest:

1. Draw a bootstrap sample Z^* of size N from the training data.
2. Grow a decision tree T_b to the bootstrapped data by recursively repeating the following steps for each terminal node of the tree, until the minimum node size n_{\min} is reached:
 - (a) Select m variables at random from the p variables.
 - (b) Pick the best variable/split-point among the m variables.
 - (c) Split the node into two daughter nodes.

The prediction of the RF then aggregates the predictions of the n trees.

For regression, it is typically the average over all trees:

$$\hat{f}_{rf}(x) = \frac{1}{n} \sum_{b=1}^n T_b(x) \quad (1)$$

For classification, it is determined by the majority vote:

$$\hat{C}_{rf}(x) = \text{majority}\{\hat{C}_b(x)\}_1^n \quad (2)$$

Here, $T_b(x)$ and $\hat{C}_b(x)$ represent the prediction of the b -th decision tree for regression and classification, respectively.

The algorithm was designed and developed using Matlab software (MatLab 2023a, The Mathworks Inc., Natick, MA, USA). Furthermore, the proposed system underwent analysis alongside other ML systems prevalent in the scientific community. These included Support Vector Machine (SVM) [14], Binary Linear Discriminant Analysis (BLDA) [26], Decision Trees (DT) [15], Linear Regression (LR) [27,28], and k-Nearest Neighbor (KNN) [16] to assess its performance. Within the ML system's learning process, it is imperative to control overtraining. To address this, the k-fold cross-validation technique was employed in our case.

As depicted in Figure 1, each iteration involves the random classification of 70% of the patients for training and 30% for testing and validation. Notably, patient data are not shared between the training and validation subsets to prevent the algorithm from being validated with data from the same patients used in the training phase.

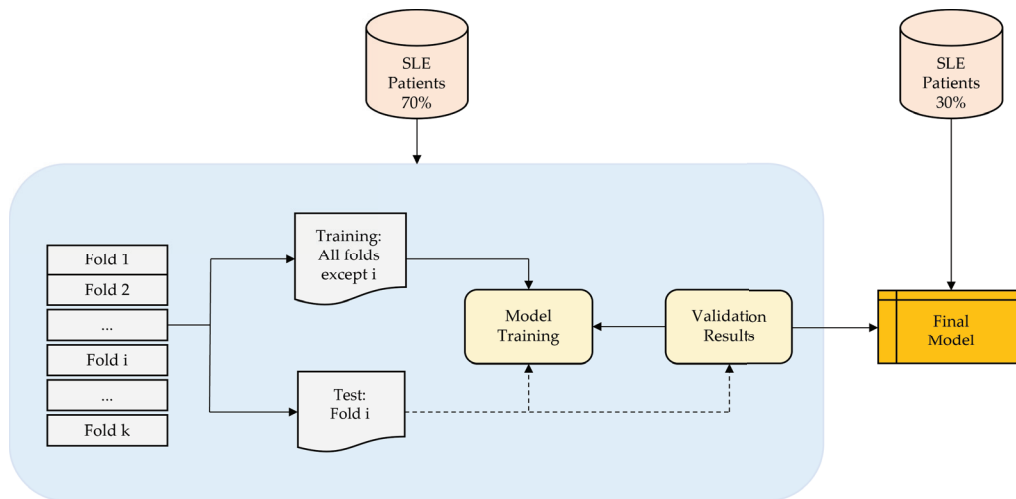


Figure 1. The figure shows the processes followed in this study for the classification of patients with SLE.

Additionally, techniques for hyperparameter optimization have been applied to fine-tune the hyperparameters of the methods. These hyperparameter values are adjusted during the training phase to maximize the accuracy of the ML method. The hyperparameters subjected to optimization encompass variables such as apprentices, neighbors, distance metric, distance weight, kernel, box constraint level, and multiclass method, each tailored to the specific requirements of the method in use. Bayesian optimization was chosen as the technique to enhance the performance of the various methods by optimizing the selection of diverse hyperparameters. Recall value and AUC were utilized as performance metrics. The entire study was iterated 100 times to obtain mean values and standard deviations for the process. Importantly, it should be emphasized that data used in each iteration were randomized, mitigating noise in the samples and ensuring the acquisition of results with statistically valid values [29].

2.3. Performance Evaluation

For this study, the most well-known metrics in artificial intelligence were implemented to test the performance of the methods [29]: balanced accuracy (BA), recall, precision, specificity (SP), degenerated Youden's index (DYI) [29], receiver operating characteristic (ROC) and area under the curve (AUC). The F_1 score is established as:

$$F_1\text{score} = 2 \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}} \quad (3)$$

To test the classification performance of the model, the Matthew correlation coefficient (MCC) has been used, which is described as follows:

$$MCC = \frac{TP \cdot TN - FP \cdot FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}} \quad (4)$$

where TP is the number of true positives, FP the number of false positives, TN the number of true negatives and FN the number of false negatives. And finally, Cohen's Kappa (CK), CK is another metric that estimates the performance of the model [29].

3. Results

The study was conducted on a group of 125 patients diagnosed with SLE. Out of these, 94 patients met the specific criteria of having a SLEDAI-2K score of less than four points, and were thus included in the study. Further analysis revealed that 77 of these 94 patients showed signs of immunodeficiency. This means that approximately 81.9% of the patients with a SLEDAI-2K score less than four exhibited signs of immunodeficiency.

The cohort of patients had a median age of 52 years, whilst the median age at diagnosis was 38 years. The group was predominantly female, with 68 female patients compared to 9 male patients. The median duration of the disease among these patients was 14 years. At the time of data collection (see Table 1), 50 patients (64.9%) were being treated with corticosteroids at an average daily dose of 2.57 mg. In addition, 25 patients (34.9%) were receiving immunosuppressants such as azathioprine, methotrexate, and mycophenolate. Two patients were on belimumab treatment. Notably, none of the patients were undergoing treatment with rituximab.

In turn, 41 patients (53.2%) exhibited patterns of immunodeficiency. Among these patients, there were a total of 51 episodes of severe infections. The breakdown of these infections is as follows:

- 17 patients were hospitalized due to lower respiratory infections.
- 4 patients were hospitalized for upper respiratory infections.
- 9 patients were treated for urinary infections.
- 10 patients had soft tissue infections.
- 4 patients suffered from digestive infections.
- 1 patient was diagnosed with tuberculous lymphadenitis.

Table 1 provides an overview of the characteristics of patients exhibiting immunodeficiency patterns. The patients under study demonstrated a decline in the count of several immune cells. This was particularly evident in the case of NK cells, a component of the innate immune system, and CD19 B-cells, a part of the adaptive immune system. The latter includes IgG subclasses and IgM, both of which also showed a decrease. These patients exhibited reduced levels of various immune cells, as illustrated in Table 1, with notable decreases observed in NK cells within the innate immune system and CD19 B-cells within the adaptive immune system, including IgG subclasses and IgM.

Table 1. Characteristics of patients with immunodeficiency patterns.

Characteristics of Patients with Immunodeficiency Patterns	
N	77
Median age (years)	52
Female/Male	68/9
SLE evolution time (years)	14
Corticosteroids (n)	50 (64.9%)
Immunosuppressants (n)	25 (32.4%)
Hydroxychloroquine (n)	37 (48%)
Severe infections (n)	51
Immunodeficiency patterns (n)	
Leucocytes (<4000 cL/ μ L)	9
Lymphocytes (<1500 cL/ μ L)	28
Neutrophils (<1800 cL/ μ L)	9
CD3 (<700 cL/ μ L)	10
CD4 (<300 cL/ μ L)	6
CD8 (<200 cL/ μ L)	3
CD19 (<100 cL/ μ L)	23
NK (<90 cL/ μ L)	13
IgG (<870 mg/dL)	17
IgG1 (<383 mg/dL)	3
IgG2 (<242 mg/dL)	36
IgG3 (<22 mg/dL)	16
IgG4 (<4 mg/dL)	7
IgA (117 mg/dL)	8
IgM (<60 mg/dL)	20
C3 (<90 mg/dL)	13
C4 (<10 mg/dL)	6

The study employed a range of ML techniques to discern patterns of innate and adaptive immunodeficiency within the SLE population. The findings derived from these techniques, coupled with several ML algorithms for identifying immunodeficiency, are detailed below. Performance metrics such as BA, recall, specificity, precision, and AUC for the investigated ML methods are exhibited in Tables 2 and 3. Both tables provide a detailed summary of performance metrics for different ML methods applied to variables IgG, IgG2, IgG3, IgG4 (Table 2), and IgM, NK, CD19, CD3 (Table 3). These variables are associated with immunoglobulins and immune cell populations, whilst the ML methods evaluated include SVM, BLDA, DT, KNN, and the RF proposed method. The results offer insights into how well each ML method performs in predicting or classifying the specified immunological variables, providing a comparative analysis of their strengths in terms of these metrics. The comprehensive nature of the data facilitates an informed selection of the most suitable method for each variable based on the desired performance criteria. Of particular note is the RF proposed method, which consistently outperforms across all variables, achieving the highest accuracy. KNN also demonstrates strong performance, particularly in IgM and CD3. LR were the lowest results obtained, whilst SVM, BLDA, and DT generally exhibit competitive results but with slightly lower accuracy than RF and KNN. In summary, the evaluation underscores the robust performance of the proposed

method across the variables related to immunoglobulins and immune cell types, being the preferred model for classifying SLE patients due to its consistently high accuracy, balanced performance metrics, ensemble learning strengths, and robustness to noisy data observed.

Table 2. The table summarises the values of BA, recall, specificity, precision and AUC for variables IgG, IgG2, IgG3 and IgG4.

IgG.					
Methods	BA	Recall	Specificity	Precision	AUC
SVM	80.85	80.95	80.76	80.28	80.00
BLDA	78.11	78.20	78.02	77.55	78.00
DT	83.85	83.95	83.75	83.25	83.00
LR	70.02	69.75	68.84	68.95	68.42
RF	93.96	94.07	93.85	93.29	94.00
KNN	86.38	86.48	86.28	85.76	86.00
IgG2.					
Methods	BA	Recall	Specificity	Precision	AUC
SVM	81.85	81.95	81.76	81.27	81.00
BLDA	77.37	77.46	77.28	76.82	77.00
DT	83.16	83.26	83.06	82.57	83.00
LR	69.51	69.24	68.33	68.44	68.42
RF	94.58	94.69	94.47	93.90	94.00
KNN	85.99	86.09	85.89	85.38	86.00
IgG3.					
Methods	BA	Recall	Specificity	Precision	AUC
SVM	81.56	81.66	81.47	80.98	81.00
BLDA	79.16	79.25	79.06	78.59	79.00
DT	83.82	83.92	83.72	83.22	83.00
LR	69.44	69.17	68.27	68.38	68.42
RF	94.42	94.53	94.31	93.75	94.00
KNN	86.57	86.67	86.47	85.95	86.00
IgG4.					
Methods	BA	Recall	Specificity	Precision	AUC
SVM	81.35	81.45	81.26	80.77	81.00
BLDA	78.93	79.02	78.83	78.36	78.00
DT	83.26	83.36	83.16	82.67	83.00
LR	70.15	69.88	68.97	69.08	68.42
RF	94.50	94.61	94.39	93.83	94.00
KNN	86.07	86.17	85.97	85.46	86.00

Table 3. The table summarises the values of BA, recall, specificity, precision and AUC for variables IgM, NK, CD19 and CD3.

IgM.					
Methods	BA	Recall	Specificity	Precision	AUC
SVM	81.24	81.34	81.15	80.67	81.00
BLDA	78.11	78.20	78.02	77.55	78.00
DT	83.35	83.45	83.25	82.76	83.00
LR	69.86	69.59	68.68	68.79	68.42
RF	94.80	94.91	94.69	94.12	94.00
KNN	86.38	86.48	86.28	85.76	86.00

Table 3. Cont.

NK.					
Methods	BA	Recall	Specificity	Precision	AUC
SVM	81.06	81.16	80.97	80.49	81.00
BLDA	77.52	77.61	77.43	76.97	77.00
DT	84.84	84.94	84.74	84.24	84.00
LR	69.51	69.24	68.33	68.44	68.42
RF	94.75	94.86	94.64	94.07	94.00
KNN	86.51	86.61	86.41	85.89	86.00
CD19.					
Methods	BA	Recall	Specificity	Precision	AUC
SVM	82.21	82.31	82.12	81.63	82.00
BLDA	76.89	76.98	76.80	76.34	76.00
DT	84.04	84.14	83.94	83.44	84.00
LR	69.65	69.38	68.47	68.58	68.42
RF	94.34	94.45	94.23	93.67	94.00
KNN	85.24	85.34	85.14	84.63	85.00
CD3.					
Methods	BA	Recall	Specificity	Precision	AUC
SVM	81.46	81.56	81.37	80.88	81.00
BLDA	77.21	77.30	77.12	76.66	77.00
DT	84.16	84.26	84.06	83.56	84.00
LR	70.41	70.14	69.22	69.33	68.42
RF	95.12	95.23	95.01	94.44	95.00
KNN	86.38	86.48	86.28	85.76	86.00

Moreover, Tables 4 and 5 present performance metrics, including F_1 score, MCC, DYI, and Kappa values, for the ML methods applied. The observed values provide insights into the models' effectiveness in classifying SLE patients. Thus, in Table 4 (variables IgG, IgG2, IgG3, and IgG4), RF consistently outperforms again other methods across all metrics, exhibiting high F_1 score, MCC, DYI, and Kappa values. This suggests RF's robustness in achieving a balanced trade-off between precision and recall, capturing the model's ability to handle both positive and negative instances effectively. Again, KNN also shows competitive performance, while SVM, BLDA, and DT demonstrate slightly lower performance across these metrics, being LR the one which obtained the lowest performance values. Similar trends are observed in the variables related to immune cell types in Table 5 (IgM, NK, CD19, and CD3), where RF again demonstrates superior performance, especially notable in achieving high F_1 score and DYI values. This reinforces RF's suitability for SLE classification, indicating its ability to maintain a balance between true positives, true negatives, false positives, and false negatives. KNN also perform well, but RF consistently stands out as the top-performing model across the diverse set of variables.

For a comprehensive view of the trade-off between the true/false positive rates between the proposed system and other ML methods, the Receiver Operating Characteristic (ROC) curves were also generated. With this purpose in mind, the ROC curve is employed to quantify sensitivity and 1-specificity at various threshold levels. As illustrated in Figure 2, which shows the ROC curve for CD19 variable as example, the system that utilizes RF generates the largest area under the curve, indicating a superior level of predictive accuracy.

Table 4. The table presents the F_1 score, MCC, DYI and Kappa values for variables IgG, IgG2, IgG3 and IgG4.

IgG.				
Methods	F_1 score	MCC	DYI	Kappa
SVM	80.61	71.74	80.85	71.98
BLDA	77.87	69.31	78.11	69.54
DT	83.60	74.40	83.85	74.65
LR	70.06	64.59	69.83	64.23
RF	93.68	83.37	93.96	83.65
KNN	86.12	76.65	86.38	76.90
IgG2.				
Methods	F_1 score	MCC	DYI	Kappa
SVM	81.61	72.63	81.85	72.87
BLDA	77.14	68.65	77.37	68.88
DT	82.91	73.79	83.16	74.04
LR	69.54	64.12	69.32	63.76
RF	94.30	83.92	94.58	84.20
KNN	85.73	76.30	85.99	76.55
IgG3.				
Methods	F_1 score	MCC	DYI	Kappa
SVM	81.32	72.37	81.56	72.61
BLDA	78.92	70.24	79.16	70.47
DT	83.57	74.38	83.82	74.62
LR	69.48	64.06	69.25	63.70
RF	94.14	83.78	94.42	84.06
KNN	86.31	76.81	86.57	77.07
IgG4.				
Methods	F_1 score	MCC	DYI	Kappa
SVM	81.11	72.19	81.35	72.43
BLDA	78.69	70.03	78.93	70.27
DT	83.01	73.88	83.26	74.13
LR	70.19	64.72	69.96	64.35
RF	94.22	83.85	94.50	84.13
KNN	85.81	76.37	86.07	76.62

Table 5. The table presents the F_1 score, MCC, DYI and Kappa values for variables IgM, NK, CD19 and CD3.

IgM.				
Methods	F_1 score	MCC	DYI	Kappa
SVM	81.00	72.09	81.24	72.33
BLDA	77.87	69.31	78.11	69.54
DT	83.10	73.96	83.35	74.21
LR	69.90	64.45	69.67	64.08
RF	94.51	84.12	94.80	84.40
KNN	86.12	76.65	86.38	76.90
NK.				
Methods	F_1 score	MCC	DYI	Kappa
SVM	80.82	71.93	81.06	72.17
BLDA	77.29	68.78	77.52	69.01
DT	84.59	75.28	84.84	75.53
LR	69.54	64.12	69.32	63.76

Table 5. Cont.

NK.				
Methods	F ₁ score	MCC	DYI	Kappa
RF	94.46	84.07	94.75	84.35
KNN	86.25	76.76	86.51	77.02
CD19.				
Methods	F ₁ score	MCC	DYI	Kappa
SVM	81.97	72.95	82.21	73.19
BLDA	76.66	68.23	76.89	68.45
DT	83.79	74.57	84.04	74.82
LR	69.68	64.25	69.45	63.89
RF	94.06	83.71	94.34	83.99
KNN	84.98	75.63	85.24	75.89
CD3.				
Methods	F ₁ score	MCC	DYI	Kappa
SVM	81.22	72.28	81.46	72.52
BLDA	76.98	68.51	77.21	68.74
DT	83.91	74.68	84.16	74.93
LR	70.45	64.96	70.22	64.59
RF	94.83	84.40	95.12	84.68
KNN	86.12	76.65	86.38	76.90

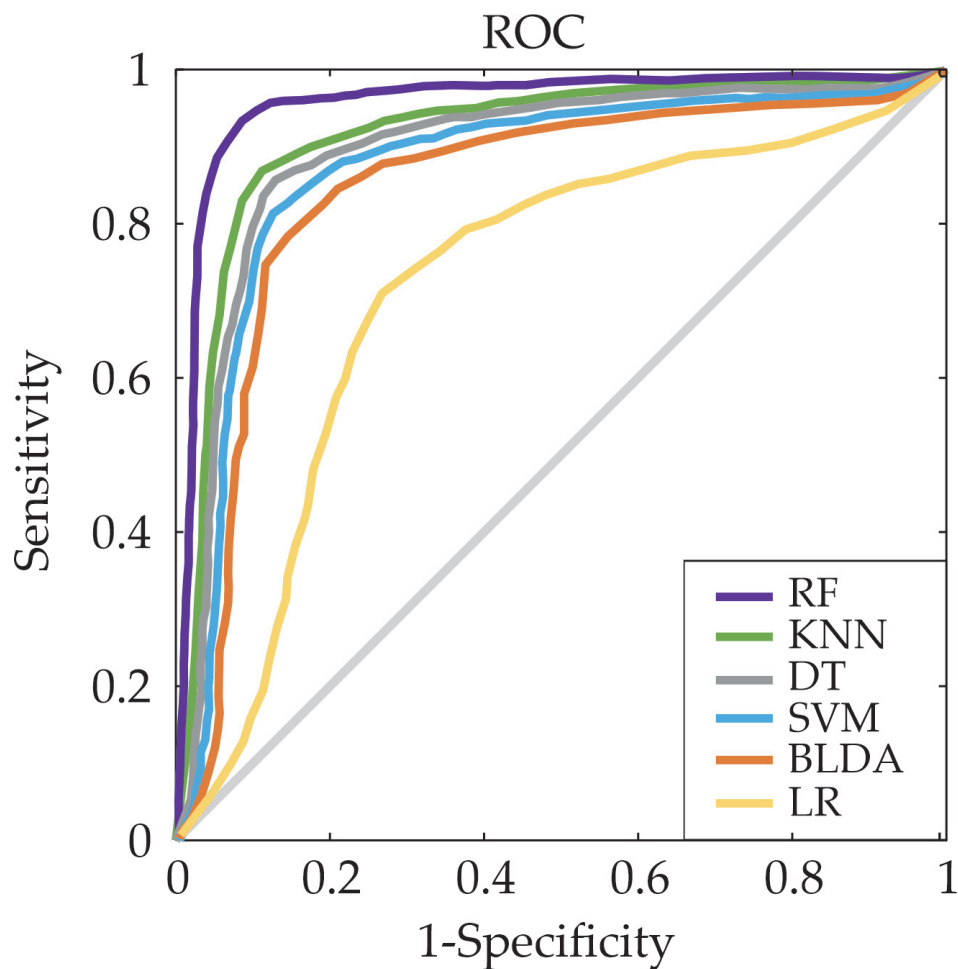


Figure 2. Example of ROC curve for the five assessed ML predictors for variable CD19.

In the study conducted, it was also observed that the subsets used for training the model exhibited high scores in the training metrics. When these models were tested, they showed a noticeable decrease in their scores. Nonetheless, as depicted in Figures 3 and 4, the RF system emerges as a well-calibrated model, attaining an optimal point in training without succumbing to overfitting or underfitting. This approach consistently delivers accurate predictions for novel inputs. The RF system's superior performance is evident, where it surpasses other methods by covering a larger area in the radar plots in both the training and testing phases.



Figure 3. The figure shows the radar plots of the variables IgG, IgG2, IgG3 and IgG4, respectively.

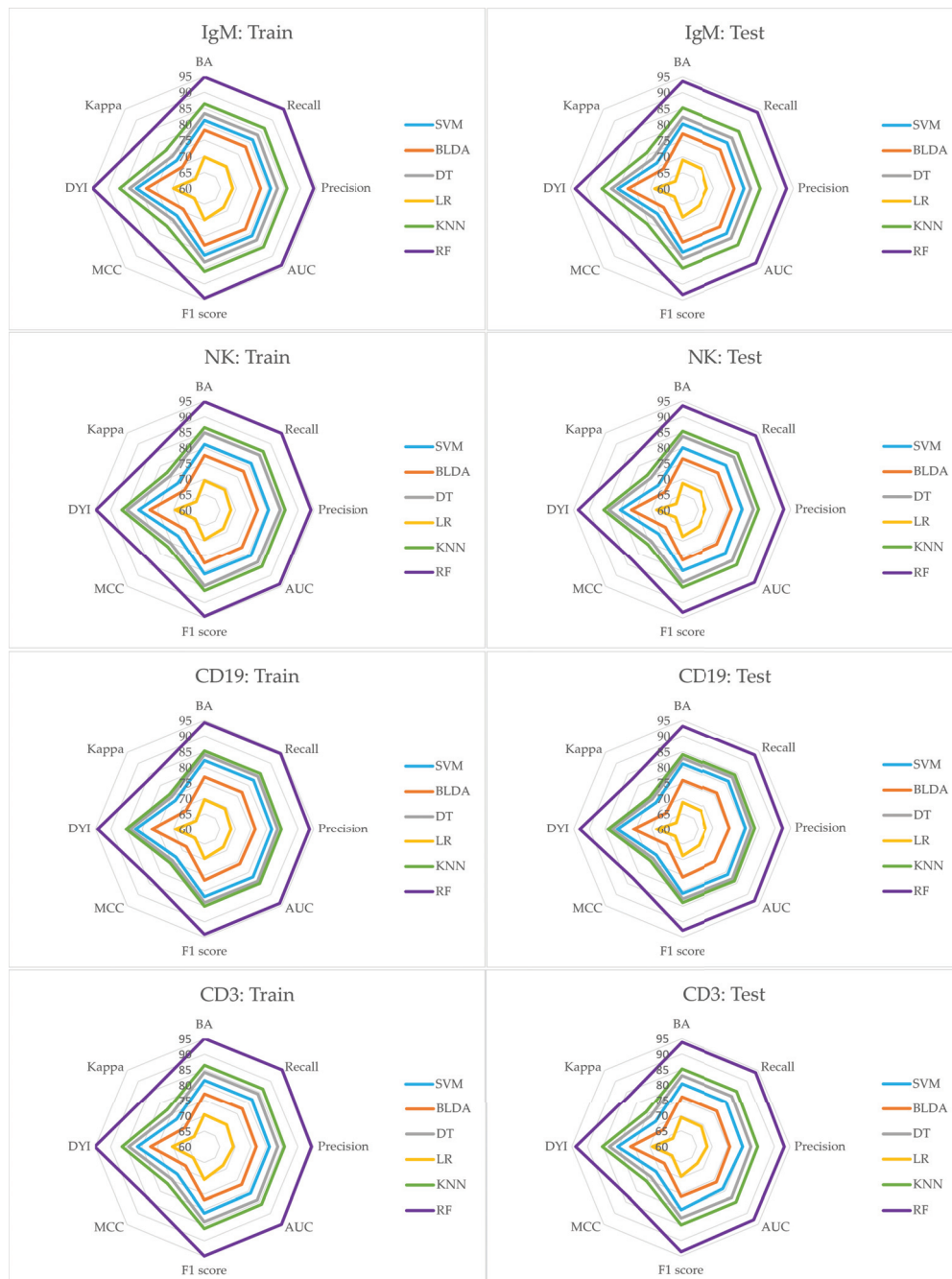


Figure 4. The figure shows the radar plots of the variables IgM, NK, CD19 and CD3, respectively.

4. Discussion

The task of managing patients with SLE is crucial in order to reduce the risk of irreversible organ damage [30,31]. This is not only vital for maintaining the health-related quality of life of the patients [32,33], but also for managing the direct costs associated with the treatment of SLE [34,35]. However, this task presents significant challenges due to the heterogeneous nature of SLE, which is characterized by variations in disease progression [36,37]. There is therefore an urgent need to improve the accuracy and classification of SLE flares, taking into account that the trigger of activity may be an infection in a situation of immunodeficiency. Numerous studies have been conducted to address this need, including recent research that has emerged over the last few years [31,33]. These studies have emphasized potential treatments for severe lupus manifestations such as lupus nephritis [31]. Despite the existence of several therapeutic agents in SLE, the disease

continues to cause significant morbidity [31]. It is encouraging that a variety of therapeutic options are currently under investigation [31].

In clinical practice, the manifestation of a malar rash, coupled with the detection of anti-DNA autoantibodies in patients, often guides healthcare professionals towards the diagnosis of SLE [38,39]. It is noteworthy that SLE is characterized by a significant degree of phenotypic diversity, which includes both systemic and localized forms. The evolution of immunological and clinical features over time underscores the dynamic nature of this disease [33,40].

A multitude of models have been established to estimate the probability of SLE occurrence, providing a degree of confidence in differentiating it from other rheumatological disorders. These models leverage unsupervised clustering based on the nature and abundance of features, mirroring diagnostic reasoning, especially during initial patient consultations [41,42]. Certain models incorporate gene analysis techniques to improve the classification of SLE patients [19]. Recent research has delved into the utilization of machine learning techniques for SLE analysis, customizing their methodologies to the specific dataset under investigation [22,43,44]. For example, Jorge et al. [20] utilized ML techniques to predict the hospitalization of SLE patients.

In the present study, the RF method, among all the ML classifiers employed, exhibited the most robust classification performance. It demonstrated superior accuracy levels and facilitated the identification of immunodeficiency patterns within the SLE population. This method offers scalability, rapid execution, and other beneficial features that enhance its classification capabilities [45]. ML models possess the capability to evaluate multiple variables and their interrelationships concurrently, accommodating non-linear patterns in the development of predictive systems [45]. Furthermore, we conducted a comparative analysis of our proposed system's performance against various ML algorithms documented in Tables 2–5. Notably, the RF method exhibited a substantial improvement, outperforming DT, BLDA and SVM, which demonstrated lower performance. Whilst the KNN method closely approached our proposed method, achieving AUC = 86% and Recall = 86%, RF demonstrated superior performance, surpassing both metrics with remarkable values of AUC and Recall, reaching around 94%. This notable improvement highlights the efficacy of the RF method in capturing complex patterns and enhancing the overall predictive capabilities.

Additionally, Figures 3 and 4 illustrate a well-balanced performance graph for our proposed system, indicating minimal disparities between training and testing phases and no signs of overfitting. This establishes the system as a dependable tool, facilitating automated analysis to aid in the classification of SLE patients. Our results affirm the efficacy of the RF system in precisely predicting SLE patients, establishing it as a valuable tool for supporting SLE diagnosis.

5. Conclusions

In conclusion, due to the complexity of this elusive autoimmune disease, the use of ML algorithms such as RF is critical for the classification and rapid detection of patients with SLE flares. SLE presents with a range of challenging symptoms that are particularly difficult to diagnose accurately in its early stages. The intricate relationship between infections and autoimmunity in SLE underscores the critical need for preventative measures and the early detection of infections in SLE patients exhibiting heightened susceptibility. This integrated approach aims to address the multifaceted challenges of SLE, providing a more holistic understanding for improved patient care.

RF's proficiency in handling diverse datasets and extracting intricate patterns makes it well-suited for identifying subtle indicators of SLE. The algorithm's swift information processing enables quick detection, allowing for timely intervention and personalized treatment plans for SLE patients. Given the rarity and importance of SLE, the use of RF and similar ML approaches not only improves the diagnostic accuracy of SLE activity, but

also contributes to improved patient outcomes, long-term monitoring, and a more effective healthcare management strategy for this devastating disease.

Thus, this investigation delves into the optimal ML technique for identifying patterns of immunodeficiency within the SLE population. It establishes that an ML system serves as a highly accurate tool for identifying diminished levels of immune parameters in individuals at a significantly elevated risk of experiencing both infections and, consequently, SLE flares. Moreover, the RF-based system proposed surpasses the performance of other studies, evident in a larger AUC, thereby affirming its superior predictive accuracy.

Author Contributions: Conceptualization: I.U., Y.A. and J.M.; methodology: I.U., Y.A., A.M.T. and J.M.; formal analysis: I.U., Y.A., A.M.T. and J.M.; investigation: I.U., Y.A., A.M.T., J.B. and J.M.; writing—original draft preparation: I.U., Y.A., A.M.T., J.B. and J.M.; writing—review and editing: I.U., Y.A., A.M.T., J.B. and J.M.; supervision: A.M.T. and J.M.; project administration: J.M.; funding acquisition: Y.A. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by UCLM-Telefónica Chair and Ministry of Economic Affairs and Digital Transformation (MINECO) grant number PID2021-125122OB-I00.

Institutional Review Board Statement: This research was approved by the ethics committee of the Valladolid Clinic Hospital.

Informed Consent Statement: Informed consent was obtained from all subjects involved in the study.

Data Availability Statement: The datasets used and/or analyzed during the present study are available from the corresponding author on reasonable request.

Acknowledgments: This work was sponsored by Institute of Technology (University of Castilla-La Mancha), the Valladolid Clinic Hospital (Spain), and the UCLM-Telefónica Chair (Spain).

Conflicts of Interest: The authors declare that they have no conflicts of interest.

References

1. Cortés Verdú, R.; Pego-Reigosa, J.M.; Seoane-Mato, D.; Morcillo Valle, M.; Palma Sánchez, D.; Moreno Martínez, M.J.; Mayor González, M.; Atxotegi Sáenz de Buruaga, J.; Urionaguena Onaindia, I.; Blanco Cáceres, B.A.; et al. Prevalence of systemic lupus erythematosus in Spain: Higher than previously reported in other countries? *Rheumatology* **2020**, *59*, 2556–2562. [PubMed] [PubMed]
2. Schmidt, R.E.; Grimbacher, B.; Witte, T. Autoimmunity and primary immunodeficiency: Two sides of the same coin? *Nat. Rev. Rheumatol.* **2018**, *14*, 7–18. [CrossRef] [CrossRef] [PubMed]
3. Bandinelli, F.; Bombardieri, S.; Matucci, M.; Delle Sedie, A. Systemic lupus erythematosus joint involvement—What does musculoskeletal ultrasound provide Us? *Eur. Musculoskelet. Rev.* **2012**, *7*, 221–223.
4. Kariburyo, F.; Xie, L.; Sah, J.; Li, N.; Lofland, J.H. Real-world medication use and economic outcomes in incident systemic lupus erythematosus patients in the United States. *J. Med. Econ.* **2020**, *23*, 1–9. [CrossRef] [PubMed] [CrossRef]
5. Piga, M.; Arnaud, L. The main challenges in systemic lupus erythematosus: Where do we stand? *J. Clin. Med.* **2021**, *10*, 243. [CrossRef] [CrossRef]
6. Rees, F.; Doherty, M.; Grainge, M.; Davenport, G.; Lanyon, P.; Zhang, W. The incidence and prevalence of systemic lupus erythematosus in the UK, 1999–2012. *Ann. Rheum. Dis.* **2016**, *75*, 136–141. [CrossRef] [PubMed] [CrossRef]
7. Adamichou, C.; Bertsias, G. Flares in systemic lupus erythematosus: Diagnosis, risk factors and preventive strategies. *Mediterr. J. Rheumatol.* **2017**, *28*, 4–12.
8. Zhou, Y.; Wang, M.; Zhao, S.; Yan, Y. Machine Learning for Diagnosis of Systemic Lupus Erythematosus: A Systematic Review and Meta-Analysis. *Comput. Intell. Neurosci.* **2022**, *2022*, 7167066. [CrossRef] [CrossRef]
9. Handelman, G.; Kok, H.; Chandra, R.; Razavi, A.; Lee, M.; Asadi, H. eDoctor: Machine learning and the future of medicine. *J. Intern. Med.* **2018**, *284*, 603–619. [CrossRef] [CrossRef]
10. Adamichou, C.; Nikolopoulos, D.; Genitsaridi, I.; Bortoluzzi, A.; Fanouriakis, A.; Papastefanakis, E.; Kalogiannaki, E.; Gergianaki, I.; Sidiropoulos, P.; Boumpas, D.T.; et al. In an early SLE cohort the ACR-1997, SLICC-2012 and EULAR/ACR-2019 criteria classify non-overlapping groups of patients: Use of all three criteria ensures optimal capture for clinical studies while their modification earlier classification and treatment. *Ann. Rheum. Dis.* **2020**, *79*, 232–241. [CrossRef] [CrossRef]
11. Suárez, M.; Martínez, R.; Torres, A.M.; Ramón, A.; Blasco, P.; Mateo, J. Personalized Risk Assessment of Hepatic Fibrosis after Cholecystectomy in Metabolic-Associated Steatotic Liver Disease: A Machine Learning Approach. *J. Clin. Med.* **2023**, *12*, 6489. [CrossRef] [PubMed] [PubMed]
12. Casillas, N.; Ramón, A.; Torres, A.M.; Blasco, P.; Mateo, J. Predictive Model for Mortality in Severe COVID-19 Patients across the Six Pandemic Waves. *Viruses* **2023**, *15*, 2184. [CrossRef] [PubMed] [CrossRef] [PubMed]

13. Soria, C.; Arroyo, Y.; Torres, A.M.; Redondo, M.Á.; Basar, C.; Mateo, J. Method for Classifying Schizophrenia Patients Based on Machine Learning. *J. Clin. Med.* **2023**, *12*, 4375. [CrossRef] [PubMed] [CrossRef] [PubMed]
14. Chen, Y.; Mao, Q.; Wang, B.; Duan, P.; Zhang, B.; Hong, Z. Privacy-Preserving Multi-Class Support Vector Machine Model on Medical Diagnosis. *IEEE J. Biomed. Health Inform.* **2022**, *26*, 3342–3353. [CrossRef] [PubMed] [CrossRef] [PubMed]
15. Sethi, M.; Ahuja, S.; Rani, S.; Bawa, P.; Zaguia, A. Classification of Alzheimer's disease using Gaussian-based Bayesian parameter optimization for deep convolutional LSTM network. *Comput. Math. Methods Med.* **2021**, *2021*, 4186666. [PubMed] [CrossRef]
16. Mahfouz, M.A.; Shoukry, A.; Ismail, M.A. EKNN: Ensemble classifier incorporating connectivity and density into kNN with application to cancer diagnosis. *Artif. Intell. Med.* **2021**, *111*, 101985. [PubMed] [CrossRef]
17. Reges, O.; Krefman, A.E.; Hardy, S.T.; Yano, Y.; Muntner, P.; Lloyd-Jones, D.M.; Allen, N.B. Decision tree-based classification for maintaining normal blood pressure throughout early adulthood and middle age: Findings from the coronary artery risk development in young adults (CARDIA) study. *Am. J. Hypertens.* **2021**, *34*, 1037–1041. [CrossRef] [CrossRef]
18. Zhao, X.; Zhang, L.; Wang, J.; Zhang, M.; Song, Z.; Ni, B.; You, Y. Identification of key biomarkers and immune infiltration in systemic lupus erythematosus by integrated bioinformatics analysis. *J. Transl. Med.* **2021**, *19*, 35. [CrossRef] [CrossRef]
19. Jiang, Z.; Shao, M.; Dai, X.; Pan, Z.; Liu, D. Identification of diagnostic biomarkers in systemic lupus erythematosus based on bioinformatics analysis and machine learning. *Front. Genet.* **2022**, *13*, 865559.
20. Jorge, A.M.; Smith, D.; Wu, Z.; Chowdhury, T.; Costenbader, K.; Zhang, Y.; Choi, H.K.; Feldman, C.H.; Zhao, Y. Exploration of machine learning methods to predict systemic lupus erythematosus hospitalizations. *Lupus* **2022**, *31*, 1296–1305.
21. Cheng, Q.; Chen, X.; Wu, H.; Du, Y. Three hematologic/immune system-specific expressed genes are considered as the potential biomarkers for the diagnosis of early rheumatoid arthritis through bioinformatics analysis. *J. Transl. Med.* **2021**, *19*, 18. [CrossRef] [PubMed] [PubMed]
22. Cicalese, P.A.; Mobiny, A.; Shahmoradi, Z.; Yi, X.; Mohan, C.; Van Nguyen, H. Kidney level lupus nephritis classification using uncertainty guided Bayesian convolutional neural networks. *IEEE J. Biomed. Health Inform.* **2020**, *25*, 315–324.
23. Aringer, M.; Brinks, R.; Dörner, T.; Daikh, D.; Mosca, M.; Ramsey-Goldman, R.; Smolen, J.S.; Wofsy, D.; Boumpas, D.T.; Kamen, D.L.; et al. European League against Rheumatism (EULAR)/American College of Rheumatology (ACR) SLE classification criteria item performance. *Ann. Rheum. Dis.* **2021**, *80*, 775–781. [CrossRef] [CrossRef] [PubMed]
24. Breiman, L. Random Forests. *Machine Learning* **2001**, *45*, 5–32. [10.1093/3404324CrossRef] [CrossRef]
25. Han, S.; Williamson, B.D.; Fong, Y. Improving random forest predictions in small datasets from two-phase sampling designs. *BMC Med. Inform. Decis. Mak.* **2021**, *21*, 322.
26. Huang, A.; Zhou, W. BLDA Approach for Classifying P300 Potential. In Proceedings of the 7th Asian-Pacific Conference on Medical and Biological Engineering, Beijing, China, 22–25 April 2008; Springer: Berlin/Heidelberg, Germany, 2008; pp. 341–343.
27. Huang, M. Theory and Implementation of linear regression. In Proceedings of the 2020 International Conference on Computer Vision, Image and Deep Learning (CVIDL), Chongqing, China, 10–12 July 2020; pp. 210–217.
28. Kuchibhotla, A.K.; Brown, L.D.; Buja, A.; Cai, J. All of Linear Regression. *arXiv* **2019**, arXiv:1910.06386.
29. Han, J.; Pei, J.; Kamber, M. *Data Mining: Concepts and Techniques*, 3rd ed.; Elsevier: Amsterdam, The Netherlands, 2016.
30. Bruce, I.N.; O'Keeffe, A.G.; Farewell, V.; Hanly, J.G.; Manzi, S.; Su, L.; Gladman, D.D.; Bae, S.C.; Sanchez-Guerrero, J.; Romero-Diaz, J.; et al. Factors associated with damage accrual in patients with systemic lupus erythematosus: Results from the Systemic Lupus International Collaborating Clinics (SLICC) Inception Cohort. *Ann. Rheum. Dis.* **2015**, *74*, 1706–1713.
31. Liossis, S.N.; Staveri, C. What is New in the Treatment of Systemic Lupus Erythematosus. *Front. Med.* **2021**, *8*, 655100. [CrossRef]
32. Ugarte-Gil, M.F.; Mendoza-Pinto, C.; Reátegui-Sokolova, C.; Pons-Estel, G.J.; Van Vollenhoven, R.F.; Bertsias, G.; Alarcon, G.S.; Pons-Estel, B.A. Achieving remission or low disease activity is associated with better outcomes in patients with systemic lupus erythematosus: A systematic literature review. *Lupus Sci. Med.* **2021**, *8*, e000542. [CrossRef] [CrossRef]
33. Yavuz, S.; Lipsky, P.E. Current Status of the Evaluation and Management of Lupus Patients and Future Prospects. *Front. Med.* **2021**, *8*, 682544.
34. Carter, E.E.; Barr, S.G.; Clarke, A.E. The global burden of SLE: Prevalence, health disparities and socioeconomic impact. *Nat. Rev. Rheumatol.* **2016**, *12*, 605–620. [CrossRef] [PubMed]
35. Aparicio-Soto, M.; Sánchez-Hidalgo, M.; Alarcón-de-la Lastra, C. An update on diet and nutritional factors in systemic lupus erythematosus management. *Nutr. Res. Rev.* **2017**, *30*, 118–137. [CrossRef] [PubMed] [PubMed]
36. Tselios, K.; Gladman, D.; Touma, Z.; Su, J.; Anderson, N.; Urowitz, M. Disease course patterns in systemic lupus erythematosus. *Lupus* **2019**, *28*, 114–122. [CrossRef] [PubMed] [PubMed]
37. Akhil, A.; Bansal, R.; Anupam, K.; Ankit, T.; Bhatnagar, A. Systemic lupus erythematosus: Latest insight into etiopathogenesis. *Rheumatol. Int.* **2023**, *43*, 1381–1393. [CrossRef] [CrossRef] [PubMed]
38. Larosa, M.; Iaccarino, L.; Gatto, M.; Punzi, L.; Doria, A. Advances in the diagnosis and classification of systemic lupus erythematosus. *Expert Rev. Clin. Immunol.* **2016**, *12*, 1309–1320. [CrossRef] [CrossRef]
39. Aringer, M.; Johnson, S.R. Classifying and diagnosing systemic lupus erythematosus in the 21st century. *Rheumatology* **2020**, *59*, v4–v11. [PubMed] [CrossRef]
40. Inês, L.; Silva, C.; Galindo, M.; López-Longo, F.J.; Terroso, G.; Romão, V.C.; Rúa-Figueroa, I.; Santos, M.J.; Pego-Reigosa, J.M.; Nero, P.; et al. Classification of systemic lupus erythematosus: Systemic Lupus International Collaborating Clinics versus American College of Rheumatology criteria. A comparative study of 2055 patients from a real-life, international systemic lupus erythematosus cohort. *Arthritis Care Res.* **2015**, *67*, 1180–1185. [CrossRef] [CrossRef]

41. Adamichou, C.; Genitsaridi, I.; Nikolopoulos, D.; Nikoloudaki, M.; Repa, A.; Bortoluzzi, A.; Fanouriakis, A.; Sidiropoulos, P.; Boumpas, D.T.; Bertsias, G.K. Lupus or not? SLE Risk Probability Index (SLERPI): A simple, clinician-friendly machine learning-based model to assist the diagnosis of systemic lupus erythematosus. *Ann. Rheum. Dis.* **2021**, *80*, 758–766. [CrossRef]
42. Donner-Banzhoff, N. Solving the diagnostic challenge: A patient-centered approach. *Ann. Fam. Med.* **2018**, *16*, 353–358. [CrossRef] [CrossRef]
43. Kinloch, A.J.; Asano, Y.; Mohsin, A.; Henry, C.; Abraham, R.; Chang, A.; Labno, C.; Wilson, P.C.; Clark, M.R. Machine learning to quantify in situ humoral selection in human lupus tubulointerstitial inflammation. *Front. Immunol.* **2020**, *11*, 593177. [CrossRef] [CrossRef]
44. Usategui, I.; Barbado, J.; Torres, A.M.; Cascón, J.; Mateo, J. Machine learning, a new tool for the detection of immunodeficiency patterns in systemic lupus erythematosus. *J. Investig. Med.* **2023**, *71*, 742–752. [CrossRef] [PubMed] [CrossRef] [PubMed]
45. Chen, W.; Lei, X.; Chakraborty, R.; Pal, S.C.; Sahana, M.; Janizadeh, S. Evaluation of different boosting ensemble machine learning models and novel deep learning and boosting framework for head-cut gully erosion susceptibility. *J. Environ. Manag.* **2021**, *284*, 112015. [CrossRef] [PubMed]

Disclaimer/Publisher’s Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.

Review

Searching for the Best Machine Learning Algorithm for the Detection of Left Ventricular Hypertrophy from the ECG: A Review

Simon W Rabkin

Department of Medicine, Division of Cardiology, University of British Columbia, 9th Floor 2775 Laurel St., Vancouver, BC V5Z 1M9, Canada; simon.rabkin@ubc.ca; Tel.: +1-(604)-875-5847; Fax: +1-(604)-875-5849

Abstract: Background: Left ventricular hypertrophy (LVH) is a powerful predictor of future cardiovascular events. **Objectives:** The objectives of this study were to conduct a systematic review of machine learning (ML) algorithms for the identification of LVH and compare them with respect to the classical features of test sensitivity, specificity, accuracy, ROC and the traditional ECG criteria for LVH. **Methods:** A search string was constructed with the operators “left ventricular hypertrophy, electrocardiogram” AND machine learning; then, Medline and PubMed were systematically searched. **Results:** There were 14 studies that examined the detection of LVH utilizing the ECG and utilized at least one ML approach. ML approaches encompassed support vector machines, logistic regression, Random Forest, GLMNet, Gradient Boosting Machine, XGBoost, AdaBoost, ensemble neural networks, convolutional neural networks, deep neural networks and a back-propagation neural network. Sensitivity ranged from 0.29 to 0.966 and specificity ranged from 0.53 to 0.99. A comparison with the classical ECG criteria for LVH was performed in nine studies. ML algorithms were universally more sensitive than the Cornell voltage, Cornell product, Sokolow-Lyons or Romhilt-Estes criteria. However, none of the ML algorithms had meaningfully better specificity, and four were worse. Many of the ML algorithms included a large number of clinical (age, sex, height, weight), laboratory and detailed ECG waveform data (P, QRS and T wave), making them difficult to utilize in a clinical screening situation. **Conclusions:** There are over a dozen different ML algorithms for the detection of LVH on a 12-lead ECG that use various ECG signal analyses and/or the inclusion of clinical and laboratory variables. Most improved in terms of sensitivity, but most also failed to outperform specificity compared to the classic ECG criteria. ML algorithms should be compared or tested on the same (standard) database.

Keywords: left ventricular hypertrophy; electrocardiogram; machine learning; artificial intelligence

1. Introduction

A number of different groups have proposed machine learning models to evaluate ECG with or without additional clinical and laboratory data to construct an approach to identify left ventricular hypertrophy (LVH). LVH, or an increased left ventricular mass, is a powerful predictor of future cardiovascular events [1–3]. LVH can serve as a marker for the severity of (occult) cardiovascular disease, thereby identifying an increased risk of stroke or, more directly, by limiting myocardial perfusion, leading to myocardial ischemia and serious cardiac arrhythmias [4–8]. The ECG has been used for decades as an indicator of the presence of LVH, with increased QRS voltage being considered to be a marker for increased left ventricular mass [9–16]. Although the ECG QRS voltage criteria are not a highly sensitive indicator of LVH [17–25], the importance of predicting the presence of LVH and the imperative of cost efficiency, i.e., utilization of a low-cost ECG compared to a more expensive echocardiogram or MRI, has focused attention on how to extract more precise indicators of LVH from ECGs. This imperative is underscored by the use of a 12-lead ECG as part of the basic assessment of patients with cardiovascular disease [26,27].

Because of the importance of LVH identification, the ML disparate approaches and the variables considered in each approach, a review of this field has become increasingly needed.

A meta-analysis, to construct a single estimate or effect size, is not realistic in fields such as machine learning, when different input variables and analytic techniques are employed by utilizing different algorithms on different datasets. Hence, an in-depth review is the best approach to evaluate different ML algorithms. The objectives of this study were to conduct a review of machine learning algorithms for the identification of LVH with respect to the classical features of sensitivity, specificity and accuracy. It also aimed to assess how each ML algorithm compares to the traditional ECG criteria for LVH, specifically Cornell voltage [13], Cornell product [28], Sokolow-Lyons [9] and Romhilt-Estes [29] criteria.

2. Methods

2.1. Literature Search

A search string was constructed using terms connected with Boolean operators “left ventricular hypertrophy AND electrocardiogram or ECG” AND machine learning to identify articles reporting a machine learning approach for the diagnosis of LVH. Medline and PubMed were systematically searched from their date of inception through to 31 October 2023. Preferred Reporting Items for Reviews and Meta-Analysis (PRISMA) was used to conduct the search [30] (Supplementary Figure S1).

Article titles and abstracts were assessed for full-text review. Papers on hypertrophic obstructive cardiomyopathy were excluded because this entity represents an asymmetric cardiac hypertrophy, which would alter ECG voltage in a different manner. The exclusion criteria were as follows: non-English studies, non-primary studies, studies without full texts, studies that have insufficient data for analysis, non-human studies and studies unrelated to the investigated topic. The review was not registered, and protocol is not available for access.

2.2. Data Extraction and Classification

Data extraction was performed by one reviewer. The following items were collected from each article: author, year of publication, recruitment center or clinical trial sampled, sample size, age and sex of participants, definition of LVH and its method of assessment. Reported sensitivity, specificity, positive predictive value, negative predictive value, area under the receiver operating curve (ROC), overall accuracy and F1 score were extracted. Input variables and the ML techniques utilized were also extracted.

3. Results

There were 14 studies that examined the detection of left ventricular hypertrophy with an approach utilizing the ECG and utilized at least one machine language approach (Table 1). Some details of the study population characteristics and input variables are summarized in Table 1.

Lin and Liu evaluated data from 2196 men, aged 17 to 45 years of age, who were in the military, and used the support vector machine (SVM) classifier as the machine learning method [31]. The prevalence of echocardiographic LVH was about 6.5%. Thirty-one input variables were utilized that included three clinical ones, age, body height and body weight, and 28 ECG parameters, such as heart rate, the durations of P wave, PR interval, QRS interval, QT interval and QTc interval in Lead II and the axes of the P, QRS and T waves in Lead II, and the voltages of R waves in all Limb Leads I, II, III, aVR, aVL, aVF and S wave in Lead aVL, and the voltages of R and S waves in all precordial leads V1–V6 [31]. The model had high sensitivity (86.7%).

Table 1. Summary of the studies, input variables and machine learning approaches.

Authors	Population	Country	Sample Size	Sex (%M)	Age (yrs)	Method LVH	Definition LVH	LVH	Variables	Machine Learning
Lin & Lui 2020 [31]	Military	Tawain	2196	100	26	Echocardiogram	$\geq 116 \text{ g/m}^2$	6.5%	31 parameters 3 clinical -age, body height, body weight 28 ECG parameters: duration P, PR, QRS, QT, QTc, P axis QRS axis, T axis plus R amplitude in all 12 leads, S amplitude in aVL, V1-6	Support vector machine classifier (SVM)
Sparapani et al., 2019 [32]	Multi-ethnic	USA	4714	46		MRI	95th percentile	NA	556 ECG variables: PR interval, P axis, QRS interval, QRS axis plus 552 amplitudes and durations per ECG	Bayesian additive regression tree
De la Garza-Salazar et al., 2020 [33]	Hospital	Mexico	432	56	67	Echocardiogram	$>115 \text{ g/m}^2$ (men)	48%	ECG p wave, QRS complex and ST waves	C5.0 supervised ML algorithm to create a multilevel binary decision tree,
							$>95 \text{ g/m}^2$ (women).			
Kwon et al., 2020 [34]	Hospital based	Korea	21,286	49	59	Echocardiogram	$>132 \text{ g/m}^2$ in men	21%	age, sex, weight, height and ECG features, heart rate, presence of atrial fibrillation or flutter, QT, QRS duration, R-wave axis, T-wave	ENN, LR and RF
De la Garza-Salazar et al., 2021 [35]	Hospital	Mexico	439	NA	67	Echocardiogram	Presumed same as 2020	46%	'Raw' ECG data with 5000 numbers from each of the 12 leads. ECG variables including T wave voltage in the lead I, peak-to-peak QRS distance (QRS PPK) in aVF, and peak-to-peak QRS distance in aVL	C5.0 supervised ML algorithm to create a multilevel binary decision tree,
Khurshid et al., 2021 [36]	UK data base	UK	32,239	47	64	MRI		2.6%		
Sabovčik et al., 2021 [37]	General population	Belgium	1407	49	51	Echocardiogram	$>115 \text{ g/m}^2$ (men)	19%	67 variables including clinical, ECG onsets, amplitudes and intervals of P waves, QRS-complexes, and T wave as well as	LR, XGBoost, Random Forest, AdaBoost, Support Vector Machines
							or 95 g/m^2 (women).		blood count, blood glucose, lipid profile, hormones (plasma renin, leptin, insulin, aldosterone, and cortisol), minerals,	
Angelaki et al. 2021 [38]	NA	Greece	528	44	61	Echocardiogram	$>115 \text{ g/m}^2$ (men)	16.8%	clinical variables (sex, age, BMI class, BSA, hypertension, and height	

Table 1. Cont.

Authors	Population	Country	Sample Size	Sex (%M)	Age (yrs)	Method LVH	Definition LVH	LVH	Variables	Machine Learning
Lim et al., 2021 [39]	Military	Singapore	17,310	100	18	Echocardiogram	>95 g/m ² (women)	26 chosen ECG-derived features	clinical variables were: body weight, height, body fat percentage, and systolic blood pressure	Random Forest Logistic Regression, GLMNet, Random Forests, Gradient Boosting Machines
							>115 g/m ² (men)	0.8%		
							ECG variables included: QT interval, mean QRS duration and R wave in lead I			
Zhao et al., 2022 [40]	Hospital based	China	3120	42	65	Echocardiogram	>115 g/m ² (men)	56%	uncertain	CNN
Sammani et al., 2022 [41]	Hospital based	The Netherlands	2456	55	61	Echocardiogram	>95 g/m ² (women).	Lab: Hgb, PLT, lipids, creatinine, Na, K	age, systolic blood pressure and body surface area	XGBoost
							>115 g/m ² (men)	0.8%		
							20 ECG data: p, QRS and T wave axes, pr, QRS, QT and QTc durations, peak amplitudes of p, Q, R, S and T waves			
Kokubo et al., 2022 [42]	Hospital based	Japan	12,008	64	57	Echocardiogram	>101 g/m ² for men	19 factors—clinical (age, sex, height and weight) and ECG features (heart rate, rhythm, pr interval, QT interval, QRS axis, p wave axis)	as well as QRS voltages in leads V1, V2, V5 and V6	ENN
							16.5%			
							Clinical—blood pressure, diabetes mellitus, lipids, cigarette and alcohol consumption			
Naderi et al., 2023 [43]	UK data base	UK	37,534	48	64	MRI	>70 g/m ² (men)	1.5%	23 ECG variables from leads I, II, V1-6	LR, RF
							>55 g/m ² (women)			
Liu et al., 2023 [44]	Military Hospital	Tawain	952	90		Echocardiogram	>115 g/m ² (men)	18%	24 features which consisted of R peak and S valley amplitudes automatically obtained from the output of ECG signal	Decision tree SVM and Back propagated Neural Network

Sparapani et al. evaluated 3774 participants from MESA (Multi-Ethnic Study of Atherosclerosis), free of clinically apparent cardiovascular disease at enrollment, using ECG and participant characteristics to predict LV mass from cardiac magnetic resonance imaging [32]. There were four global ECG measurements (PR interval, P axis, QRS interval and QRS axis) plus 552 amplitude and duration measurements per ECG, which resulted in 556 ECG variables. The machine learning technique Bayesian Additive Regression Trees (BART) was used [32]. This model showed the highest sensitivity (29.0%), greater than the other criteria, including the Sokolow-Lyon criterion (21.7%), Peguero-Lo Presti (14.5%), Cornell voltage product (10.1%) and Cornell voltage (5.8%). The specificity was >93% for all criteria [32].

Garza-Salazar et al. conducted an observational, retrospective case-control study that included data from a representative sample of consecutive adult patients who underwent an ECG and an echocardiogram at their institution [33]. They evaluated 432 patients, of whom 47% had LVH [33]. The ECG variables included S-wave voltage and R-wave voltage in all ECG leads (I, II, III, aVL, aVF, aVR and V1-V6), P-wave duration and voltage in the V1 lead, left atrial enlargement, QRS complex duration in lead V1, QRS axis (using leads I and aVL), intrinsicoid deflection in lead V6 and “ST strain” (downward ST depression and asymmetric T-wave inversion) [33]. The logistic regression (LR) model was used as well as a supervised ML algorithm to create a multilevel binary decision tree, using the ECG features that provided the greatest information to classify patients as having LVH [33]. Their five-level binary decision tree used only six predictive variables and had an accuracy of 71.4%, a sensitivity of 79.6% and specificity of 53% [33].

De la Garza Salazar et al. reported another observational, retrospective, case-control study on 439 patients who underwent an echocardiogram and an ECG [35]. Sixteen ECG parameters, including T voltage in lead I, peak-to-peak QRS distance in aVL (>1.235 mV) and peak-to-peak QRS distance in aVF (>0.178 mV), were fed into a C5.0 ML algorithm, a method that defines a decision tree structure model (or criteria). Their model had an accuracy of 70.5%, a sensitivity of 74.3% and a specificity of 68.7%.

Kwon et al. conducted a retrospective cohort study of 12,648 patients who underwent 12-lead ECG and echocardiography [34]. LVH was present in 21% of the group. An ensemble neural network (ENN) combining a convolutional (CNN) and deep neural network (DNN) was developed. Two other machine learning-based algorithms—LR and RF—were also developed. The model was developed using 3162 ECGs from 3162 patients. They used four clinical variables (age, sex, weight and height) and ECG features, such as heart rate, presence of atrial fibrillation or flutter, QT interval, QTc, QRS duration, R-wave axis and T-wave axis. In addition, they used raw ECG data with 5000 numbers from each of the 12 leads. The area under the ROC curve for ENN was 0.880, which significantly outperformed the Romhilt-Estes point system, Cornell voltage criteria and the Sokolow-Lyon criteria [34].

Lim et al. examined the ECGs and echocardiograms of 17,310 male military conscripts, aged 16 to 23 years [39]. The prevalence of echocardiographic LVH was 0.82%. Several machine learning models (Logistic Regression, GLMNet, Random Forests and Gradient Boosting Machines) were used. Their clinical variables were body weight, height, body fat percentage and systolic blood pressure. Their ECG variables included QT interval, mean QRS duration and R wave in lead I, ECG parameters not used in the classical criteria but deemed important to the machine learning algorithms, both when ECG parameters alone were included and when all predictive parameters were included. Considering AUC, ML methods achieved superior performance: logistic regression (0.811), GLMNet (0.873), Random Forest (0.824) and Gradient Boosting Machines (0.800).

Two studies used the UK Biobank with individuals aged 40 to 69 years, with a mean age of 64 years, of which 52% were female, who had LV mass index assessed by MRI [36]. Khurshid et al. also tested a Massachusetts General/Brigham Hospital database, but more information was available for the UK Biobank so it was selected in this analysis. Khurshid et al. trained an ML model on 32,239 participants [36]. The input variables were

demographic factor, age, sex, race, height, weight and body mass index (BMI), plus ECG waveform data. Their model had a sensitivity of 34% and a specificity of 96% [36].

Naderi et al. also explored the UK Biobank [43]. There was a low prevalence of LVH, specifically 1.5%. Demographic factors included age, sex and race, and physical measurements included height, weight and body mass index (BMI). Clinical variables included blood pressure, and 23 ECG variables used the independent ECG leads (I, II, V1–6). ECG variables consisted of ECG waveform data. Three supervised machine learning algorithms, logistic regression (LR), support vector machine (SVM) and Random Forest (RF), were used. For the SVM classifier, the Gaussian kernel function was applied to deal with potential non-linear data. The three models were comparable in classifying LVH. Classification of LVH with logistic regression had an accuracy of 81%, sensitivity of 70%, specificity of 81% and an AUC of 0.86. Analysis with SVM showed 81% accuracy, sensitivity of 72%, specificity of 81% and AUC of 0.85. RF analysis showed 72% accuracy, sensitivity of 74%, specificity of 72% and AUC of 0.83 [43].

Sabovčik et al. evaluated 1407 individuals (mean age 51 years, 51% women), randomly recruited from the general population, of whom an echocardiographically determined LV mass was present in 19% [37]. A large number of clinical and laboratory variables (blood count, blood glucose, lipids, renin activity, leptin, insulin, aldosterone and cortisol) were used. From the ECG tracing, the onsets, amplitudes and intervals of P waves, QRS complexes and T waves were extracted. They used five standard ML methods, XGBoost, AdaBoost, RF, SVM and LR, to build classifiers based on 67 clinical, biochemical and ECG variables. A high area under the ROC was found for XGBoost (0.785), RF classifiers (0.783), AdaBoost (0.771), SVM (0.783) and LR (0.783) for predicting LVH. Age, body mass index, different components of blood pressure, history of hypertension, antihypertensive treatment and various electrocardiographic variables were the top features for predicting LVH [37].

Angelaki et al. evaluated 528 patients with and without essential hypertension but no other indications of cardiovascular disease [38]. LVH, assessed by echocardiogram, was present in 16.8% of cases. Clinical variables were used. ECG waveform measurements from each lead included peak voltages, area of the QRS complex, planar frontal QRS-T angle and QTc duration. A Random Forest ML algorithm consisting of a collection of de-correlated decision trees was used. They calculated SHAP (SHapley Additive exPlanations). Hypertension, age and BMI were the most significant factors predicting the presence of LVH. The area under the QRS complex summed over all 12 leads, the Planar Frontal QRS-T angle and QTc duration, among others, was important in predicting risk. For the identification of LVH, their model noted 87% accuracy, 75% specificity, 97% sensitivity and area under the receiver operating curve (AUC/ROC) of 0.91 [38]. However, some of the patients did not have LVH but rather concentric remodeling [38].

Kokubo et al. analyzed data from patients aged 18 years or older, with a mean age 64.2 years, 57% men, who had an echocardiogram and ECG at The University of Tokyo Hospital [42]. LVH was defined as an LVMI > 101 g/m² for men and > 85 g/m² for women, consistent with recommendations for the Japanese population, and was present in 16.5% of cases [42]. The data were derived from a training set of 12,008 persons. Nineteen factors—clinical (age, sex, height and weight) and ECG features (heart rate, rhythm, pr interval, QT interval, QRS axis, P wave axis as well as QRS voltages in leads V1, V2, V5 and V6)—were used as input variables. They developed an ensemble neural network (ENN) model, which consisted of a convolutional neural network (CNN) and a deep neural network (DNN) as well as a LR and RF approaches to detect LVH. For the detection of LVH, the area under the ROC curve was 0.784 for the deep learning model, which was significantly greater than that of the LR, RF or conventional ECG criteria [42].

Zhao et al. utilized data from 3120 patients who had an echocardiogram and an ECG within one week after hospital admission [40]. The input variables included clinical factors, such as age, sex and medical history; laboratory factors, such as hemoglobin, platelet count, lipids, creatinine, Na, K⁺; ECG factors, such as R in AVL, V5 and V6, and S in V1 and

V3. The ECG final dataset included 36,350 ECG segments in the control and LVH groups. They constructed and built a deep learning (DL) model based on convolutional neural network–long short-term memory (CNN-LSTM) to detect LVH. LVH was predicted by the CNN-LSTM model with an area under the curve (AUC) of 0.62, with a sensitivity of 68% and specificity of 57%. The CNN-LSTM model predicted LVH by 12-lead ECG performed better in male than female patients [40].

Sammani et al. developed an ML algorithm for echocardiographically detected LVH that utilized a variety of clinical factors (age, systolic blood pressure and body surface area) and over 20 ECG data variables (P, QRS and T wave axes, PR, QRS, QT and QTc durations, peak amplitudes of P, Q, R, S and T waves in three different ECG leads) [41]. There were 26,954 subjects (median age 61 years, 55% male), of whom 0.8% had LVH, and of those with LVH, a very small number had amyloidosis; only two had Anderson-Fabry Disease. XGBoost was the only machine learning logarithm used [41].

Liu et al. studied 952 individuals, mainly men, from a military hospital and used a back-propagation neural network (BPN) on 24 features, which consisted of R peak and S valley amplitudes, automatically obtained from the output of the ECG signal. This group found a prevalence of 18% with echocardiographic LVH. Their combination of sensitivity and specificity was the highest of any approach [44].

Sensitivity and specificity were reported in 13 of the 14 studies. There was a wide range of sensitivity for the ML approaches across all studies. The range is from 0.29 to 0.966 (Figure 1). The highest sensitivity was 0.966 using the algorithm proposed by Liu et al., 2023 [44], followed by 0.867 using the algorithm proposed by Lin and Liu [31], followed by the one proposed by De la Garza-Salazar et al. [33]. Specificity ranged from 0.53 to 0.99 (Figure 2). The highest specificity was found using the algorithm proposed by Sammani et al. [41], followed closely by that of Liu et al. [44] and then Khurshid et al. [36].

An overall assessment indicated by AUC was reported in nine studies and ranged from 0.705 to 0.89, with the highest AUC reported for the algorithms of Angelaki et al. [38] followed by Kwon et al. [34] (Figure 3). Overall accuracy was detailed in eight studies, with the highest value of 0.961 from Liu et al. [44] followed by that of Angelaki et al. [38] (Figure 4). The next best was that of Kwon et al. [34]. Positive and negative predictive values were reported in less than one half, or only six studies (Figure 4). Four studies presented their F1 score, which is a composite indicator of sensitivity, the true-positive rate, taking into account false positives and false negatives. The values ranged from 0.294 [32] to 0.3314 [31] and 0.458–0.488 (depending on the ML method) [37] to 0.64, which was the highest value and was reported by Zhao et al., 2022 [40].

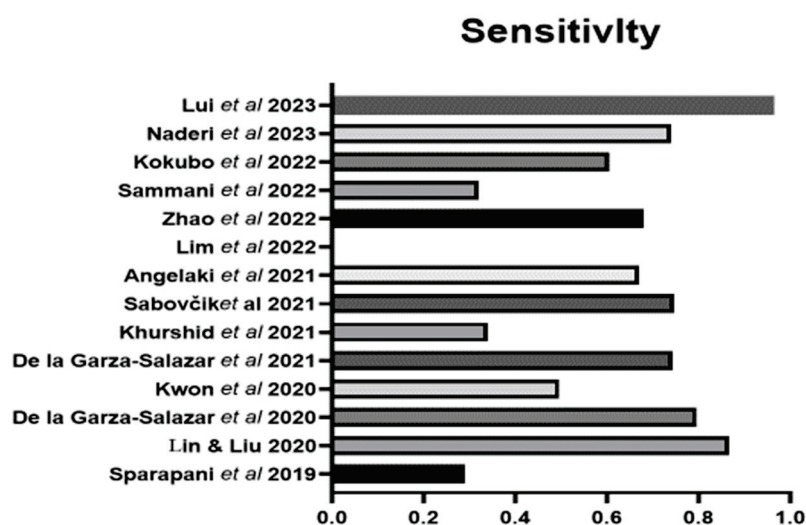


Figure 1. The sensitivity of the ML algorithms for LVH. Lin & Liu 2020 [31], Sparapani et al., 2019 [32], De la Garza-Salazar et al., 2020 [33], Kwon et al., 2020 [34], De la Garza-Salazar et al., 2021 [35],

Khurshid et al., 2021 [36], Sabovčik et al., 2021 [37], Angelaki et al., 2021 [38], Lim et al., 2021 [39], Zhao et al., 2022 [40], Sammani et al., 2022 [41], Kokubo et al., 2022 [42], Naderi et al., 2023 [43], Liu et al., 2023 [44].

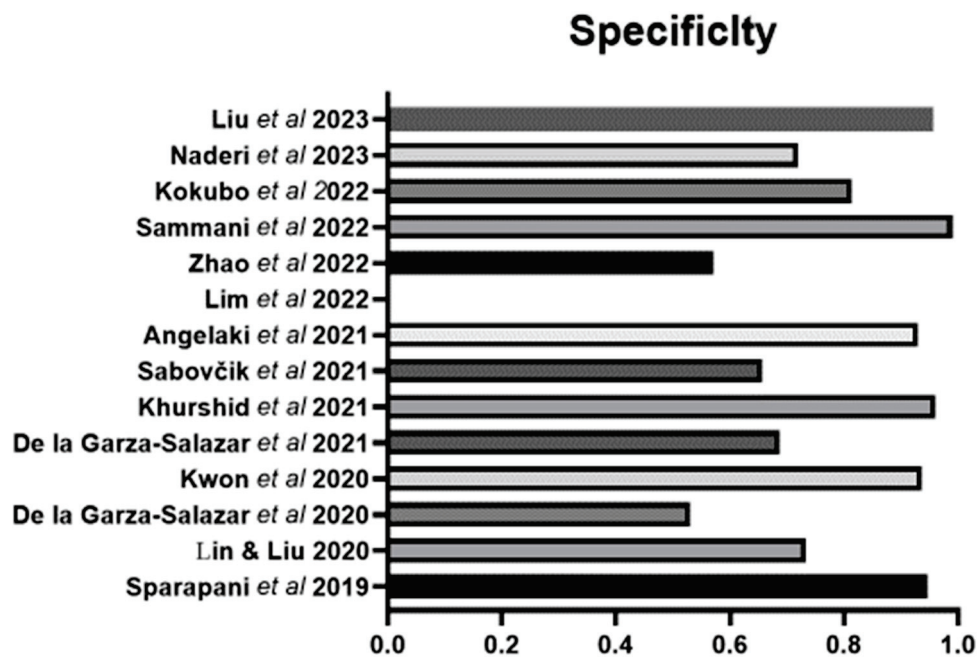


Figure 2. The specificity of the ML algorithms for LVH. Lin & Lui 2020 [31], Sparapani et al., 2019 [32], De la Gar-za-Salazar et al., 2020 [33], Kwon et al., 2020 [34], De la Gar-za-Salazar et al., 2021 [35], Khurshid et al., 2021 [36], Sabovčik et al., 2021 [37], Angelaki et al., 2021 [38], Lim et al., 2021 [39], Zhao et al., 2022 [40], Sammani et al., 2022 [41], Kokubo et al., 2022 [42], Naderi et al., 2023 [43], Liu et al., 2023 [44].

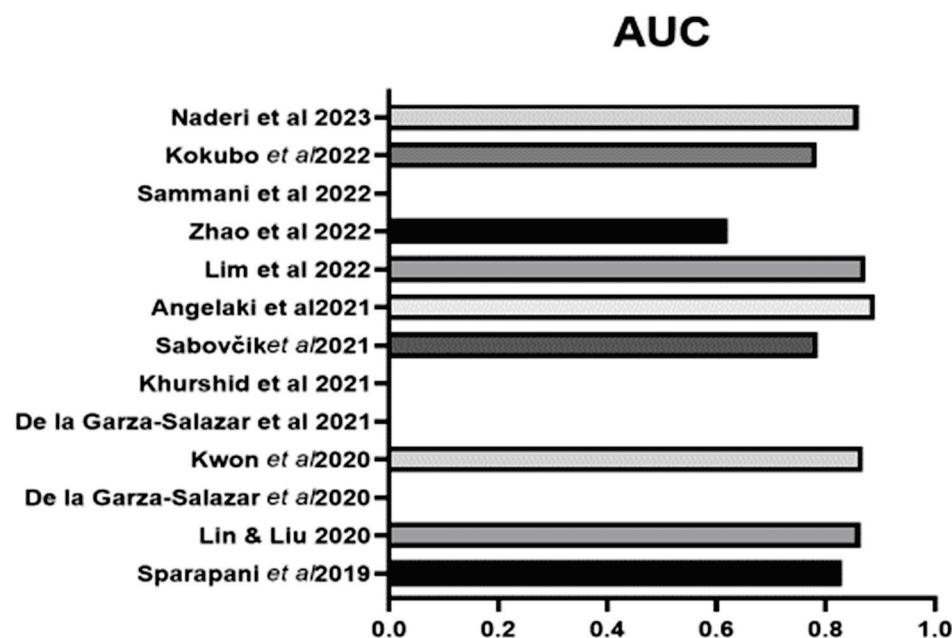


Figure 3. The AUC of the ML algorithms for LVH in those studies that reported such data. Lin & Lui 2020 [31], Sparapani et al., 2019 [32], De la Gar-za-Salazar et al., 2020 [33], Kwon et al., 2020 [34], De la Gar-za-Salazar et al., 2021 [35], Khurshid et al., 2021 [36], Sabovčik et al., 2021 [37], Angelaki et al., 2021 [38], Lim et al., 2021 [39], Zhao et al., 2022 [40], Sammani et al., 2022 [41], Kokubo et al., 2022 [42], Naderi et al., 2023 [43].

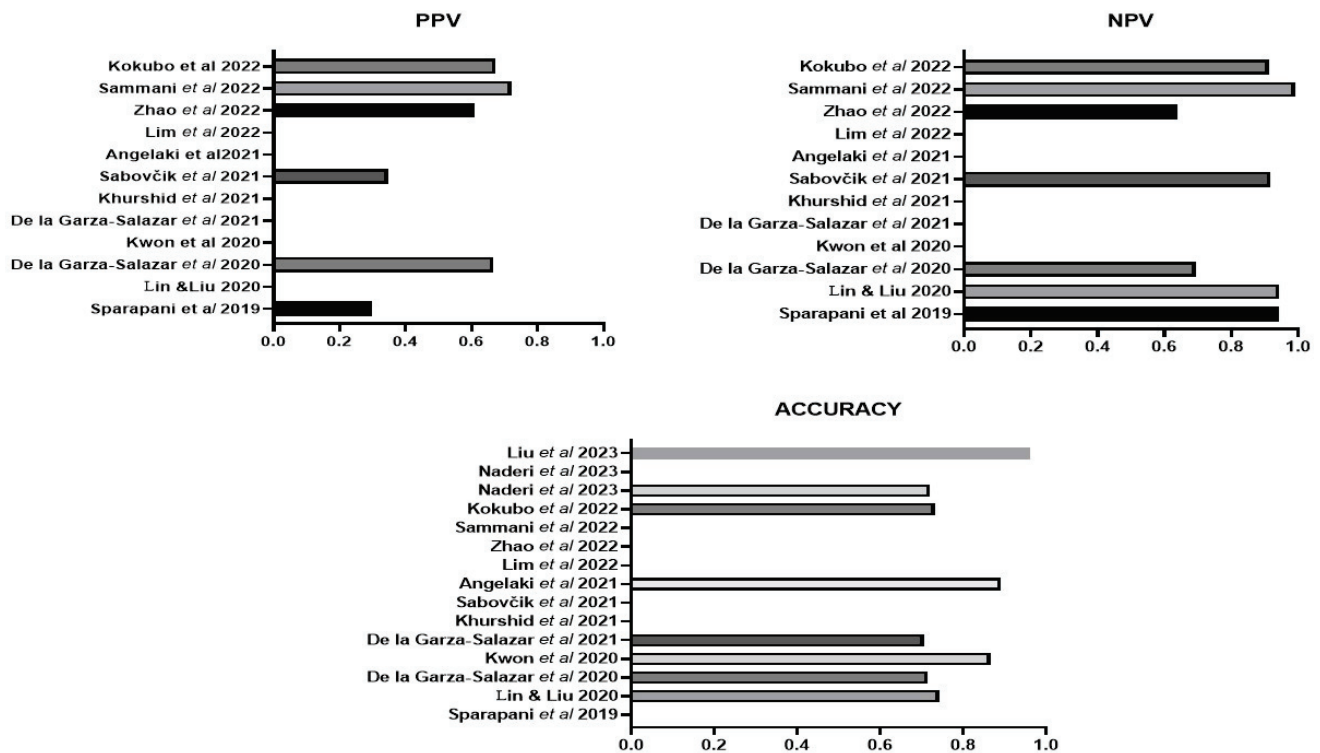


Figure 4. The positive predictive value (PPV), negative predictive value (NPV) and accuracy of the ML algorithms for LVH in those studies that reported such data. Lin & Lui 2020 [31], Sparapani et al., 2019 [32], De la Garza-Salazar et al., 2020 [33], Kwon et al., 2020 [34], De la Garza-Salazar et al., 2021 [35], Khurshid et al., 2021 [36], Sabovčik et al., 2021 [37], Angelaki et al., 2021 [38], Lim et al., 2021 [39], Zhao et al., 2022 [40], Sammani et al., 2022 [41], Kokubo et al., 2022 [42], Naderi et al., 2023 [43], Liu et al., 2023 [44].

Three studies used different ML algorithms and compared them. Kwon et al. found that their AI algorithm based on ENN significantly outperformed the DNN, CNN, RF and LR ones using AUC as the metric [34]. Using the same metric (AUC), Sabovcik et al. reported that XGBoost and RF classifiers exhibited a high area under the receiver operating characteristic curve, with values between 77.7% and 78.5%, for predicting LVH, and these approaches were better than AdaBoost, support vector machines and logistic regression [37]. They did not use an ENN approach. Kokubo et al. found values of 78.4% for the deep learning model (ENN), which was significantly higher than that of the logistic regression and Random Forest methods [42]. Thus, based on the two studies that used ENN, ENN offers a competitive advantage over other ML approaches [34,42].

Nine studies compared their ML approach to the classic ECG approach. The ML algorithm of Zhao et al. outperformed Cornell voltage criteria (AUC 0.57, sensitivity 48%, specificity 72%) and Sokolow-Lyon voltage (AUC 0.51, sensitivity 14%, specificity 96%). [40]. The ML algorithm proposed by Liu et al. reported sensitivity, specificity and accuracy values that were better than the Cornell voltage criteria, Sokolow-Lyons, Peguero, Framingham and Gubner criteria [44]. Of the two ML algorithms presented by De la Garza-Salazar et al., the first had better results than the Romhilt-Estes score, with an accuracy of 61.3%, a sensitivity of 23.2% and a specificity of 94.8% [33], while the second one had an accuracy better than Romhilt-Estes (57%), Cornell (59%) and Sokolow-Lyon (53.9%) [35].

Eight studies reported better sensitivity for their ML algorithm compared to assessment with the Romhilt-Estes point system, Cornell voltage criteria or Sokolow-Lyon criteria (Figure 5). Four of the eight studies reported a specificity of equal to or better than the classic ECG criteria [32,34,36,42] (Figure 6). For some ML algorithms, specificity was higher than the classic ECG criteria, while others did not find a significant difference. Seven

studies reported better AUC for their ML algorithm compared to an assessment with the Romhilt-Estes point system, Cornell voltage criteria or Sokolow-Lyon criteria (Figure 7).

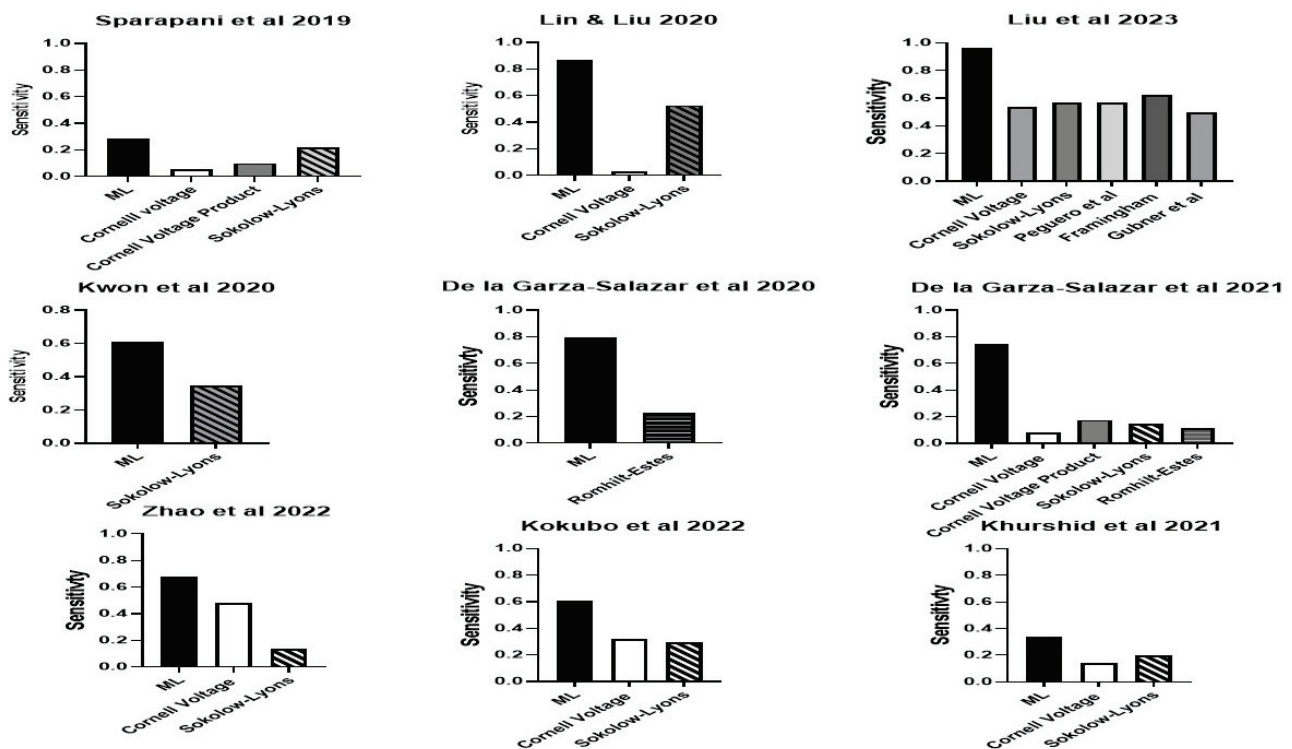


Figure 5. The sensitivity of the ML algorithms for LVH compared to standard ECG LVH criteria in studies that reported such data. Lin & Lui 2020 [31], Sparapani et al., 2019 [32], De la Gar-za-Salazar et al., 2020 [33], Kwon et al., 2020 [34], De la Gar-za-Salazar et al., 2021 [35], Khurshid et al., 2021 [36], Zhao et al., 2022 [40], Kokubo et al., 2022 [42], Liu et al., 2023 [44].

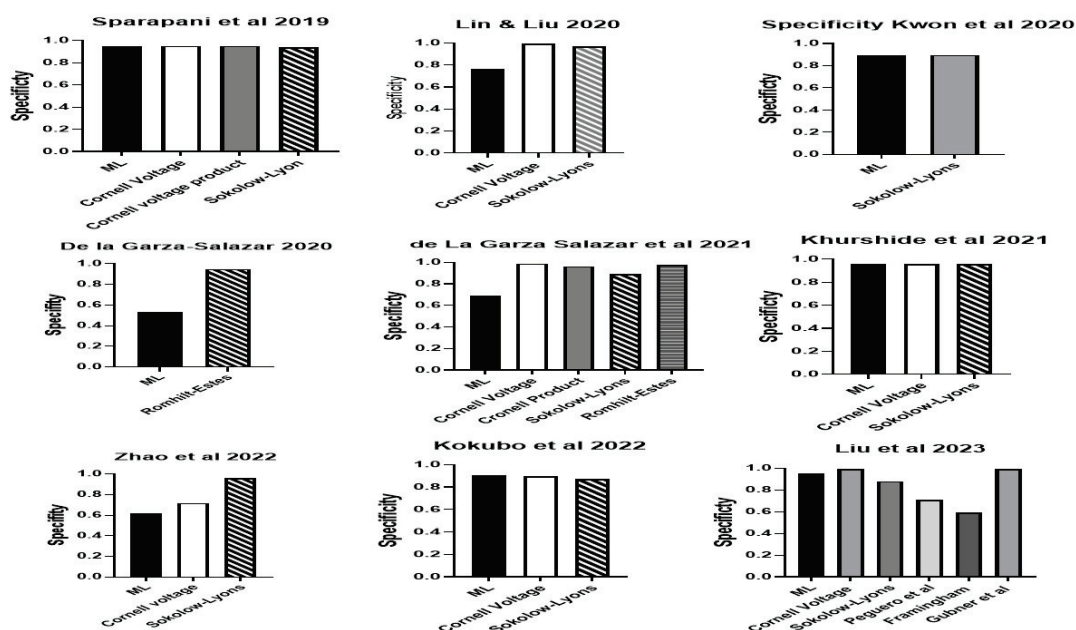


Figure 6. The specificity of the ML algorithms for LVH compared to standard ECG LVH criteria in studies that reported such data. Lin & Lui 2020 [31], Sparapani et al., 2019 [32], De la Gar-za-Salazar et al., 2020 [33], De la Gar-za-Salazar et al., 2021 [35], Khurshid et al., 2021 [36], Zhao et al., 2022 [40], Kokubo et al., 2022 [42], Liu et al., 2023 [44].

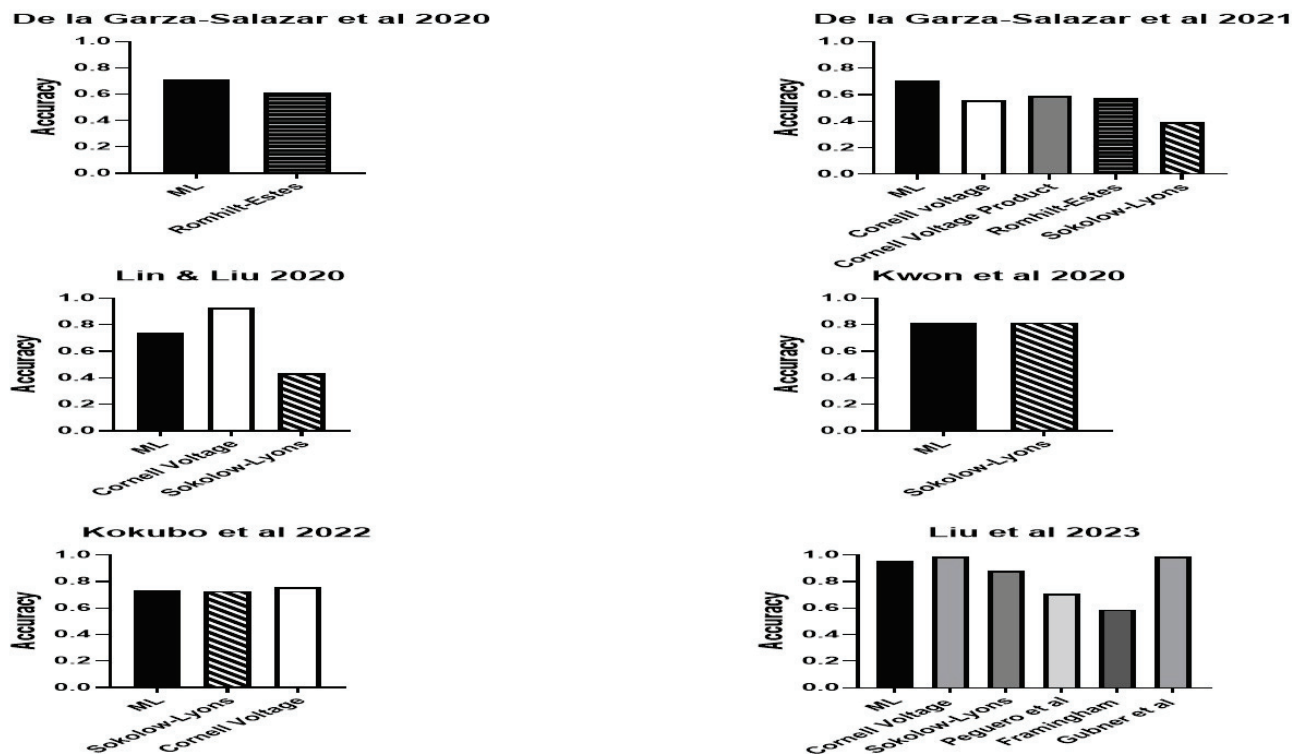


Figure 7. The accuracy of the ML algorithms for LVH compared to standard ECG LVH criteria in studies that reported such data. Lin & Lui 2020 [31], De la Garza-Salazar et al., 2020 [33], Kwon et al., 2020 [34], De la Garza-Salazar et al., 2021 [35], Kokubo et al., 2022 [42], Liu et al., 2023 [44].

Several studies listed the important factors in their ML models. Ignoring the QRS voltage, Lin and Liu reported that there were other significant predictors of LVH, including age, heart rate, PR interval, uncorrected QT interval and QRS axis in Lead II [31]. Systolic and diastolic BP values were in the top-40 predictors of LVH in the algorithm proposed by Naderi et al. [43]. Age and blood pressure were key predictors of LVH in the ML model of Sammani et al., along with P- and T-wave characteristics [41]. Age, waist circumference, different components of BP, history of hypertension, serum renin and antihypertensive treatment were the top predictors of LVH in the algorithm of Sabovic et al. [37].

4. Discussion

This study demonstrates the wide variety of machine learning techniques that have been used to assess the presence of an increased left ventricular mass or cardiac hypertrophy. It demonstrates the differences in sensitivity, specificity and predictive accuracy between ML algorithms. It further identifies large differences in the input variables between algorithms. These differences underscore the necessity to conduct an in-depth evaluation.

The sources of datasets for left ventricular hypertrophy from the ECG in the literature varied widely between studies. Two studies derived data from military recruits, who were essentially young men with a low prevalence of LVH [31,39]. There were three population-based studies, with two studies on the same UK database, with a greater prevalence of LVH [32,36,37,43]. There were eight hospital-based studies, which had, on average, the oldest mean age and the highest prevalence of LVH, with the proportion of men ranging from 42 to 64% [33–35,38,40–42] and one military hospital with a predominance of men (90%) [44]. Overall, the proportion of men and women varied greatly but mainly because of the predominance of men in the studies of young military recruits and in a military hospital. Young male military recruits may not be generalizable to the general population or to older patients admitted to hospital. The prevalence of LVH between studies ranged

from 0.8 to 48% and may have influenced the precision of LVH detection. The majority of studies used echocardiograms to assess the prevalence of LVH, but the LVH criteria varied between studies in Asian or European populations.

ML algorithms may be differentiated by the manner in which they select the boundary that distinguishes different groupings. SVM was used by several groups [31,37,43]. The SVM classifier can use linear or non-linear functions, although linear functions are usually selected. The decision boundary in this method is called the maximum margin classifier, maximum margin hyperplane or the maximum margin hyper plane [45]. Other studies relied on logistic regression [33], a simpler method that tries to maximize the conditional likelihoods; however, it is more prone to outliers than SVMs, which mostly prioritize the points that are closest to the decision boundary. However, LR and SVM often yield similar results [46]. Some studies used RF [37,39,43]. Random forests are a classification algorithm using an ensemble of decision trees, such that each tree depends on the values of a random vector sampled independently, and the generalizability depends on the strength of each tree and the correlation between them [47]. In several clinical diagnosis conditions, RF showed the highest accuracy followed by SVM [48].

Angelaki et al. used SHAP (SHapley Additive exPlanations), a game theoretic approach that connects optimal credit allocation with local explanations, using the classic Shapley values from game theory and their related extensions [38]. A number of studies used multiple ML algorithms [37,39]. Some investigators employed deep learning methods [34,40,42]. The explosive growth of deep learning for ECG data led to the conclusion that a hybrid architecture of a convolutional neural network and recurrent neural network ensemble yielded the best results [49]. However, there are some new challenges and problems related to interpretability, scalability, and efficiency, in addition to differences in the perspectives of datasets and methods [49]. This hybrid combination has been used in a few studies for LVH detection [34,42].

Liu et al. reported both very high sensitivity and specificity. Usually, the higher a test sensitivity, the lower its specificity. They used detailed QRS analysis, but other studies that did not attain as high a sensitivity and specificity also used detailed ECG signal analysis [43]; for example, Zhao et al. had 36,350 ECG segments in their final dataset [40], and another ML algorithm used 552 amplitude and duration measurements per ECG [32]. The findings of Liu et al. [44] showed both very high sensitivity and specificity, but this may relate to their decision that they had too few LVH cases “for designing a machine-learning model. Therefore, the beat segmentation method, Pan-Tompkins technique was performed to increase the ECG data amount to improve the detection performances” [44].

The crucial test of the ML algorithms is the comparative ability to predict LVH. The best or highest sensitivity was the algorithm proposed by Liu et al. [44], followed by Lin and Liu [31] and then by De la Garza-Salazar et al. [33]. If one wants a specific diagnosis, the highest specificity was found using the algorithm proposed by Sammani et al. [41], followed closely by that of Liu et al. [44] and then Khurshid et al. [36]. However, algorithms with high specificity often have low sensitivity. Combining sensitivity and specificity using ROC curves suggests the best approach would be the algorithms of Angelaki et al. [38] followed by Kwon et al. [34]. Several studies compared different ML models to predict LVH [34,39,42]. The differences were usually not large. Two studies compared at least four ML approaches, and both found that ENN had the highest AUC; ENN offered a competitive advantage over other ML approaches [34,42]. Kokopo et al. developed an ensemble neural network (ENN) model, which consisted of a convolutional neural network (CNN) and a deep neural network (DNN) [42]. Kwan et al. used a deep neural network (DNN). Based on the two studies that used ENN, this ML approach (ENN) offered a competitive advantage over other ML approaches [34,42].

Comparisons with classical ECG criteria for LVH showed that ML algorithms were usually more sensitive than the standard Cornell voltage [13], Cornell product [28], Sokolow-Lyons [9] or Romhilt-Estes criteria [29] for the detection of LVH. In contrast, generally, the ML algorithms were not more specific than the classic criteria, as four ML algorithms were

no better and four were worse than these classic criteria for LVH. The ML algorithms of Sparapani et al. [32], Kokubo et al. [42], Kwan et al. [34] and Khurshid et al. [36] had a specificity equal to the classical ECG criteria.

A major theoretical issue with the ML algorithms for the detection of LVH is the use of different kinds of input data. There are several lines of reasoning for the use of ML for LVH diagnosis. The first is whether ML can improve LVH detection based on QRS complexes and especially QRS voltage, which was historically the first attempt to electrocardiographically identify LVH [9]. The second approach is to utilize all aspects of the ECG signal. This was embodied by the work of Romhilt and Estes [29], who added QRS axis and ST-T waves to QRS voltage to identify LVH. As such, ML algorithms can point to the classical approach to justify the inclusion of other ECG factors. The incorporation of clinical factors becomes more problematic in ECG assessment. When age and history of hypertension are included, sensitivity increases markedly, but is that a fair test of the use of ECGs in diagnosis? The addition of an extensive list of clinical and laboratory variables further removes the question from the utility of the ECG but satisfies the question of how to more accurately predict the presence of LVH. For example, Sabovčik et al. inputted a large number of clinical and laboratory variables, including blood count, blood glucose, lipids, renin activity, leptin, insulin, aldosterone and cortisol [37]. Zhao et al. included the input variables of clinical factors, age, sex and medical history, as well as laboratory factors, like hemoglobin, PLT, lipids, creatinine sodium and potassium [40]. The inclusion of such extensive clinical and laboratory data precludes the use of the ECG as a screening test for the presence of LVH, as all the clinical and other laboratory data, which are usually not available, would have to be inputted to utilize the algorithms.

Studies on machine learning-based prediction models have been criticized because of poor methodological quality and a high risk of bias [50]. The criticism relates to the frequent failure ‘to report key information to help readers judge the methods and have a complete, transparent and clear picture of the . . . content of the model’ [51]. This criticism has some validity in the assessment of ML algorithms for the detection/diagnosis of LVH. These kinds of models, because of their complexity, have been labelled as a ‘black box’, certainly compared to regression-based models that can be more recognizable [51]. For example, it is challenging to compare algorithms that state they are derived from 24 features, which consist of R peak and S valley amplitudes automatically obtained from the output of an ECG signal [44] versus raw ECG data, with 5000 numbers from each of the 12 leads [34]. Recognizing the limitations of each of the studies, it is worth discussing the implications of the results. First, ML algorithms can improve the sensitivity of the ECG for the detection of LVH. Improving sensitivity is important for a screening technique, and the ECG fulfills that requirement. Second, simplicity warrants using an algorithm that only relies on ECG variables to add to the ECG interpretation with respect to LVH. Third, algorithms that were developed utilizing a neural network approach appear to offer a competitive advantage over other ML approaches.

There are several limitations of this analysis that warrant discussion. First, the studies usually utilized ML approaches from available ‘packages’. Wallace et al. cautioned that the “near-ubiquitous reliance on ‘out of bag’ approaches may provide ‘misleading results’” [52]. Second, many of the algorithms use ECG variables from most or all of the ECG leads, but in LVH detection, the QRS criteria from multiple leads often provide similar data [53]. Third, not all publications provided the same outputs to compare accuracy, F1 or ROC data. Fourth, it is difficult to compare and select the ‘best’ ML algorithms when one algorithm employs an extensive list of laboratory variables and another uses only ECG factors. One is left with the question whether one approach would be better if it also included an extensive list of laboratory tests.

In summary, it is important to re-emphasize the potential of the ECG to identify LVH because LVH is a significant predictor of cardiovascular events [3,8,19,24] and because a better approach for LVH detection would be an important contribution. Several ML algorithms improve the sensitivity, but most do not improve specificity for LVH diagnosis

compared to classical ECG criteria. Future research is needed to obtain a more standardized approach for the evaluation and comparison of all ML algorithms using the same dataset to determine the competitive advantage of each and identify the best one. In addition, the separation of LVH diagnosis into two stages—an ECG interpretation that uses an ML algorithm and a second step with a simple application—can add further clinical variables.

Supplementary Materials: The following supporting information can be downloaded at: <https://www.mdpi.com/article/10.3390/bioengineering11050489/s1>, Figure S1: Preferred Reporting Items for Reviews and Meta-Analysis (PRISMA).

Funding: No funding to report. There was no sponsor or any contribution from a sponsor.

Data Availability Statement: The data used are available in the literature.

Conflicts of Interest: The author declares no conflicts of interest.

Abbreviations

SVM	Support vector machine
LR	logistic regression
RF	Random Forest
ENN	Ensemble neural network
CNN	Convolutional neural network
DNN	Deep neural network
AUC	Area under the receiver operating curve

References

1. Levy, D.; Garrison, R.J.; Savage, D.D.; Kannel, W.B.; Castelli, W.P. Prognostic implications of echocardiographically determined left ventricular mass in the Framingham Heart Study. *N. Engl. J. Med.* **1990**, *322*, 1561–1566. [CrossRef] [PubMed]
2. Brown, D.W.; Giles, W.H.; Croft, J.B. Left ventricular hypertrophy as a predictor of coronary heart disease mortality and the effect of hypertension. *Am. Heart J.* **2000**, *140*, 848–856. [CrossRef] [PubMed]
3. Abdi-Ali, A.; Miller, R.J.H.; Southern, D.; Zhang, M.; Mikami, Y.; Knudtson, M.; Heydari, B.; Howarth, A.G.; Lydell, C.P.; James, M.T.; et al. LV Mass Independently Predicts Mortality and Need for Future Revascularization in Patients Undergoing Diagnostic Coronary Angiography. *JACC Cardiovasc. Imaging* **2018**, *11*, 423–433. [CrossRef] [PubMed]
4. Rabkin, S.W.; Shiekh, I.A.; Wood, D.A. The Impact of Left Ventricular Mass on Diastolic Blood Pressure Targets for Patients with Coronary Artery Disease. *Am. J. Hypertens.* **2016**, *29*, 1085–1093. [CrossRef] [PubMed]
5. Chatterjee, S.; Bavishi, C.; Sardar, P.; Agarwal, V.; Krishnamoorthy, P.; Grodzicki, T.; Messerli, F.H. Meta-analysis of left ventricular hypertrophy and sustained arrhythmias. *Am. J. Cardiol.* **2014**, *114*, 1049–1052. [CrossRef]
6. Varvarousis, D.; Kallistratos, M.; Poulimenos, L.; Triantafyllis, A.; Tsinivizov, P.; Giannakopoulos, A.; Kyfnidis, K.; Manolis, A. Cardiac arrhythmias in arterial hypertension. *J. Clin. Hypertens.* **2020**, *22*, 1371–1378. [CrossRef] [PubMed]
7. Rabkin, S.W. Considerations in understanding the coronary blood flow- left ventricular mass relationship in patients with hypertension. *Curr. Cardiol. Rev.* **2017**, *13*, 75–83. [CrossRef]
8. Yi, S.; Wang, F.; Wan, M.; Yi, X.; Zhang, Y.; Sun, S. Prediction of stroke with electrocardiographic left ventricular hypertrophy in hypertensive patients: A meta-analysis. *J. Electrocardiol.* **2020**, *61*, 27–31. [CrossRef]
9. Sokolow, M.; Lyon, T.P. The ventricular complex in left ventricular hypertrophy as obtained by unipolar precordial and limb leads. *Am. Heart J.* **1949**, *37*, 161–186. [CrossRef]
10. Koito, H.; Spodick, D.H. Accuracy of the RV6:RV5 voltage ratio for increased left ventricular mass. *Am. J. Cardiol.* **1988**, *62*, 985–987. [CrossRef]
11. Crow, R.S.; Prineas, R.J.; Rautaharju, P.; Hannan, P.; Liebson, P.R. Relation between electrocardiography and echocardiography for left ventricular mass in mild systemic hypertension (results from Treatment of Mild Hypertension Study). *Am. J. Cardiol.* **1995**, *75*, 1233–1238. [CrossRef] [PubMed]
12. Levy, D.; Labib, S.B.; Anderson, K.M.; Christiansen, J.C.; Kannel, W.B.; Castelli, W.P. Determinants of sensitivity and specificity of electrocardiographic criteria for left ventricular hypertrophy. *Circulation* **1990**, *81*, 815–820. [CrossRef] [PubMed]
13. Casale, P.N.; Devereux, R.B.; Kligfield, P.; Eisenberg, R.R.; Miller, D.H.; Chaudhary, B.S.; Phillips, M.C. Electrocardiographic detection of left ventricular hypertrophy: Development and prospective validation of improved criteria. *J. Am. Coll. Cardiol.* **1985**, *6*, 572–580. [CrossRef] [PubMed]
14. Peguero, J.G.; Lo Presti, S.; Perez, J.; Issa, O.; Brenes, J.C.; Tolentino, A. Electrocardiographic criteria for the diagnosis of left ventricular hypertrophy. *J. Am. Coll. Cardiol.* **2017**, *69*, 1694–1703. [CrossRef] [PubMed]
15. Pewsner, D.; Jüni, P.; Egger, M.; Battaglia, M.; Sundström, J.; Bachmann, L.M. Accuracy of electrocardiography in diagnosis of left ventricular hypertrophy in arterial hypertension: Systematic review. *Br. Med. J. Br. Med. J. Publ. Group* **2007**, *335*, 711. [CrossRef] [PubMed]

16. Rautaharju, P.M.; Soliman, E.Z. Electrocardiographic left ventricular hypertrophy and the risk of adverse cardiovascular events: A critical appraisal. *J. Electrocardiol.* **2014**, *47*, 649–654. [CrossRef] [PubMed]
17. Fagard, R.H.; Staessen, J.A.; Thijs, L.; Celis, H.; Birkenhäger, W.H.; Bulpitt, C.J.; de Leeuw, P.W.; Leonetti, G.; Sarti, C.; Tuomilehto, J.; et al. Prognostic significance of electrocardiographic voltages and their serial changes in elderly with systolic hypertension. *Hypertension* **2004**, *44*, 459–464. [CrossRef] [PubMed]
18. Kannel, W.B.; Gordon, T.; Castelli, W.P.; Margolis, J.R. Electrocardiographic left ventricular hypertrophy and risk of coronary heart disease: The Framingham Study. *Ann. Intern. Med. Am. Coll. Physicians* **1970**, *72*, 813–822. [CrossRef] [PubMed]
19. Rabkin, S.W.; Mathewson, F.A.L.; Tate, R.B. The electrocardiogram in apparently healthy men and the risk of sudden death. *Br. Heart J.* **1982**, *47*, 546–552. [CrossRef]
20. De Bacquer, D.; De Backer, G.; Kornitzer, M.; Blackburn, H. Prognostic value of ECG findings for total, cardiovascular disease, and coronary heart disease death in men and women. *Heart* **1998**, *80*, 570–577. [CrossRef]
21. Lonn, E.; Mathew, J.; Pogue, J.; Johnstone, D.; Danisa, K.; Bosch, J.; Baird, M.; Dagenais, G.; Sleight, P.; Yusuf, S.; et al. Relationship of electrocardiographic left ventricular hypertrophy to mortality and cardiovascular morbidity in high-risk patients. *Eur. J. Cardiovasc. Prev. Rehabil.* **2003**, *10*, 420–428. [CrossRef]
22. Hsieh, B.P.; Pham, M.X.; Froelicher, V.F. Prognostic value of electrocardiographic criteria for left ventricular hypertrophy. *Am. Heart J.* **2005**, *150*, 161–167. [CrossRef] [PubMed]
23. Sullivan, J.M.; Vander Zwaag, R.V.; el-Zeky, F.; Ramanathan, K.B.; Mirvis, D.M. Left ventricular hypertrophy: Effect on survival. *J. Am. Coll. Cardiol.* **1993**, *22*, 508–513. [CrossRef]
24. Hawkins, N.M.; Wang, D.; McMurray, J.J.V.; Pfeffer, M.A.; Swedberg, K.; Granger, C.B.; Yusuf, S.; Pocock, S.J.; Ostergren, J.; Michelson, E.L.; et al. Prevalence and prognostic implications of electrocardiographic left ventricular hypertrophy in heart failure: Evidence from the CHARM programme. *Heart* **2007**, *93*, 59–64. [CrossRef] [PubMed]
25. You, Z.; He, T.; Ding, Y.; Yang, L.; Jiang, X.; Huang, L. Predictive value of electrocardiographic left ventricular hypertrophy in the general population: A meta-analysis. *J. Electrocardiol.* **2020**, *62*, 14–19. [CrossRef] [PubMed]
26. Schlant, R.C.; Adolph, R.J.; DiMarco, J.P.; Dreifus, L.S.; Dunn, M.I.; Fisch, C.; Garson, A., Jr.; Haywood, L.J.; Levine, H.J.; Murray, J.A. Guidelines for electrocardiography. A report of the American College of Cardiology/American Heart Association Task Force on Assessment of Diagnostic and Therapeutic Cardiovascular Procedures (Committee on Electrocardiography). *Circulation Am. Heart Assoc.* **1992**, *85*, 1221–1228. [CrossRef]
27. Unger, T.; Borghi, C.; Charchar, F.; Khan, N.A.; Poulter, N.R.; Prabhakaran, D.; Ramirez, A.; Schlaich, M.; Stergiou, G.S.; Tomaszewski, M.; et al. 2020 International Society of Hypertension global hypertension practice guidelines. *J. Hypertens.* **2020**, *38*, 982–1004. [CrossRef]
28. Okin, P.M.; Roman, M.J.; Devereux, R.B.; Kligfield, P. Electrocardiographic identification of increased left ventricular mass by simple voltage-duration products. *J. Am. Coll. Cardiol.* **1995**, *25*, 417–423. [CrossRef]
29. Romhilt, D.W.; Bove, K.E.; Norris, R.J.; Conyers, E.; Conradi, S.; Rowlands, D.T.; Scott, R. A critical appraisal of the electrocardiographic criteria for the diagnosis of left ventricular hypertrophy. *Circ. Am. Heart Assoc.* **1969**, *40*, 185–196. [CrossRef]
30. Page, M.J.; McKenzie, J.E.; Bossuyt, P.M.; Boutron, I.; Hoffmann, T.C.; Mulrow, C.D.; Shamseer, L.; Tetzlaff, J.M.; Akl, E.A.; Brennan, S.E.; et al. The PRISMA 2020 statement: An updated guideline for reporting systematic reviews. *BMJ* **2021**, *372*, n71. [CrossRef]
31. Lin, G.-M.; Liu, K. An Electrocardiographic System with Anthropometrics via Machine Learning to Screen Left Ventricular Hypertrophy among Young Adults. *IEEE J. Transl. Eng. Health Med.* **2020**, *8*, 1800111. [CrossRef] [PubMed]
32. Sparapani, R.; Dabbouseh, N.M.; Gutterman, D.; Zhang, J.; Chen, H.; Bluemke, D.A.; Lima, J.A.C.; Burke, G.L.; Soliman, E.Z. Detection of Left Ventricular Hypertrophy Using Bayesian Additive Regression Trees: The MESA. *J. Am. Heart Assoc.* **2019**, *8*, e009959. [CrossRef] [PubMed]
33. De la Garza-Salazar, F.; Romero-Ibarguengoitia, M.E.; Rodriguez-Diaz, E.A.; Azpiri-Lopez, J.R.; Gonzalez-Cantu, A. Improvement of electrocardiographic diagnostic accuracy of left ventricular hypertrophy using a Machine Learning approach. *PLoS ONE* **2020**, *15*, e0232657. [CrossRef]
34. Kwon, J.-M.; Jeon, K.-H.; Kim, H.M.; Kim, M.J.; Lim, S.M.; Kim, K.-H.; Song, P.S.; Park, J.; Choi, R.K.; Oh, B.-H. Comparing the performance of artificial intelligence and conventional diagnosis criteria for detecting left ventricular hypertrophy using electrocardiography. *EP Eur.* **2020**, *22*, 412–419. [CrossRef] [PubMed]
35. De la Garza Salazar, F.; Romero Ibarguengoitia, M.E.; Azpiri López, J.R.; González Cantú, A. Optimizing ECG to detect echocardiographic left ventricular hypertrophy with computer-based ECG data and machine learning. *PLoS ONE* **2021**, *16*, e0260661. [CrossRef]
36. Khurshid, S.; Friedman, S.; Pirruccello, J.P.; Di Achille, P.; Diamant, N.; Anderson, C.D.; Ellinor, P.T.; Batra, P.; Ho, J.E.; Philippakis, A.A.; et al. Deep Learning to Predict Cardiac Magnetic Resonance-Derived Left Ventricular Mass and Hypertrophy From 12-Lead ECGs. *Circ. Cardiovasc. Imaging* **2021**, *14*, e012281. [CrossRef]
37. Sabovic, F.; Cauwenberghs, N.; Kouznetsov, D.; Haddad, F.; Alonso-Betanzos, A.; Vens, C.; Kuznetsova, T. Applying machine learning to detect early stages of cardiac remodelling and dysfunction. *Eur. Hear. J. Cardiovasc. Imaging* **2021**, *22*, 1208–1217. [CrossRef]

38. Angelaki, E.; Marketou, M.E.; Barmparis, G.D.; Patrianakos, A.; Vardas, P.E.; Parthenakis, F.; Tsironis, G.P. Detection of abnormal left ventricular geometry in patients without cardiovascular disease through machine learning: An ECG-based approach. *J. Clin. Hypertens.* **2021**, *23*, 935–945. [CrossRef] [PubMed]
39. Lim, D.Y.; Sng, G.; Ho, W.H.; Hankun, W.; Sia, C.-H.; Lee, J.S.; Shen, X.; Tan, B.Y.; Lee, E.C.; Dalakoti, M.; et al. Machine learning versus classical electrocardiographic criteria for echocardiographic left ventricular hypertrophy in a pre-participation cohort. *Kardiol. Pol.* **2021**, *79*, 654–661.
40. Zhao, X.; Huang, G.; Wu, L.; Wang, M.; He, X.; Wang, J.-R.; Zhou, B.; Liu, Y.; Lin, Y.; Liu, D.; et al. Deep learning assessment of left ventricular hypertrophy based on electrocardiogram. *Front. Cardiovasc. Med.* **2022**, *9*, 952089. [CrossRef]
41. Sammani, A.; Jansen, M.; de Vries, N.M.; de Jonge, N.; Baas, A.F.; Te Riele, A.S.J.M.; Asselbergs, F.W.; Oerlemans, M.I.F.J. Automatic Identification of Patients with Unexplained Left Ventricular Hypertrophy in Electronic Health Record Data to Improve Targeted Treatment and Family Screening. *Front. Cardiovasc. Med.* **2022**, *9*, 768847. [CrossRef] [PubMed]
42. Kokubo, T.; Kodera, S.; Sawano, S.; Katsushika, S.; Nakamoto, M.; Takeuchi, H.; Kimura, N.; Shinohara, H.; Matsuoka, R.; Nakanishi, K.; et al. Automatic Detection of Left Ventricular Dilatation and Hypertrophy from Electrocardiograms Using Deep Learning. *Int. Heart J.* **2022**, *63*, 939–947. [CrossRef]
43. Naderi, H.; Ramírez, J.; van Duijvenboden, S.; Pujadas, E.R.; Aung, N.; Wang, L.; Anwar Ahmed Chahal, C.; Lekadir, K.; Petersen, S.E.; Munroe, P.B. Predicting left ventricular hypertrophy from the 12-lead electrocardiogram in the UK Biobank imaging study using machine learning. *Eur. Heart J. Digit. Health* **2023**, *4*, 316–324. [CrossRef]
44. Liu, C.-W.; Wu, F.-H.; Hu, Y.-L.; Pan, R.-H.; Lin, C.-H.; Chen, Y.-F.; Tseng, G.-S.; Chan, Y.-K.; Wang, C.-L. Left ventricular hypertrophy detection using electrocardiographic signal. *Sci. Rep.* **2023**, *13*, 2556. [CrossRef]
45. Boser, B.; Guyon, I.; Vapnik, V. A training algorithm for optimal margin classifiers. In Proceedings of the fifth Annual Workshop on Computational Learning Theory—COLT '92, Pittsburgh, PA, USA, 27–29 July 1992; ACM Press: New York, NY, USA, 1992; pp. 144–152.
46. Golpour, P.; Ghayour-Mobarhan, M.; Saki, A.; Esmaily, H.; Taghipour, A.; Tajfard, M.; Ghazizadeh, H.; Moohebat, M.; Ferns, G.A. Comparison of Support Vector Machine, Naïve Bayes and Logistic Regression for Assessing the Necessity for Coronary Angiography. *Int. J. Environ. Res. Public Health* **2020**, *17*, 6449. [CrossRef] [PubMed]
47. Breiman, L. Random Forests. *Mach. Learn.* **2001**, *45*, 5–32. [CrossRef]
48. 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*, 281. [CrossRef]
49. Hong, S.; Zhou, Y.; Shang, J.; Xiao, C.; Sun, J. Opportunities and challenges of deep learning methods for electrocardiogram data: A systematic review. *Comput. Biol. Med.* **2020**, *122*, 103801. [CrossRef] [PubMed]
50. Andaur Navarro, C.L.; Damen, J.A.A.; Takada, T.; Nijman, S.W.J.; Dhiman, P.; Ma, J.; Collins, G.S.; Bajpai, R.; Riley, R.D.; Moons, K.G.M.; et al. Risk of bias in studies on prediction models developed using supervised machine learning techniques: Systematic review. *BMJ* **2021**, *375*, n2281. [CrossRef]
51. Collins, G.S.; Dhiman, P.; Andaur Navarro, C.L.; Ma, J.; Hooft, L.; Reitsma, J.B.; Logullo, P.; Beam, A.L.; Peng, L.; Van Calster, B.; et al. Protocol for development of a reporting guideline (TRIPOD-AI) and risk of bias tool (PROBAST-AI) for diagnostic and prognostic prediction model studies based on artificial intelligence. *BMJ Open* **2021**, *11*, e048008. [CrossRef]
52. Wallace, M.L.; Mentch, L.; Wheeler, B.J.; Tapia, A.L.; Richards, M.; Zhou, S.; Yi, L.; Redline, S.; Buysse, D.J. Use and misuse of random forest variable importance metrics in medicine: Demonstrations through incident stroke prediction. *BMC Med. Res. Methodol.* **2023**, *23*, 144. [CrossRef] [PubMed]
53. Rabkin, S.W.; Zhou, J. Estimating left ventricular mass from the electrocardiogram across the spectrum of LV mass from normal to increased LV mass in an older aged group. *Cardiol. Res. Pract.* **2024**, 6634222. [CrossRef] [PubMed]

Disclaimer/Publisher’s Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.

Article

Exploring Diagnostic Precision and Triage Proficiency: A Comparative Study of GPT-4 and Bard in Addressing Common Ophthalmic Complaints

Roya Zandi ¹, Joseph D. Fahey ¹, Michael Drakopoulos ¹, John M. Bryan ¹, Siyuan Dong ², Paul J. Bryar ¹, Ann E. Bidwell ¹, R. Chris Bowen ¹, Jeremy A. Lavine ¹ and Rukhsana G. Mirza ^{1,*}

¹ Department of Ophthalmology, Feinberg School of Medicine, Northwestern University, Chicago, IL 60611, USA

² Division of Biostatistics, Department of Preventive Medicine, Feinberg School of Medicine, Northwestern University, Chicago, IL 60611, USA

* Correspondence: r-mirza@northwestern.edu

Abstract: In the modern era, patients often resort to the internet for answers to their health-related concerns, and clinics face challenges to providing timely response to patient concerns. This has led to a need to investigate the capabilities of AI chatbots for ophthalmic diagnosis and triage. In this *in silico* study, 80 simulated patient complaints in ophthalmology with varying urgency levels and clinical descriptors were entered into both ChatGPT and Bard in a systematic 3-step submission process asking chatbots to triage, diagnose, and evaluate urgency. Three ophthalmologists graded chatbot responses. Chatbots were significantly better at ophthalmic triage than diagnosis (90.0% appropriate triage vs. 48.8% correct leading diagnosis; $p < 0.001$), and GPT-4 performed better than Bard for appropriate triage recommendations (96.3% vs. 83.8%; $p = 0.008$), grader satisfaction for patient use (81.3% vs. 55.0%; $p < 0.001$), and lower potential harm rates (6.3% vs. 20.0%; $p = 0.010$). More descriptors improved the accuracy of diagnosis for both GPT-4 and Bard. These results indicate that chatbots may not need to recognize the correct diagnosis to provide appropriate ophthalmic triage, and there is a potential utility of these tools in aiding patients or triage staff; however, they are not a replacement for professional ophthalmic evaluation or advice.

Keywords: artificial intelligence; ophthalmology; triage; chatbots; ChatGPT; bard; large language models

1. Introduction

Conversational artificial intelligence (AI) chatbots have gained significant momentum over the last few years. OpenAI's Chat Generative Pre-trained Transformer (ChatGPT), released November 2022, and Google's Bard, launched March 2023, are two chatbots that are publicly available. These systems use large language models (LLMs) to process and generate text similar to human language. LLMs constitute a growing field of technology where computer models are pre-trained on large-scale data to then be adapted to a variety of tasks [1]. While there are several LLMs available today, this work will focus its efforts on ChatGPT and Bard due to their widespread presence and public availability. There are a few key operational differences between the two systems. Namely, ChatGPT uses GPT-3.5 or GPT-4 chatbot models, whereas Bard uses PaLM 2 (Pathways Language Model 2). In addition, Bard draws its data live and directly from Google, whereas ChatGPT operates based on data from 2021, and must search papers to gather information [2,3].

These powerful tools are increasingly being considered for efficiency improvements across medicine, including in such applications as supporting clinical practice, scientific writing, image analysis, or immediate medical advice [4,5]. However, they are not without risk. It has been noted that LLMs can produce biased or harmful content due to the vast variability in quality of the data used to power them; in medicine in particular, the quality of

chatbot output is of concern as it relates to patient care [6]. There has recently been a strong interest in exploring the capabilities of these LLMs in medicine. Early work demonstrated that OpenAI's GPT-3.5 performed at or near passing for all three exams in the United States Medical Licensing Exam (USMLE) series [7]. With newer iterations, GPT-4, released March 2023, was found to outperform GPT-3.5 in correctly answering USMLE questions involving communication skills, ethics, empathy, and professionalism [8]. More recent studies have compared GPT-4 and Bard in their performance of answering board-style questions in various subspecialties, and these showed that GPT-4 was superior to Bard [9–11]. Within ophthalmology, there has been an interest as well, where in one study GPT-4 demonstrated an excellent performance, significantly better than GPT-3.5, in answering practice questions to the Ophthalmology Knowledge Assessment Program (OKAP) examination [12].

As AI chatbots evolve, become more widely used by the general population, and are integrated into common internet search engines, it is increasingly imperative to assess their role in the patient care journey. It is a well-established trend that patients look to the internet for seeking information about their health [13,14] and often turn to the internet first for health advice before contacting health professionals [15,16]. Moreover, patients often have long wait times when they do contact their health providers, which is especially true in ophthalmology. A recent study projected that there will be a sizable shortage of ophthalmologists relative to demand by the year 2035 [17], with limited ophthalmology coverage in emergency departments, especially in rural settings [18]. As a result, non-ophthalmology providers, busy triage call centers, and patients may begin to look to technological solutions such as AI chatbots that can support addressing ophthalmic complaints and triage. It is therefore critical that the strengths and potential risks of these tools are evaluated thoroughly.

Recent studies have begun to explore the capabilities of AI chatbots as ocular symptom checkers or ophthalmic triage tools. Specifically, Pushpanathan et al. investigated accuracy and quality of responses (without examining triage capabilities) for GPT-3.5, GPT-4, and Bard in answering direct questions about specific ocular symptoms and found that GPT-4 had the highest accuracy [19]. Lim et al. benchmarked performance of ChatGPT and Bard for myopia-related queries specifically, and also found that GPT-4.0 had superior accuracy [20]. Lyons et al. compared the triage capabilities of GPT-4, Bing Chat, and WebMD Symptom Checker with ophthalmology trainees across 24 ophthalmic diagnoses. Notably, GPT-4 performed comparably with the trainees in diagnostic and triage accuracy [21].

In this work, we aimed to evaluate and compare GPT-4 and Bard in their responses to commonly encountered ophthalmic complaints corresponding to 40 critical diagnoses in the form of simulated patient vignettes with targeted questions about proposed diagnoses and triage recommendations. We additionally analyzed how prompt descriptiveness impacts response quality with the aim to better understand how they would best be used in future patient-oriented settings. As we gain a better understanding of the values and limitations of this technology, we can move closer to determining how conversational AI can potentially be implemented in day-to-day society for meeting the demands of delivering timely, accurate, and safe ophthalmic health information for patient use and decision making.

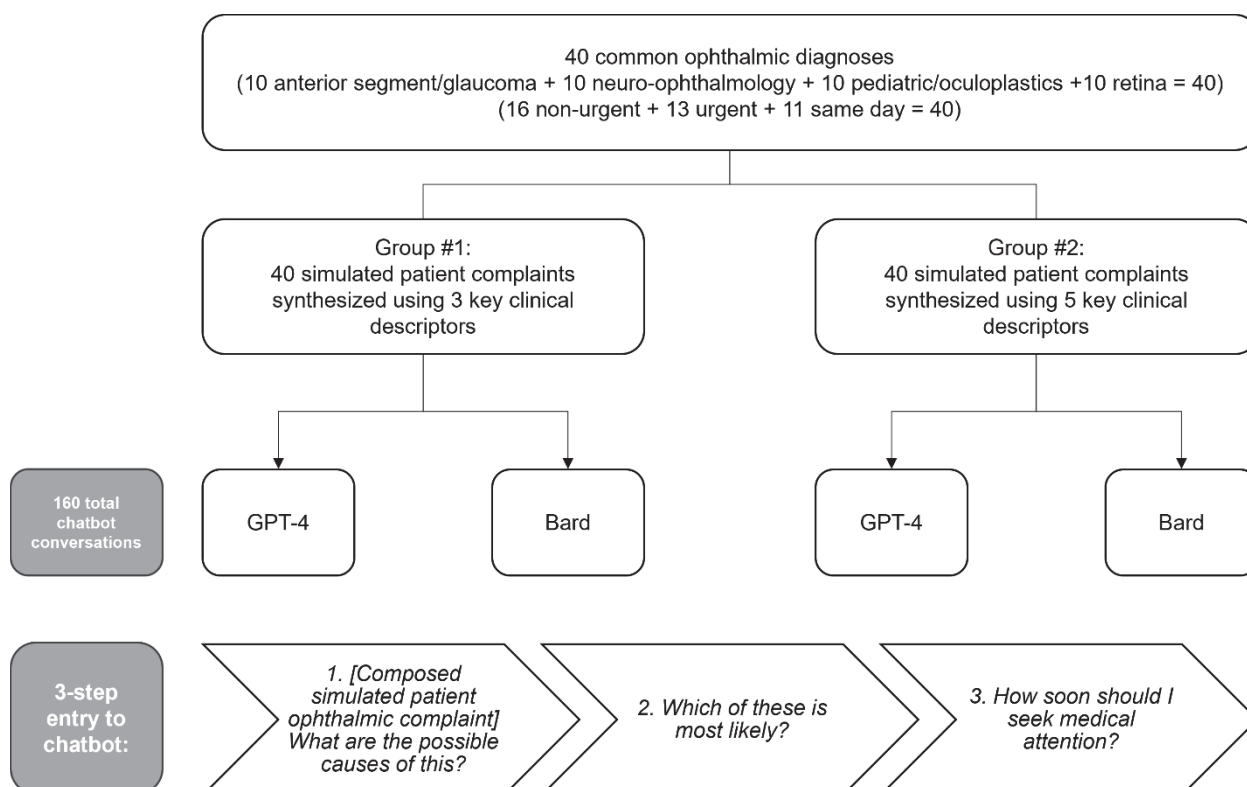
2. Materials and Methods

The Northwestern University Institutional Review Board determined that this *in silico* research did not involve human subjects. At the time of data collection, GPT-4 was publicly available by paid subscription through ChatGPT Plus, and Bard was freely accessible.

2.1. Creation of Simulated Patient Prompts in Ophthalmology

We systematically constructed common scenarios encountered in ophthalmology from the perspective of a patient. Forty common diagnoses, including “cannot miss diagnoses” [22], were identified and distributed evenly among four groups of ophthalmic specialties: anterior segment/glaucoma, neuro-ophthalmology, pediatric ophthalmol-

ogy/oculoplastics, and retina. An urgency level to seek care was designated for each diagnosis as either same day, urgent (<1 week), or non-urgent (>1 week). For each diagnosis, two prompts were created, one with three key clinical descriptors and one with five descriptors (Scheme 1). A descriptor was defined as a clinically relevant piece of information that addressed any of the following: relevant history, onset, duration, laterality, mention of specific ocular anatomy, vision, dyschromatopsia, pain, photophobia, visual disturbances, or any other clinical characteristic. For each patient scenario, consensus was reached among experienced ophthalmologists (P.B., A.E.B., and R.C.B) regarding both intended diagnosis and urgency level based on expert opinion (Supplemental Tables S1–S4).



Scheme 1. Flowchart of overall study design of chatbot prompts.

2.2. Input to Artificial Intelligence Chatbots

The simulated patient prompts were entered into the AI chatbots between 14 June 2023 and 20 June 2023 using GPT-4 version 2023.05.24 and Bard version 2023.06.07. While many chatbots exist, including Microsoft Bing AI, Claude AI, or Meta’s LLaMA, we chose these two as they are among the most commonly used and referenced chatbots at the time of publication, and they are publicly available. Each prompt was entered into the chatbot using a standardized 3-part stepwise approach. First, the simulated patient scenario followed by the question “What are the possible causes of this?” was entered to the chatbot. The second entry was “Which of these is most likely?”. Finally, the third input was “How soon should I seek medical attention?” Chatbot history was reset prior to starting each 3-part entry. These sequential questions were to ensure that the chatbot addressed a differential diagnosis, leading diagnosis, and provided triage recommendations.

2.3. Grading of Chatbot Responses

To evaluate the response generated by the chatbots, a seven-question questionnaire was designed using both a 4-point Likert scale and binary (yes/no) style questions (Supplemental Table S5). Prior to grading, chatbot identifiers (e.g., “I’m an AI developed by OpenAI”) were removed from chatbot responses to eliminate potential grader bias toward

a particular chatbot. Two experienced ophthalmologists graded each chatbot conversation for accuracy of diagnosis and appropriateness of triage recommendations (primary outcomes), as well as relevance of differential diagnosis, satisfaction with quality of responses for real patient use, and potential harm that responses may pose to real patients (secondary outcomes). A third experienced ophthalmologist served as arbiter for any grading disagreements. All graders were blinded to the chatbot source. For the purposes of this analysis, responses of 3 or 4 on the 4-point Likert scale were treated as “agree” and responses of 1 or 2 were treated as “disagree” to provide binary data.

2.4. Statistical Analysis

Descriptive statistics were generated for all variables of interest where frequencies along with percentages were reported. To compare the outcomes of interest between ChatGPT and Bard, as well as between 5 and 3 descriptors, Pearson’s Chi-squared test or Fisher’s exact test was used when appropriate. Sub-analyses for sub-specialty categories and urgency levels were conducted using the same method. Logistic regression models were applied to the primary outcomes with the degree of detail in prompt as the predictor, and the models were fit separately for ChatGPT and Bard. Model performance was estimated using Area under Curve (AUC) and receiver operating characteristic (ROC) curves. All analyses were conducted using R version 4.3.1.

3. Results

Eighty unique entries were supplied to both GPT-4 and Bard, resulting in a total of 160 chatbot generated responses. 40 entries had 3 prompt descriptors and a counterpart 40 entries had 5 prompt descriptors. The 40 diagnoses were broken down into 4 sub-specialty categories (10 general, 10 neuro-ophthalmology, 10 pediatrics/oculoplastics, and 10 retina) and 3 urgency levels (16 non-urgent, 13 urgent, and 11 same day) (Scheme 1). The diagnosis rates (i.e., providing the correct diagnosis as the stated most likely cause of the patient’s symptoms) for GPT-4 and Bard were 53.8% and 43.8%, respectively ($p = 0.2$). Interestingly, both chatbots were significantly better at providing triage recommendations than at providing the correct leading diagnosis (GPT-4: $p < 0.001$, Bard: $p < 0.001$). The rates of generally appropriate triage recommendations for GPT-4 and Bard were 96.3% and 83.8%, respectively ($p = 0.008$) (Table 1).

Table 1. Primary and secondary outcomes of Bard and GPT-4.

	Overall N = 160 ¹	Bard N = 80 ¹	GPT-4 N = 80 ¹	p-Value ²
Primary outcomes of Bard and GPT-4 overall				
Correct diagnosis as most likely cause of symptoms	78 (48.75%)	35 (43.75%)	43 (53.75%)	0.2
Correct diagnosis somewhere in the chatbot conversation	125 (78.13%)	58 (72.50%)	67 (83.75%)	0.085
“Somewhat” or “completely” appropriate triage recommendations	144 (90.00%)	67 (83.75%)	77 (96.25%)	0.008 *
“Completely” appropriate triage recommendations	123 (76.88%)	55 (68.75%)	68 (85.00%)	0.015 *
Secondary outcomes of Bard and GPT-4 overall				
“Somewhat” or “very” relevant differential diagnosis	140 (87.50%)	66 (82.50%)	74 (92.50%)	0.056
Graders “somewhat” or “very” satisfied with quality of chatbot response for actual patient use	109 (68.13%)	44 (55.00%)	65 (81.25%)	<0.001 *
Potentially harmful for patients	21 (13.13%)	16 (20.00%)	5 (6.25%)	0.010 *

* p -value < 0.05. ¹ N (%). ² Pearson’s Chi-squared test between Bard and GPT-4.

Secondary outcomes included relevance of differential diagnosis, grader satisfaction with chatbot responses, and expert opinion as to whether the chatbot response could pose harm if provided to an actual patient. Of all the 160 responses, the differential diagnoses were generally relevant (87.5%). Additionally, graders indicated satisfaction with 109 responses (68.1%); the satisfaction rate was significantly higher for responses from GPT-4 than from Bard (81.3% vs. 55.0%, respectively; $p < 0.001$). Graders reported that 21 of 160 chatbot responses (13.1%) would pose harm if provided to an actual patient; GPT-4 had a lower potential harm rate than Bard (6.3% vs. 20.0%; $p = 0.010$) (Table 1).

Additional sub-analyses were performed comparing the 3 and 5 descriptor cohorts. Notably, increasing the degree of prompt descriptiveness resulted in significant improvement in diagnosis rates for both GPT-4 and Bard (GPT-4: 42.5% vs. 65.0%, respectively; $p = 0.044$, Bard: 32.5% vs. 55.0%, respectively; $p = 0.043$), whereas triage recommendations did not significantly improve (Table 2). However, the model performance of prompt descriptiveness predicting appropriate triage for ChatGPT (AUC 0.587) was better than Bard (AUC 0.523) (Supplemental Figure S1). Of note, increasing the number of descriptors (from 3 to 5) resulted in significantly higher grader satisfaction for GPT-4 (70.0% vs. 92.5%; $p = 0.010$), but not for Bard (47.5% vs. 62.5%; $p = 0.2$) (Table 2).

Table 2. Primary and secondary outcomes between 3 and 5 descriptors for Bard and GPT-4.

	Bard			GPT-4		
	3 Descriptors N = 40 ¹	5 Descriptors N = 40 ¹	<i>p</i> -Value ²	3 Descriptors N = 40 ¹	5 Descriptors N = 40 ¹	<i>p</i> -Value ³
Primary outcomes of Bard and GPT-4 overall						
Correct diagnosis as most likely cause of symptoms	13 (32.50%)	22 (55.00%)	0.043 *	17 (42.50%)	26 (65.00%)	0.044 *
Correct diagnosis somewhere in the chatbot conversation	29 (72.50%)	29 (72.50%)	>0.9	30 (75.00%)	37 (92.50%)	0.034 *
“Somewhat” or “completely” appropriate triage recommendations	33 (82.50%)	34 (85.00%)	0.8	38 (95.00%)	39 (97.50%)	>0.9
“Completely” appropriate triage recommendations	24 (60.00%)	31 (77.50%)	0.091	32 (80.00%)	36 (90.00%)	0.2
Secondary outcomes of Bard and GPT-4 overall						
“Somewhat” or “very” relevant differential diagnosis	35 (87.50%)	31 (77.50%)	0.2	36 (90.00%)	38 (95.00%)	0.7
Graders “somewhat” or “very” satisfied with quality of chatbot response for actual patient use	19 (47.50%)	25 (62.50%)	0.2	28 (70.00%)	37 (92.50%)	0.010 *
Potentially harmful for patients	10 (25.00%)	6 (15.00%)	0.3	5 (12.50%)	0 (0.00%)	0.055

* p -value < 0.05. ¹ N (%). ² Pearson’s Chi-squared test. ³ Fisher’s exact test; Pearson’s Chi-squared test.

Further sub-analyses were performed to compare GPT-4 with Bard and this demonstrated that GPT-4 performed better than Bard in the 5 descriptor cohort when considering the responses that listed the correct diagnosis anywhere within the chatbot response (92.5% vs. 72.5%; $p = 0.019$). GPT-4 also performed significantly better than Bard in the 5 descriptor cohort in generating relevant differential diagnoses (95.0% vs. 77.5%; $p = 0.023$). The satisfaction rate was also significantly higher for GPT-4 than Bard (3 descriptor group: 70.0% vs. 47.5%; $p = 0.041$, 5 descriptor group: 92.5% vs. 62.5%; $p = 0.001$). Within the 5 descriptor cohort, the rate of potential to cause patient harm was zero for GPT-4 and 15.0% for Bard ($p = 0.026$) (Table 3).

Table 3. Primary and secondary outcomes between Bard and GPT-4 for 3 and 5 descriptor cohorts.

	3 Descriptors			5 Descriptors		
	Bard N = 40 ¹	GPT-4 N = 40 ¹	<i>p</i> -Value ²	Bard N = 40 ¹	GPT-4 N = 40 ¹	<i>p</i> -Value ²
Primary outcomes of Bard and GPT-4 overall						
Correct diagnosis as most likely cause of symptoms	13 (32.50%)	17 (42.50%)	0.4	22 (55.00%)	26 (65.00%)	0.4
Correct diagnosis somewhere in the chatbot conversation	29 (72.50%)	30 (75.00%)	0.8	29 (72.50%)	37 (92.50%)	0.019 *
“Somewhat” or “completely” appropriate triage recommendations	33 (82.50%)	38 (95.00%)	0.2	34 (85.00%)	39 (97.50%)	0.11
“Completely” appropriate triage recommendations	24 (60.00%)	32 (80.00%)	0.051	31 (77.50%)	36 (90.00%)	0.13
Secondary outcomes of Bard and GPT-4 overall						
“Somewhat” or “very” relevant differential diagnosis	35 (87.50%)	36 (90.00%)	>0.9	31 (77.50%)	38 (95.00%)	0.023 *
Graders “somewhat” or “very” satisfied with quality of chatbot response for actual patient use	19 (47.50%)	28 (70.00%)	0.041 *	25 (62.50%)	37 (92.50%)	0.001 *
Potentially harmful for patients	10 (25.00%)	5 (12.50%)	0.2	6 (15.00%)	0 (0.00%)	0.026 *

* *p*-value < 0.05. ¹ N (%). ² Fisher’s exact test; Pearson’s Chi-squared test.

Additional sub-analyses of the GPT-4–5 descriptor cohort revealed that the chatbot performed similarly for all outcome measures regardless of urgency level and subspecialty of diagnosis (Supplemental Tables S6 and S7).

4. Discussion

To our knowledge, this is the first work to investigate both the diagnostic accuracy and appropriateness of triage recommendations of GPT-4 and Bard in response to simulated ophthalmic complaints of varying degrees of descriptiveness. It also uses the largest sample size of responses. Overall, the chatbots were significantly better at ophthalmic triage than at providing the correct diagnosis; notably, GPT-4 displayed high rates of appropriate triage—which supports data found in another recent study [21]. Our work demonstrates that GPT-4 performed significantly better than Bard in the domains of appropriate triage recommendations, responses that experts were satisfied with for patient use, and responses that were not considered to cause harm if given to real patients. While some of the results were not statistically significant in the sub-analyses, this was likely due to the smaller sample size. It should be highlighted, however, that in the 5 descriptor sub-analysis, GPT-4 performed significantly better than Bard in considering the correct diagnosis as either the most likely diagnosis or as one of the possible diagnoses in the differential diagnosis (92.5% vs. 72.5%; *p* = 0.019). This is in agreement with recent research demonstrating the superiority of GPT-4 to Bard in correctly answering questions related to ocular symptoms [19] and myopia-related queries [20]. In addition, our work uniquely reveals that increasing the detail of chatbot input (more descriptors) generally improved the quality of output. It should be emphasized that the chatbots were able to provide appropriate triage recommendations without necessarily recognizing the exact diagnosis which better lends itself as an ophthalmic triage tool than as a diagnostic tool.

Another critical question to consider is the performance of these chatbots in the context of do-not-miss diagnoses that are vision- or life-threatening, such as an oculomotor nerve palsy, endophthalmitis, or acute angle closure crisis. In these cases, humans might be trained to take extreme caution when giving guidance to patients, and adoption of conversational AI tools in this space may depend on the responses in such cases. Here, we examined the superior performing chatbot (GPT-4) under optimal conditions (5 descriptor

prompts), and we found that all 11 entries with do-not-miss diagnoses resulted in generally appropriate triage recommendations and responses that senior ophthalmologists were satisfied with for patient use. Moreover, there were no responses (0 of 40) in the GPT-4-5 descriptor subgroup that were considered to be potentially harmful. This is particularly valuable as we consider the potential applications of this technology for future patient use, either as a self-inquiry tool or as an adjunct tool for medical staff to execute timely, appropriate, and safe patient triage.

This study's moderate sample size of chatbot responses is one of its many strengths. We used a highly systematic approach to develop and input all chatbot entries, with chatbot history being reset following each entry to eliminate the variable of chatbot growth over the course of data collection. In addition, our 3-step approach to inputting entries for each "conversation" attempted to take advantage of the conversational capabilities of these chatbots. Lastly, the chatbot responses were all gathered within a one-week time-frame during which all responses were generated from a single iteration of either GPT-4 or Bard, then statistical analysis was performed.

Based on the results found in this work, chatbot responses in the current state of technology are promising but not a sufficient substitute for professional medical advice, yet only in a handful of chatbot responses, GPT-4 more often than Bard, were there such explicit disclaimers. Some examples from GPT-4 include: "I'm an AI developed by OpenAI and while I can help suggest some potential causes for your symptoms, I'm not a substitute for professional medical advice", "Please note that this advice does not substitute professional medical advice. Always consult with a healthcare provider for medical concerns", or "Remember that while the internet can provide useful general advice, it's no substitute for the professional judgment of a healthcare provider who can evaluate your child in person". It should also be added that in less than half of the GPT-4 responses (28 of 80) and in only one Bard response, was there a specific comment about being an AI; interestingly, Bard only indicated itself as such when unable to respond ["I'm a text-based AI and can't assist with that."]. In this work, we inputted the chatbot prompts from the perspective of a simulated patient; in the future, it would be worthwhile to assess how chatbot responses would differ if the chatbot was prompted to answer questions while identifying itself as an AI-generated ophthalmic triage staff. Nonetheless, the chatbot's recognition of self-limitations as an AI and its recommendation to seek professional medical evaluation are important elements that serve as safety checks to the general public who may already be using this technology to answer their own health-related questions.

5. Limitations

Prior to considering the implementation of such technology for clinical use, further studies with larger data sets should be performed. Another shortcoming of this study is the relatively subjective definition of a "descriptor", where some modifiers might be more informative than others; nonetheless, our results overall do show that a more descriptive chatbot prompt is desirable. A general concern of conversational AI that must also be highlighted is the risk of generating "hallucinations", seemingly accurate information that are in fact false [23]. While our study found that chatbot responses, especially GPT-4, were typically not harmful, we did not specifically investigate the number of hallucinatory responses. The potential of generating convincing misinformation is a serious concern that should not be taken lightly and should be further explored, especially as different iterations of these software are developed.

6. Conclusions

In this work we have evaluated two AI chatbots for their use in ophthalmology; however, as the usage of these tools increases over the coming years, it is imperative to continue their evaluation. As future iterations of GPT-4, Bard, or other LLMs are published, each of these should be tested anew. Additionally, more reviewers could assess a larger sample size of chatbot responses in order to provide greater accountability for the variability

in AI responses. In addition, the language of prompts can be more varied in future studies to allow for variability from the user. Finally, more in-depth studies of chatbots' use may be performed over a longitudinal study following patients' real-time diagnoses and the chatbots' capabilities in diagnosis and triage.

Another aspect of AI chatbots to study further includes incorporating images with written text of patient concerns to assess how diagnostic accuracy and triage recommendations vary with the added variable of clinical photos. Given that ophthalmology is a highly visual discipline, it would be interesting to assess how external photos of the eye (i.e., the type of photo that a patient could realistically provide) in conjunction with clinical context would impact chatbot responses.

While this work sheds light on the performance and potential utility of GPT-4 and Bard in the domain of ophthalmic diagnostics and triage, the broader scientific understanding of conversational AI in medicine is still in its infancy as there is an endless number of ways to engage with these chatbots. Recognizing the optimal approach of feeding information into the chatbot and evaluating the quality of the resultant response is imperative to advancing towards real-world application of conversational AI for patient use, either as a self-triage tool or as an adjunct triage tool for medical staff. Our results herein suggest that currently, GPT-4 outperforms Bard and that a greater number of key clinical descriptors for chatbot input is desirable. But only until we have established a full understanding of the strengths and weaknesses of AI chatbots and have been able to consistently achieve a high level of excellence in the quality of responses should we consider their incorporation in patient care. As the world quickly moves towards greater use of conversational AI and a greater need in clinical settings for technological solutions, the urgency for investigative studies like this one will only increase.

Supplementary Materials: The following supporting information can be downloaded at: <https://www.mdpi.com/article/10.3390/bioengineering11020120/s1>, Table S1: Compilation of anterior segment/glaucoma/comprehensive simulated patient complaints; Table S2: Compilation of neuro-ophthalmology simulated patient complaints; Table S3: Compilation of pediatric/oculoplastic simulated patient complaints; Table S4: Compilation of retina simulated patient complaints; Table S5: Questionnaire to grade chatbot responses; Table S6: Sub-analysis of primary and secondary outcomes of GPT-4–5 descriptor cohort by urgency level; Table S7: Sub-analysis of primary and secondary outcomes of GPT-4–5 descriptor cohort by subspecialty; Figure S1: Receiver operating characteristic curves for Bard and GPT-4 in providing triage recommendations.

Author Contributions: Conceptualization: R.G.M.; Data curation: R.Z., J.D.F., M.D. and J.M.B.; Funding acquisition: R.G.M.; Formal analysis: S.D.; Investigation: R.Z.; Methodology: R.Z., M.D., P.J.B., A.E.B., R.C.B., J.A.L. and R.G.M.; Project Administration: R.Z. and R.G.M.; Software: R version 4.3.1; Supervision: R.G.M.; Validation: P.J.B., A.E.B. and R.C.B.; Visualization: R.Z. and R.G.M.; Writing—original draft preparation: R.Z. and S.D.; Writing—review and editing: R.G.M. All authors have read and agreed to the published version of the manuscript.

Funding: This work was funded in part by an unrestricted departmental grant from Research to Prevent Blindness. JAL was supported by NIH grant K08 EY030923, R01 EY034486, and the Research to Prevent Blindness Sybil B. Harrington Career Development Award for Macular Degeneration. The funding agency had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Institutional Review Board Statement: The Institutional Review Board of Northwestern University determined that this work was not human research and therefore did not require ethical approval.

Informed Consent Statement: Not applicable.

Data Availability Statement: The original contributions presented in the study are included in the article/supplementary material, further inquiries can be directed to the corresponding author.

Conflicts of Interest: J.A.L. is a consultant for Genentech, Inc. R.C.B. is a cofounder of Stream Dx, Inc. R.G.M. has received research support from Google Inc. No party had any role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

References

1. Tian, S.; Jin, Q.; Yeganova, L.; Lai, P.-T.; Zhu, Q.; Chen, X.; Yang, Y.; Chen, Q.; Kim, W.; Comeau, D.C. Opportunities and challenges for ChatGPT and large language models in biomedicine and health. *Brief. Bioinform.* **2024**, *25*, bbad493. [CrossRef]
2. Singh, S.K.; Kumar, S.; Mehra, P.S. Chat GPT & Google Bard AI: A Review. In Proceedings of the 2023 International Conference on IoT, Communication and Automation Technology (ICICAT), Online, 23–24 June 2023; pp. 1–6.
3. Thirunavukarasu, A.J.; Ting, D.S.J.; Elangovan, K.; Gutierrez, L.; Tan, T.F.; Ting, D.S.W. Large language models in medicine. *Nat. Med.* **2023**, *29*, 1930–1940. [CrossRef]
4. Cascella, M.; Montomoli, J.; Bellini, V.; Bignami, E. Evaluating the feasibility of ChatGPT in healthcare: An analysis of multiple clinical and research scenarios. *J. Med. Syst.* **2023**, *47*, 33. [CrossRef] [PubMed]
5. Zheng, Y.; Wang, L.; Feng, B.; Zhao, A.; Wu, Y. Innovating healthcare: The role of ChatGPT in streamlining hospital workflow in the future. *Ann. Biomed. Eng.* **2023**, *18*, 1–4. [CrossRef] [PubMed]
6. Ouyang, L.; Wu, J.; Jiang, X.; Almeida, D.; Wainwright, C.; Mishkin, P.; Zhang, C.; Agarwal, S.; Slama, K.; Ray, A. Training language models to follow instructions with human feedback. *Adv. Neural Inf. Process. Syst.* **2022**, *35*, 27730–27744.
7. Kung, T.H.; Cheatham, M.; Medenilla, A.; Sillos, C.; De Leon, L.; Elepaño, C.; Madriaga, M.; Aggabao, R.; Diaz-Candido, G.; Maningo, J. Performance of ChatGPT on USMLE: Potential for AI-assisted medical education using large language models. *PLoS Digit. Health* **2023**, *2*, e0000198. [CrossRef] [PubMed]
8. Brin, D.; Sorin, V.; Vaid, A.; Soroush, A.; Glicksberg, B.S.; Charney, A.W.; Nadkarni, G.; Klang, E. Comparing ChatGPT and GPT-4 performance in USMLE soft skill assessments. *Sci. Rep.* **2023**, *13*, 16492. [CrossRef]
9. Ali, R.; Tang, O.Y.; Connolly, I.D.; Fridley, J.S.; Shin, J.H.; Sullivan, P.L.Z.; Cielo, D.; Oyelese, A.A.; Doberstein, C.E.; Telfeian, A.E. Performance of ChatGPT, GPT-4, and Google bard on a neurosurgery oral boards preparation question bank. *Neurosurgery* **2022**, *93*, 1090–1098. [CrossRef]
10. Patil, N.S.; Huang, R.S.; van der Pol, C.B.; Larocque, N. Comparative performance of ChatGPT and bard in a text-based radiology knowledge assessment. *Can. Assoc. Radiol. J.* **2023**. [CrossRef] [PubMed]
11. Noda, R.; Izaki, Y.; Kitano, F.; Komatsu, J.; Ichikawa, D.; Shibagaki, Y. Performance of ChatGPT and Bard in Self-Assessment Questions for Nephrology Board Renewal. *medRxiv* **2023**. [CrossRef]
12. Teebagay, S.; Colwell, L.; Wood, E.; Yaghy, A.; Faustina, M. Improved Performance of ChatGPT-4 on the OKAP Examination: A Comparative Study with ChatGPT-3.5. *J. Acad. Ophthalmol.* **2023**, *15*, e184–e187. [CrossRef] [PubMed]
13. Thapa, D.K.; Visentin, D.C.; Kornhaber, R.; West, S.; Cleary, M. The influence of online health information on health decisions: A systematic review. *Patient Educ. Couns.* **2021**, *104*, 770–784. [CrossRef]
14. Calixte, R.; Rivera, A.; Oridota, O.; Beauchamp, W.; Camacho-Rivera, M. Social and demographic patterns of health-related Internet use among adults in the United States: A secondary data analysis of the health information national trends survey. *Int. J. Environ. Res. Public Health* **2020**, *17*, 6856. [CrossRef] [PubMed]
15. Hesse, B.W.; Nelson, D.E.; Kreps, G.L.; Croyle, R.T.; Arora, N.K.; Rimer, B.K.; Viswanath, K. Trust and sources of health information: The impact of the Internet and its implications for health care providers: Findings from the first Health Information National Trends Survey. *Arch. Intern. Med.* **2005**, *165*, 2618–2624. [CrossRef]
16. Fox, S.D. Maeve. In *Health Online 2013*; Pew Research Center: Washington, DC, USA, 2013.
17. Berkowitz, S.T.; Finn, A.P.; Parikh, R.; Kuriyan, A.E.; Patel, S. Ophthalmology Workforce Projections in the United States, 2020–2035. *Ophthalmology* **2023**, *131*, 133–139. [CrossRef]
18. Wedekind, L.; Sainani, K.; Pershing, S. Supply and perceived demand for teleophthalmology in triage and consultations in California emergency departments. *JAMA Ophthalmol.* **2016**, *134*, 537–543. [CrossRef]
19. Pushpanathan, K.; Lim, Z.W.; Yew, S.M.E.; Chen, D.Z.; Lin, H.A.H.E.; Goh, J.H.L.; Wong, W.M.; Wang, X.; Tan, M.C.J.; Koh, V.T.C. Popular Large Language Model Chatbots’ Accuracy, Comprehensiveness, and Self-Awareness in Answering Ocular Symptom Queries. *iScience* **2023**, *26*, 108163. [CrossRef]
20. Lim, Z.W.; Pushpanathan, K.; Yew, S.M.E.; Lai, Y.; Sun, C.-H.; Lam, J.S.H.; Chen, D.Z.; Goh, J.H.L.; Tan, M.C.J.; Sheng, B. Benchmarking large language models’ performances for myopia care: A comparative analysis of ChatGPT-3.5, ChatGPT-4.0, and Google Bard. *EBioMedicine* **2023**, *95*, 104770. [CrossRef]
21. Lyons, R.J.; Arepalli, S.R.; Fromal, O.; Choi, J.D.; Jain, N. Artificial intelligence chatbot performance in triage of ophthalmic conditions. *Can. J. Ophthalmol.* **2023**, in press. [CrossRef]
22. Deaner, J.D.; Amarasekera, D.C.; Ozzello, D.J.; Swaminathan, V.; Bonafede, L.; Meeker, A.R.; Zhang, Q.; Haller, J.A. Accuracy of referral and phone-triage diagnoses in an eye emergency department. *Ophthalmology* **2021**, *128*, 471–473. [CrossRef]
23. Azamfiri, R.; Kudchadkar, S.R.; Fackler, J. Large language models and the perils of their hallucinations. *Crit. Care* **2023**, *27*, 120. [CrossRef] [PubMed]

Disclaimer/Publisher’s Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.

Article

An Effective Methodology for Diabetes Prediction in the Case of Class Imbalance

Borislava Toleva ¹, Ivan Atanasov ¹, Ivan Ivanov ^{1,*} and Vincent Hooper ²

¹ Faculty of Economics and Business Administration, Sofia University, St. Kl. Ohridski, 1113 Sofia, Bulgaria; vrigazova@uni-sofia.bg (B.T.); iv.atanasov89@gmail.com (I.A.)

² SP Jain Global School of Management, Academic City, Dubai P.O. Box 502345, United Arab Emirates; vincent.hooper@spjain.org

* Correspondence: i_ivanov@feb.uni-sofia.bg

Abstract: Diabetes causes an increase in the level of blood sugar, which leads to damage to various parts of the human body. Diabetes data are used not only for providing a deeper understanding of the treatment mechanisms but also for predicting the probability that one might become sick. This paper proposes a novel methodology to perform classification in the case of heavy class imbalance, as observed in the PIMA diabetes dataset. The proposed methodology uses two novel steps, namely resampling and random shuffling prior to defining the classification model. The methodology is tested with two versions of cross validation that are appropriate in cases of class imbalance—k-fold cross validation and stratified k-fold cross validation. Our findings suggest that when having imbalanced data, shuffling the data randomly prior to a train/test split can help improve estimation metrics. Our methodology can outperform existing machine learning algorithms and complex deep learning models. Applying our proposed methodology is a simple and fast way to predict labels with class imbalance. It does not require additional techniques to balance classes. It does not involve preselecting important variables, which saves time and makes the model easy for analysis. This makes it an effective methodology for initial and further modeling of data with class imbalance. Moreover, our methodologies show how to increase the effectiveness of the machine learning models based on the standard approaches and make them more reliable.

Keywords: class imbalance; classification; cross validation; resample; shuffle

1. Introduction

The topic of diabetes disease prediction has been an extremely popular topic lately. Diabetes causes an increase in the level of blood sugar, which leads to damage to various parts of the human body. Many researchers collect medical data on the physiological, social, and environmental factors that would cause diabetes. These data are used not only for providing a deeper understanding of the treatment mechanisms but also for predicting the probability that one might become sick. This prediction is important for reversing the course of the possible development of diabetes by adjusting the related factors that impact the individual. Therefore, diabetes prediction can be crucial not only for decreasing the number of cases but also for developing a better clinical approach for the individual based on the specifics of their case.

The aim of the article is to propose a novel methodology for predicting whether an individual would develop diabetes over time given a set of biological and social indicators. The proposed algorithms create effective classification models to predict the risk of diabetes.

The proposed methodology works on public clinical data on diabetes in the women of the PIMA Indians dataset [1]. Our results outperform other existing research and extend the practical tools for researchers to provide a holistic approach for each patient. Based on the predictions, public healthcare specialists can create and implement specific strategies for the prevention and treatment of diabetes.

Predicting the risk of diabetes using novel computing techniques has been an expanding topic in academic literature with practical applications in medicine. Many of these techniques are centered around the application of Machine Learning (ML) models are the decision tree, support vector machines (SVM), Random Forest (RF) and Naive Bayes (NB) models [2]. For example, Traymbak et al. [3] used SVM to model the PIMA diabetes dataset and achieved a high classification accuracy of 73.86%, sensitivity of 83%, and specificity of 56.60%. In more advanced cases, deep learning techniques are applied to grasp the underlying complexity of the data and the connections among them [2]. The selection of the classification algorithm depends on the complexity of the data. The most common dataset for ML and deep learning experiments is the PIMA Indian diabetes dataset although other datasets exist as well [1]. An effective approach is to apply the ensemble and bagging methods for handling class imbalance in machine learning for healthcare datasets [4,5]. The high results emphasize the role of ensemble and bagging methods in enhancing model performance on classification models with imbalanced datasets.

For example, the RF classifier with feature selection has been applied to the PIMA diabetes dataset by Zou et al. [6]. They have concluded that the Random Forest model after the feature reduction achieves the best accuracy of 77.4% for the PIMA dataset. The same authors applied the RF model to another clinical dataset about diabetes—the Luzhou dataset. Random Forest confirms the efficiency with an accuracy value of 80.8% for the Luzhou dataset. Zhou et al. [7] have also worked on the PIMA dataset. They built an ensemble learning model based on the Boruta feature selection. The authors have applied the grid search approach to optimize the parameters of the proposed model. The accuracy they achieve is 98.0%. Machine learning models such as linear discriminant analysis (LDA), k-nearest neighbors (kNN), and Adaboost have also been applied for diabetes prediction by Traymbak et al. [3]. Different methodologies are commented on and compared with effective applications to healthcare modeling and prediction [8–11]. A detailed study on the application of SMOTE-based machine learning algorithms to predict diabetes can be found in [8,9]. In addition, Wu et al. [10] have applied machine learning modeling for imbalanced datasets based on the local interpretable model-agnostic explanation (LIME). In fact, LIME is applied to each sample independently and estimated. Then, conclusions are extended to the dataset.

Deep neural networks (DNN) are often used in complex cases. Deep neural networks are a deep learning algorithm. They combine the advantages of both deep learning and neural networks. Deep neural networks significantly increase the capabilities and quality of models applied in artificial intelligence, including diabetes prediction. Often deep learning algorithms outperform machine learning algorithms due to their flexibility and ability to capture and model more complex data structures including such in the PIMA diabetes dataset [2,12–16].

Regardless of the methodology selected, model estimation is a critical step in evaluating the performance of a classification model, especially when working with imbalanced data like the PIMA dataset. The standard tools for evaluating model performance are the confusion matrix and classification metrics like accuracy, precision, specificity (recall), sensitivity, and F1 score [17–19]. The confusion matrix provides insights into the extent to which the two classes are correctly predicted. The elements of the confusion matrix are denoted as true positive (TP), false positive (FP), true negative (TN), and false negative (FN).

The measures precision, specificity, and sensitivity can be calculated based on the elements of the confusion matrix as shown in [17–19]. The measures specificity and sensitivity provide information on how the model predicts both classes of the dataset. The high values of these measures confirm the efficiency of the model. The accuracy provides an insight into the overall performance of the classification models. In this study, we adhere to the standard formulas for classification evaluation [17–19].

Using these measures, we compare our methodology to existing ones. We show that our methodology may outperform other existing machine learning algorithms, while producing competitive results to deep learning algorithms. The advantages are simplicity, easy interpretation of the results, and improved model performance.

Next section describes the proposed methodology on the PIMA Indian diabetes dataset. The results section details the results of our experiments in terms of model quality and considering other papers on this dataset, while Section 4 concludes.

2. Materials and Methods

Our methodology is applied on the public PIMA Indian diabetes dataset from [1,20]. The dataset contains 768 observations and 9 variables for female patients from Arizona, USA. The dataset consists of nine medical variables (predictors) and one target variable. The target variable for the dataset represents 268 observations that are positives for diabetes. They are denoted by value ‘1’ whereas value ‘0’ is used for negative results for diabetes observations. The number of negative observations is 500. This dataset is representative of the class imbalance problem in the ML theory and practice [21–23]. Class imbalance is an issue in classification problems where the target variable has one class dominating over the other [23]. The structure of the PIMA dataset demonstrates heavy class imbalance as the label ‘0’ is the predominant class, accounting for about 2/3s of the observations in the target variable. Therefore, the prediction of diabetes in the PIMA dataset is sensitive to class imbalance, which needs to be handled using appropriate tools.

This paper presents a novel methodology for handling the class imbalance issue in the PIMA dataset. To interpret the results from the novel methodology, we also run experiments with the classical methodology. The first methodology models the original observations without preprocessing them for class imbalance. Class imbalance is handled using the built-in parameter in Python called “class_weight” = balanced. This is a standard tool to handle class imbalance without preprocessing the data. This is the classical approach. The second methodology introduces novel steps for data preprocessing to handle the class imbalance in the target variable before applying a classification model. The novelty in our methodology lies in resampling and shuffling the data as a preprocessing step prior to cross validation and model fitting. The classical methodology is tested with k-fold cross validation, while the proposed methodology is run by k-fold and stratified k-fold cross validation.

K-fold cross validation is the most often used validation strategy. However, it is suitable for large datasets with no class imbalance. Some authors [16,18] argue that k-fold cross validation can be appropriate for datasets with class imbalance only if the dataset is large enough. The definition of ‘large enough’ is not provided but the PIMA dataset has less than 1000 observations, so it may be considered a ‘small’ dataset. As the definitions of ‘small’ and ‘large’ datasets are not clear, we test the proposed methodology with both k-fold cross validation and stratified k-fold cross validation. The aim is to understand which cross validation strategy is better for the PIMA dataset.

Another advantage of the proposed methodology is the unique combination of (1) re-sampling and shuffling, (2) setting ‘class_weight’ = ‘balanced’, and (3) using (stratified)

k-fold cross validation as simple steps to handle the class imbalance issue without further complicating the classification model.

Figure 1 demonstrates the difference between the classical methodology and the proposed methodology.

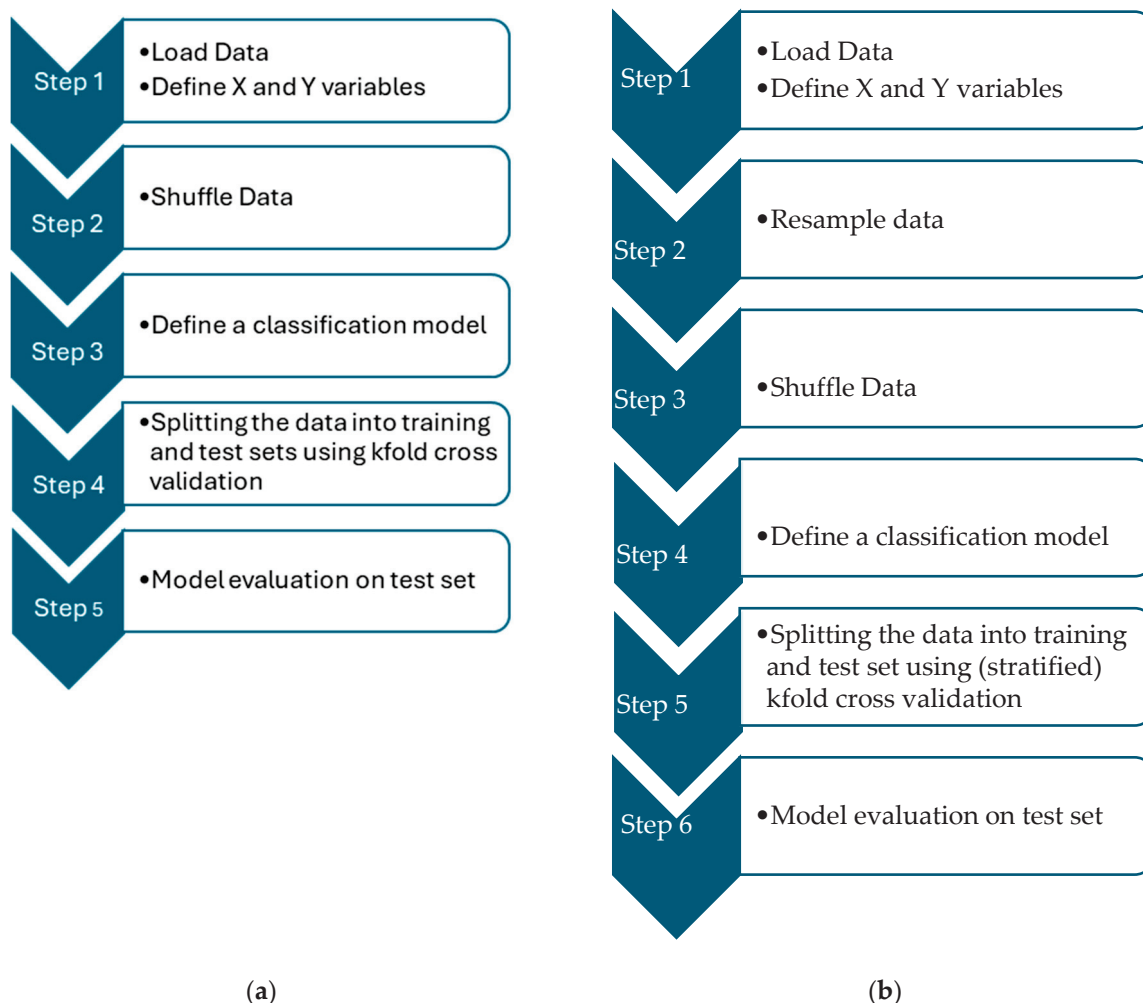


Figure 1. (a) left image summarizes Methodology 1, and (b) right image summarizes Methodology 2.

2.1. Methodology 1: Algorithms 1–3—Classical Approach

Methodology 1 is presented via Algorithms 1, 2 and 3 in the investigation. The approach of Algorithms 1–3 is as follows:

Step 1: Data loading and initial processing—the initial step involves loading the dataset and delineating the independent (X) and dependent (y) variables. The y variable is transformed into categories to facilitate analysis.

Step 2: Data shuffling—the indices of all independent variables (X) and the target variable (y) have been shuffled so that their place in the dataset is rearranged. To perform the shuffle a seed of 99 is set and the numpy command `np.random.permutations` is used as shown below:

```
np.random.seed(99)
permuted_indices = np.random.permutation(len(Y))
```

Random shuffling in train-test splitting is employed to ensure that the training and testing datasets are representative of the overall dataset [24]. By shuffling, the data is randomized, preventing the model from learning potential patterns that may be due to the order of the data rather than the underlying relationships between the variables.

Step 3: Define two classification models—the Random Forest classifier (RF) and the support vector machines model (SVM) with parameters:

RandomForestClassifier(n_estimators = 50, max_depth = 5, random_state = 0, class_weight = 'balanced')

SVC (C = 10, kernel = 'linear', gamma = 0.01, probability = True, class_weight = "balanced").

The parameter 'class_weight' is set to 'balanced' because of the class imbalance in y.

In this step the parameters as well as the type of model can be changed.

Step 4: Split the data into training and test sets using k-fold cross validation—the k-fold cross validation function is applied using different commands for classification models:

RF: KFold (n_splits = 4, shuffle = True, random_state = 555) (Algorithm 1),

SVM: KFold (n_splits = 4, shuffle = True, random_state = 763) (Algorithm 2),

SVM: KFold (n_splits = 5, shuffle = True, random_state = 673) (Algorithm 3).

Step 5. Model evaluation on the test set. The confusion matrix-model evaluation is performed using the confusion matrix and calculating accuracy, precision, specificity, and sensitivity using the formulae [17–19]:

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN};$$

$$\text{Precision} = \frac{TP}{TP + FP};$$

$$\text{Sensitivity (Recall)} = \frac{TP}{TP + FN};$$

$$\text{Specificity} = \frac{TN}{TN + FP}$$

Measures like specificity and sensitivity provide information on how the model predicts both classes of the dataset. The high values of these measures confirm the efficiency of the model.

2.2. Proposed Methodology: Algorithms 4 and 5—Classification with Data Preprocessing for Class Imbalance

This methodology contains two essential steps: (a) it applies two approaches to split training and test subsets, and (b) it applies the support vector machines model. Algorithm 4 applies k-fold cross validation to separate training and test subsets, while Algorithm 5 applies stratified k-fold cross validation for the same purpose. The approaches to Algorithms 4 and 5 are as follows:

Step 1: Data loading and initial processing—the initial step involves loading the dataset and defining the independent (X) and dependent (y) variables. The y variable is transformed into categorical labels to facilitate analysis.

Step 2: Data resampling—the second methodology applies the resampling procedure to supplement the smaller class in the set with observations. The aim is to increase their number to avoid inequality between the two classes in terms of the number of observations. The resampling function is applied with the parameters shown below:

resample(data2, replace = True, n_samples = 500, random_state = 605)

Step 3: Data shuffling—the indices of all independent variables (X) and the target variable (y) are shuffled so that their place in the dataset is rearranged. To perform the shuffle a seed of 31 is set and the numpy command np.random.permutations is used as shown below:

np.random.seed (31)

permuted_indices = np.random.permutation (len(Y))

Step 4: Define a classification model—a support vector machines classifier is defined with the parameter ‘class_weight’ set to ‘balanced’ because of the class imbalance in y. The settings of the two classifications are shown below:

SVC (C = 10, kernel = ‘rbf’, gamma = ‘auto’, probability = True, class_weight = “balanced”)

In this step, the parameters as well as the type of model can be changed.

Step 5: Split the data into training and test sets using k-fold cross validation—the k-fold cross validation function is applied using different commands for classification models:

SVM: KFold (n_splits = 5, shuffle = True, random_state = 42) (Algorithm 4),

SVM: KFold (n_splits = 5, shuffle = True, random_state = 73) (Algorithm 5).

Step 6: The same as Step 5 from Methodology 1.

The next section describes the output from the proposed algorithms and compares the results to others.

3. Results

3.1. Comparison of Classification Metrics

Our experiments are conducted on a laptop with 1.50 GHz Intel(R) Core (TM) and 8 GB RAM, running on Windows with Python 3.7 in the Anaconda environment. The results discuss two groups of methodologies. The first methodology is represented by Algorithms 1–3, where only a shuffling method is applied at the preprocessing stage. While the second methodology is represented by Algorithms 4 and 5 and contains a shuffling and a resampling method. To consider the output from each methodology effective, the values of accuracy, precision, sensitivity, and specificity should be high enough. The results for the classes can be averaged, so we present the average precision, sensitivity, and specificity for each algorithm.

As shown by Table 1, the first methodology results in accuracies of 83.85% (Algorithm 1), 84.9% (Algorithm 2), and 85.06% (Algorithm 3). The accuracies from the second methodology are much higher as Table 2 shows. The two Algorithms (4 and 5) in the second methodology result in accuracies of 95.5% and 90.5%. Algorithms 1–3 are considered a classical approach that underperforms when compared to the second methodology, which is a novel approach.

Table 1. Results from methodology 1. Authors’ calculations.

Algorithm 1 (%) Random Forest	Algorithm 2 (%) SVM, KFOLD(N_SPLITS = 4)	Algorithm 3 (%) SVM, KFOLD(N_SPLITS = 5)
Accuracy = 83.85	Accuracy = 84.90	Accuracy = 85.06
Precision = 90.82	Precision = 92.56	Precision = 90.91
Sensitivity = 82.50	Sensitivity = 84.85	Sensitivity = 86.54
Specificity = 86.11	Specificity = 85.00	Specificity = 82.00

Table 2. Results from the proposed Methodology 2. Authors’ calculations.

Algorithm 4 (%)	Algorithm 5 (%)
Accuracy = 95.5	Accuracy = 95.05
Precision = 91.35	Precision = 91.74
Sensitivity = 100.00	Sensitivity = 100.00
Specificity = 91.35	Specificity = 91.00

This finding is also visible from the precision, sensitivity, and specificity of the first methodology (Algorithms 1–3) compared to Algorithms 4 and 5. The proposed novel methodology outperforms the classical methodology when the metrics are averaged

(Tables 1 and 2). The precision for the two methodologies is similar—it varies between 90.82% and 92.56%. The high value for all algorithms shows that all of them predict correctly the positive class in more than 90% of the cases. However, methodology 1 has sensitivity scores lower than Methodology 2. The second methodology has a sensitivity of 100%, while methodology 1 achieves the highest sensitivity scores (86.54%) via Algorithm 3. As Tables 1 and 2 show, the second methodology improved the sensitivity scores by more than 13 p.p., which is a significant improvement. A high sensitivity score shows that the model classifies correctly the observations in each class. Therefore, the proposed methodology predicts whether the patient has diabetes or not better than the classical methodology.

The second methodology also results in better specificity. Specificity demonstrates the model's ability to correctly classify all patients that do not have diabetes. The highest score for specificity in the proposed methodology is 91.35% (Algorithm 4), whereas Algorithms 1–3 exhibit specificity scores between 82% and 86.5%. The proposed methodology improved the model's ability to predict correctly the cases of healthy patients. The classical methodology correctly predicted healthy patients in 82% to 86.5% of the cases. While the proposed methodology captures healthy patients in more than 95% of cases.

The significant improvements in the model's ability to predict correctly healthy patients and to identify sick patients lead to improved overall accuracy of Algorithms 4 and 5. Therefore, the finding that the proposed methodology outperforms the classical one is a result of the overall improvement of the model's performance. The overall improvement of the model's performance in the proposed methodology can be attributed to two factors.

The first one is handling the class imbalance issue. The classical methodology does not perform data preprocessing aimed at class imbalance. It tries to handle this issue by only setting the Python parameter 'class_weight' to 'balanced'. Although this approach provides good results, our methodology offers an effective solution to the class imbalance that significantly improves the classification ability of the model. The novel steps to shuffle and resample data prior to setting the parameter 'class_weight' = 'balanced' helps to get a more even distribution of the two classes so that the training/test split is undertaken in an unbiased way where the predominant class does not affect the split. Therefore, handling the class imbalance issue proves to be a key part in the quality of the model.

The second factor is the type of cross validation. As mentioned in the methods section, some authors recommend using stratified k-fold cross validation in relatively small datasets with class imbalance [23,25]. Although no explicit definition of a 'small' and 'large' dataset is given, we tested the proposed methodology with k-fold (Algorithm 4) and stratified k-fold (Algorithm 5) cross validation. As Table 2 shows, the accuracy and the rest of classification metrics are very similar in the two cases. One key finding is that the two types of cross validation can be used with this dataset, which is further proof for the lack of overfitting in our algorithms.

Another key finding is that the suitability of stratified and k-fold cross validation in target variables with class imbalance may not be defined by the size of the dataset rather than the characteristics of the data. This finding may be further explored in additional research. A third finding is that the methodology we propose is robust in terms of steps to handle the class imbalance issue. The proposed steps to handle the class imbalance issue involve shuffling and resampling data and setting the parameter 'class_weight' = 'balanced'. We tested these steps with k-fold and stratified k-fold cross validation and the results were similar (Table 2). This means that the novel steps we propose to handle class imbalance in the PIMA dataset are robust and are not affected by the strategy chosen for training and testing the model.

The overall quality of the models in the proposed methodology is improved and that is also confirmed by looking at the prediction for individual classes as described by confusion matrices and AUC–ROC curves.

3.2. Confusion Matrices and AUC–ROC Curves

The ROC–AUC (Receiver Operating Characteristic—Area Under the Curve) is another key metric for evaluating the performance of binary classification models. The ROC curve plots the true positive rate (TPR) against the false positive rate (FPR) at various threshold settings, while the AUC quantifies the overall ability of the model to distinguish between classes. Unlike accuracy, ROC–AUC is threshold-independent and remains informative even with imbalanced datasets, making it a robust tool for model comparison and selection.

Tables 3–5 show the AUC–ROC curves and confusion matrices for Algorithms 1–3, which confirm that the classical methodology performs well. High values for accuracy, precision, sensitivity, and specificity are the first indicator that the classical methodology results in correct predictions. The confusion matrix along with the AUC–ROC curve demonstrates the reliability of the standard methodology as they also confirm the quality of the model.

Table 3. Results from Methodology 1. Random Forest, (Algorithm 1). Authors’ calculations.

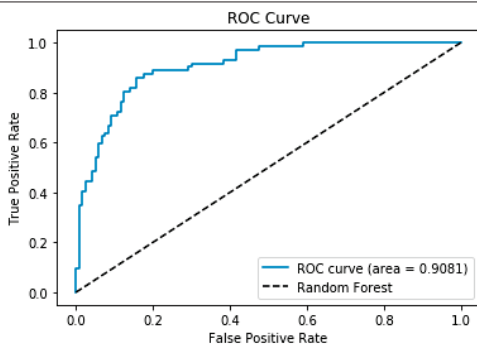
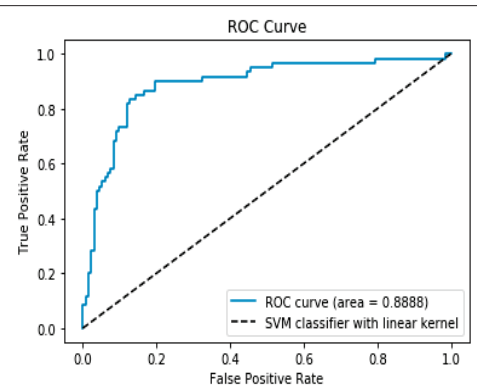
Confusion Matrix	ROC Curve
$\begin{bmatrix} 100 & 20 \\ 10 & 62 \end{bmatrix}$ <p>ROC curve AUC= 90.81%</p>	

Table 4. Results from Methodology 1. SVM and kFold (n_splits = 4), (Algorithm 2). Authors’ calculations.

Confusion Matrix	ROC Curve
$\begin{bmatrix} 112 & 20 \\ 9 & 51 \end{bmatrix}$ <p>ROC curve AUC= 88.88%</p>	

However, the data suffers from class imbalance. All tables related to methodology 1 (1, 3–5) show that the classical methodology performs well despite the class imbalance. But as shown in Tables 2, 6 and 7 the model can perform better when the class imbalance is handled appropriately. Therefore, handling class imbalance using the proposed methodology is a better approach in the PIMA dataset.

Table 5. Results from Methodology 1. SVM and kFold (n_splits = 5), (Algorithm 3). Authors' calculations.

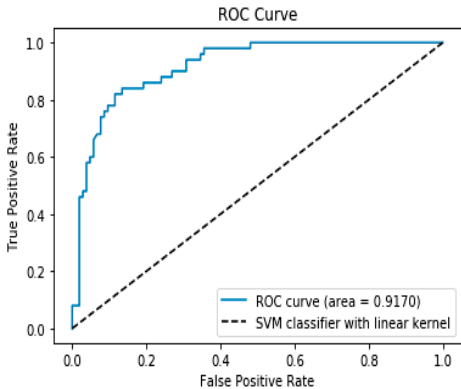
Confusion Matrix	ROC Curve
$\begin{bmatrix} 90 & 14 \\ 9 & 41 \end{bmatrix}$ <p>ROC curve AUC= 91.70%</p>	

Table 6. Results from Methodology 2. Algorithm 4. Authors' calculations.

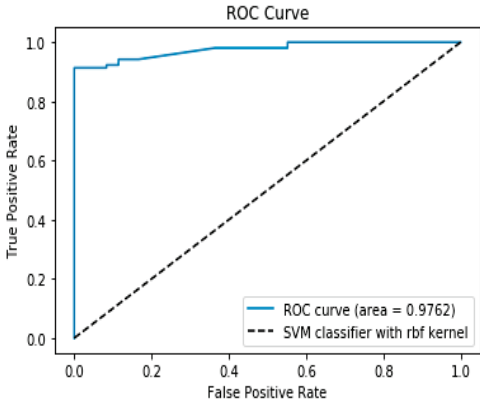
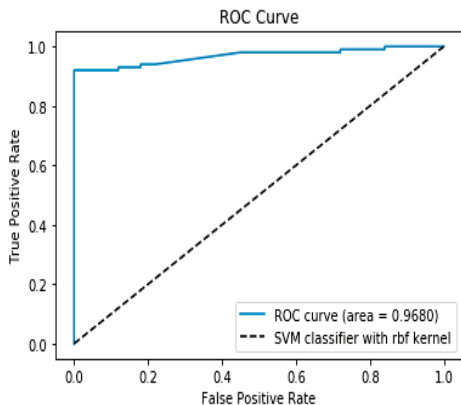
Confusion Matrix	ROC Curve
$\begin{bmatrix} 96 & 0 \\ 9 & 95 \end{bmatrix}$ <p>ROC curve AUC= 97.62%</p>	

Table 7. Results from Methodology 2. Algorithm 5. Authors' calculations.

Confusion Matrix	ROC Curve
$\begin{bmatrix} 100 & 0 \\ 9 & 91 \end{bmatrix}$ <p>ROC curve AUC= 96.80%</p>	

The proposed methodology classifies the minority class more accurately, while keeping the same prediction rate for the majority class as seen by the confusion matrices of Algorithms 4 and 5 (Tables 6 and 7). Tables 6 and 7 show that Algorithms 4 and 5 provide the best prediction for the two classes, although the results for individual classes are close to those from Algorithms 1 to 3. Therefore, the second methodology results in more accurate

predictions of the two classes, especially the minority class. This finding is key as heavy class imbalance usually leads to better prediction of the majority class and poor prediction of the minor class. As Tables 6 and 7 show, the proposed two Algorithms (4 and 5) overcome the issue of class imbalance effectively and improve the overall quality of the model compared to the classical methodology.

3.3. Comparison to Other Research

As Table 2 shows, Algorithm 4 achieves an accuracy of 95.50%, an overall precision of 91.43%, an overall sensitivity of 91.35%, and an overall specificity of 100.00%. Similarly, Gupta and Goel [18] have applied several machine learning models. Their best results are obtained from the Random Forest model. The estimated parameters have the following values: the accuracy is 80.52%, the precision value is 74.47%, the sensitivity value is 72.72%, and the specificity is 90.74%. The output from the algorithms we propose outperforms Gupta's results as seen in Tables 2 and 8.

Table 8. Results obtained by Gupta et al. [18], Chang et al. [19], and Tigga et al. [21].

Table 6. [18] (%)	Table 3. [21] (%)	Table 13. [19] (%)
Accuracy = 80.52	Accuracy = 75.0	Accuracy = 79.57
Precision = 74.47	Precision = 84.0	Precision = 89.40
Sensitivity = 72.72	Sensitivity = 78.95	Sensitivity = 81.33
Specificity = 90.74	Specificity = 66.10	Specificity = 75.0
		ROC-AUC = 86.24

Table 8 also shows Tigga's [21] and Chang's [19] results on the PIMA dataset. Tigga's [21] article does not apply techniques to handle class imbalance. Instead, they aim to find the most appropriate machine learning algorithm that can predict the patient's condition correctly. Their findings suggest that the most appropriate machine learning algorithm for the PIMA dataset is the Random Forest classifier. Table 8 presents their results from the Random Forest. Compared to Gupta [18] and Chang [19], their results are worse in terms of accuracy and specificity. The second methodology we propose (Algorithms 4 and 5) outperforms Tigga's results. A key finding in this case is that using a classifier like the Random Forest, which is known to work good with class imbalance, may not be enough to handle the bias coming from imbalanced classes

Table 8 also shows a similar case with the results of Chang [19]. They also aim to find the most appropriate machine learning model to accurately predict the PIMA dataset in the context of feature selection. They conclude that a naïve Bayes model works well when features are carefully selected, while Random Forest works better when more features are added. Although both Chang [19] and Tigga [21] conclude that the Random Forest is the most appropriate model for the PIMA dataset, Tables 2 and 8 show that Algorithms 4 and 5, proposed in this article, outperform the Random Forest. As shown by Tables 2 and 8, the second methodology outperforms all models used by Tigga [21] and Chang [19] (this paper presents their best results). This finding is also valid for Gupta's [18] experiments.

The advantage of our methodology is its simplicity and fast calculation. Also, our methodology can be applied to larger or smaller datasets related to PIMA or any other diabetes dataset as we test two versions of cross validation. The k-fold cross validation, that is usually applied even in class imbalance datasets with the remark that the dataset should be large enough. We also provide a version of our methodology (Algorithm 5) with stratified k-fold cross validation, which is recommended for class imbalance in smaller datasets. Although in the case of the PIMA dataset, the type of cross validation is not the

key for improved model performance, we propose two alternatives that can be the effective solution to class imbalance in other diabetes datasets with class imbalance.

Also, adding two simple steps (shuffling and resampling) before data preprocessing has proven to be an effective way to handle class imbalance as our results suggest. This finding extends the applications of shuffling and resampling in machine learning, which uncovers a new path for the discovery of new potential applications for the two of them.

We compare the results from Table 9 (based on the values of Table 7 [22]) to those obtained by Methodology 2. Table 10 presents the entries of confusion matrices when methodology 2 is applied. We calculate the percentage of the sum (FN+FP). The values are displayed in the last column of Tables 9 and 10. Thus, the values of Table 10 are smaller than the corresponding ones in Table 9. Thus, methodology 2 minimizes the sum of false cases (FN+FP).

Table 9. Entries of confusion matrices were obtained by Ejiyi and coauthors [22].

Models	TP	FN	FP	TN	Total	(FN + FP)/Total (%)
Extra Tree	141	12	18	129	300	0.1
RF	142	11	11	136	300	0.073
AdaBoost	142	11	5	142	300	0.053
GB	143	12	4	141	300	0.053

Table 10. Entries of confusion matrices after methodology 2. Our calculations.

Models	TP	FN	FP	TN	Total	(FN + FP)/Total (%)
Algorithm 4	95	9	0	96	200	0.045
Algorithm 5	91	9	0	100	200	0.045

Further on, applying the entries of confusion matrices the values of parameters accuracy, precision, sensitivity, and specificity are computed and presented in Tables 2 and 11.

Table 11. Results were obtained according to Table 6 [22].

Models	Accuracy (%)	Precision (%)	Sensitivity (%)	Specificity (%)
Extra Tree	90.0	88.68	92.20	87.76
RF	92.67	92.81	92.80	92.52
AdaBoost	94.67	96.60	92.80	96.60
GB	94.67	97.28	92.30	97.24

Ejiyi and coauthors [22] also experimented with the PIMA dataset. They aim to propose a robust methodology for diabetes prediction. Their approach is different from [18,19,21] as they perform feature extraction using the Shapley Additive Explanation (SHAP). Then, they use the subset of the most important features to find the most appropriate machine learning model. They conclude that Xgboost and Adaboost perform best with SHAP. Table 11 shows their results. Their results are better than [18,19,21] as seen in Table 8. However, this is not the case when comparing the values from Tables 2 and 11. The two tables demonstrate that methodology 2 achieves higher values of the metrics of accuracy, precision, and specificity than those obtained by Ejiyi and coauthors [22]. The measure sensitivity is 91.30%, which is slightly smaller than the corresponding value of 92.80% from Table 11. This point is another important finding as it demonstrates that combining complex techniques like SHAP, Xgboost, and Adaboost may handle class imbalance well, but a

simpler and more effective methodology may still exist. The second proposed methodology is an example of a simpler and more effective methodology to predict accurately the target in the PIMA dataset, while outperforming a wide variety of existing machine learning algorithms and complex methodologies

Another advantage of the proposed methodology is that the performance is close to that of deep neural networks (DNN), which are much more flexible and reliable in capturing data anomalies. Yet, they are more complicated to use. As Table 8 shows, the accuracy of the proposed Algorithms 4 and 5 is better than that of [2,13]. The highest accuracy we achieved (Table 2) was 95.59%, while DNN achieved 98.07% [13,15,16]. The DNN accuracy is bigger than ours by about 2.5 percentage points, which is a small difference. In terms of sensitivity, the proposed methodology achieves a result of 100%, which outperforms the DNN models in Tables 9–11. For specificity, we achieve similar results to Hounguè [2].

However, classical models (Table 1) fail to handle class imbalance in a way that the prediction ability of the models does not suffer. Therefore, the results in Table 1 are not close to the results of other authors (Table 12) and the proposed methodology (Table 2). This is another evidence that handling class imbalance is a key point in modeling the PIMA dataset.

Table 12. Other authors obtained results via deep learning models.

Models	Measures
DNN [12]	Accuracy = 94.39%
DNN [13]	Accuracy: 98.04% Sensitivity: 98.80%. Specificity: 96.64%
DNN [14]	Accuracy: 99.4%
DNN + DT [15]	Accuracy: 98.07% Sensitivity: 95.52% Specificity: 99.29%
DNN + 10-fold cross-validation [2]	Accuracy: 89% Sensitivity: 87% Specificity: 91%
DNN [16]	Accuracy: 98.07%

As seen in Tables 2 and 12, other authors achieve similar results to the methodology we propose, or in some cases, slightly better results. These results are not surprising as deep learning models also capture the unstructured connections in the data and use them to train and test the model. However, machine learning models derive the boundaries between classes based on the distance among observations. Machine learning models are better at modeling structured datasets, while deep learning models capture hidden complex connections in the data.

Also, machine learning models require more involvement of the research in model tuning, while deep learning algorithms reduce the involvement of the researcher in model tuning. Therefore, deep learning models may reduce the bias coming from the researcher's experience and knowledge when tuning the model. Deep learning models are often a good alternative to machine learning models, especially in complex datasets. Deep learning models usually perform better than machine learning models. That is why the results from the proposed methodology are remarkable.

On the one hand, the machine learning methodology we propose improves the classification ability of the model so that the results are close to complex deep learning models like neural networks. On the other hand, we achieved a specificity of 100%, a result that was not achieved by any of the deep learning models in Table 12. The proposed methodology (Algorithms 4 and 5) manages to correctly identify non-sick patients in 100% of the cases. The methodology we propose is also notable as it represents a significant improvement in the prediction ability of machine learning algorithms that become comparable to that of deep learning models. The key to this result is the steps we propose to handle class imbalance.

Another important fact is that the authors in Table 12 do not provide their confusion matrices to observe the performance of the individual classes. Confusion matrices are an essential part of the analysis of the model quality. They show whether each class is predicted correctly and to what extent. A sign for a good model is not only high classification metrics (accuracy, precision, sensitivity, and specificity) but also the correct prediction of each class. Cases when the classification metrics are high, but the confusion matrices show that one of the classes is not predicted correctly may be a sign of overfitting. Therefore, we compare our results to other authors based on the classification metrics, but the comparison of the quality of prediction for each class cannot be done thoroughly. Despite this, we achieve similar results to Tables 8 and 12 with a much simpler methodology that is not computationally exhaustive and does not require a complex hardware setup to run. This finding presents another advantage of the proposed methodology.

Based on the results above, the key finding is that the proposed methodology is competitive with other existing machine learning and deep learning algorithms. In some cases, the proposed methodology can outperform existing ML algorithms, while having similar performance to deep learning algorithms. Resampling and shuffling the data prior to defining the classification model can improve the prediction ability of the classification in case of class imbalance. The proposed methodology can be used with either k-fold cross-validation or stratified cross-validation. In the two cases, the results are similar, which validates the importance of the two novel steps in improving the predictions for class imbalance. This result also validates the assumption made by other authors [25–28] that both k-fold and stratified k-fold cross-validation can be used in class imbalance classification. Our results also suggest that the size of the dataset in the case of class imbalance may not be the key factor determining the type of cross-validation, which opens a new field for exploration of the role of cross-validation in the class imbalance issue. The proposed methodology improves not only the classification ability of the model but the ability of the model to improve the prediction for each class. This makes it competitive with deep learning models without further complicating the applications and interpretation of the model. Therefore, we consider the proposed methodology effective and efficient in handling class the class imbalance issue for the PIMA diabetes data.

A future extension of this work would be testing the proposed methodology on larger datasets with class imbalance as well as multilabel classification. Also, the proposed methodology can be tested in datasets with many features without feature selection. We consider our methodology flexible and able to adapt to the characteristics of the data as model tuning can easily be performed. Therefore, the proposed methodology can easily be adapted to other datasets and multiclass classification.

Although some recent diabetes research focusing on various aspects of diabetes prediction exists [29–32] they use different datasets. This makes their authors' results incomparable to ours. However, in the publication of Mohanty [33] the same dataset as ours is analyzed. The authors have studied a few machine learning models and ensemble models for classification analysis of the dataset. The proposed ensemble model in the paper [33] has

achieved the following values: accuracy 84%, precision 80%, sensitivity 92%, and specificity 77% (see Table 19 [33]). Our findings with the SVM model exceed these values (see Table 2).

4. Discussion

In this paper we present a simple and not computationally exhaustive methodology for improving the prediction ability of classification models for the minority class in the case of class imbalance. We can summarize the key findings from this research as follows:

1. Using Python's built-in functions can successfully tackle class imbalance. The used parameter in question is 'class_weight' = balance. However, additional steps need to be taken to handle class imbalance so that the model can become more efficient.
2. When having imbalanced data, resampling and shuffling the data randomly before the train/test split can help improve estimation metrics. This result is robust regardless of the type of cross validation used.
3. Applying our proposed algorithm is a simple and fast way to predict labels with class imbalance. It does not require additional techniques to balance classes. It does not involve preselecting important variables, which saves time and makes the model easy for analysis. This makes it an effective algorithm for the initial and further modeling of data with heavy class imbalance.
4. Our algorithm does not need a feature selection procedure, therefore avoiding the bias that can be introduced with the method of feature selection.
5. Two types of cross validation can be used as shown. The results are similar, suggesting that the type of cross validation may not be key for class imbalance. Rather, the overall strategy to eliminate the influence of the dominant class may be more important.
6. Despite the relatively small size of the PIMA datasets, both k-fold and stratified k-fold cross validation are appropriate. This finding contrasts with some researchers suggesting using k-fold cross validation for class imbalance only in large datasets. We highlight that the type of cross validation used may not be dependent on the size of the dataset rather than its characteristics.
7. This property of the model makes it flexible to adjust to other issues in the data, not only class imbalance. Therefore, other types of cross validation can be used.

As a conclusion, testing the proposed methodology on the PIMA diabetes dataset has shed light on the importance of resampling and shuffling data as a data processing step for handling class imbalance. With this discovery, we extend the applications of resampling and shuffling data and introduce a new line of academic research in the field of class imbalance. Moreover, our methodologies show how to increase the effectiveness of the machine learning models based on the standard approaches and make them more reliable.

Currently, the use of artificial intelligence (AI) applications in healthcare is growing rapidly. AI implementation allows healthcare professionals to devote more attention to patient care prescreening and individual treatment. As a result, the quality of life and rates of recovery increase. Therefore, the issue of the legal and ethical consequences of adopting AI technology in medicine is relevant. For example, Geantă et. al. [34] compared the efficiency of artificial intelligence models in healthcare to answer the question of which platform is better equipped to produce health-safe diagnostic models. They discovered that AI models generated by ChatGPT 3.5 in some healthcare fields may be more trustworthy and secure in diagnosing some diseases than AI models pre-defined by researchers. However, other diseases are diagnosed and treated more accurately using AI models defined by the researcher. Therefore, an ethical issue arises to discover and make sure that the safest AI algorithm for the detection of the disease is used. This results in public trust [34–36] in the healthcare system. However, the questions of what AI models are safe and ethical to use, how hospitals can better protect the personal data they collect as

a result of more sophisticated AI models, and how can AI models be tested in reality are posed. Such scientific research supports the public debate on the use of AI in healthcare and seeks to produce new evidence in building a coherent ethical framework for AI medical technologies [36]. Therefore, proof of the ethical, clinically effective, and safe applications of each AI model should be provided before using them for diagnosis, treatment, and pre-screening [36].

Further on, in our future construction of machine learning classification methodologies for medical datasets, we will expand our research in the directions: (a) we will apply nested stratified k-fold cross-validation, (b) we will introduce additional evaluation metrics like the Matthews Correlation Coefficient, and (c) we will use the confidence intervals for applied metrics.

Author Contributions: Conceptualization, B.T., I.I. and V.H.; methodology, B.T. and I.I.; software, B.T. and I.A.; validation, B.T., I.A. and I.I.; formal analysis, B.T.; investigation, B.T. and V.H.; resources, B.T. and I.A.; data curation, B.T. and I.A.; writing—original draft preparation, B.T. and V.H.; writing—review and editing, B.T. and V.H.; visualization, B.T.; supervision, I.I.; project administration, I.I. and V.H.; funding acquisition, V.H. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: This study was conducted using publicly available data, which does not contain identifiable personal information and therefore did not require ethical approval. The data sources are described in the Data Availability Statement section. All analyses were performed in accordance with relevant ethical guidelines and regulations.

Informed Consent Statement: Informed consent was obtained from all subjects involved in the study.

Data Availability Statement: The data presented in this study are openly available here <https://www.kaggle.com/uciml/pima-indians-diabetes-database> (accessed on 7 October 2024) and here <https://www.openml.org/search?type=data&status=active&id=43582&sort=runs> (accessed on 30 December 2024).

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Kaggle. Pima Indians Diabetes Database. Available online: <https://www.kaggle.com/uciml/pima-indians-diabetes-database> (accessed on 30 June 2024).
2. Houngue, P.; Bigirimana, A. Leveraging Pima Dataset to Diabetes Prediction: Case Study of Deep Neural Network. *J. Comput. Commun.* **2022**, *10*, 15–28. [CrossRef]
3. Traymbak, S.; Issar, N. Data Mining Algorithms in Knowledge Management for Predicting Diabetes After Pregnancy by Using R. *Indian J. Comput. Sci. Eng.* **2021**, *12*, 1542–1558. [CrossRef]
4. Gurcan, F.; Soylu, A. Learning from Imbalanced Data: Integration of Advanced Resampling Techniques and Machine Learning Models for Enhanced Cancer Diagnosis and Prognosis. *Cancers* **2024**, *16*, 3417. [CrossRef] [PubMed]
5. John, A.; Isnin, I.F.B.; Madni, S.H.H.; Muchtar, F.B. Enhanced intrusion detection model based on principal component analysis and variable ensemble machine learning algorithm. *Intell. Syst. Appl.* **2024**, *24*, 200442. [CrossRef]
6. Zou, Q.; Qu, K.; Luo, Y.; Yin, D.; Ju, Y.; Tang, H. Predicting diabetes mellitus with machine learning techniques. *Front. Genet.* **2018**, *9*, 515. [CrossRef]
7. Zhou, H.; Xin, Y.; Li, S. A diabetes prediction model based on Boruta feature selection and ensemble learning. *BMC Bioinform.* **2023**, *24*, 224. [CrossRef]
8. Alghamdi, M.; Al-Mallah, M.; Keteyian, S.; Brawner, C.; Ehrman, J.; Sakr, S. Predicting diabetes mellitus using SMOTE and ensemble machine learning approach: The Henry Ford Exercise Testing (FIT) project. *PLoS ONE* **2017**, *12*, e0179805. [CrossRef] [PubMed]
9. Rezki, M.K.; Mazdadi, M.I.; Indriani, F.; Muliadi, M.; Saragih, T.H.; Athavale, V.A. Application of SMOTE to address class imbalance in diabetes disease classification utilizing C5.0, Random Forest, and SVM. *J. Electron. Electromed. Eng. Med. Inform.* **2024**, *6*, 343–354. [CrossRef]

10. Wu, Y.; Zhang, L.; Bhatti, U.A.; Huang, M. Interpretable Machine Learning for Personalized Medical Recommendations: A LIME-Based Approach. *Diagnostics* **2023**, *13*, 2681. [CrossRef] [PubMed]
11. Kitova, K.; Ivanov, I.; Hooper, V. Stroke Dataset Modeling: Comparative Study of Machine Learning Classification Methods. *Algorithms* **2024**, *17*, 571. [CrossRef]
12. Mhaskar, H.N.; Pereverzyev, S.V.; Van der Walt, M.D. A Deep Learning Approach to Diabetic Blood Glucose Prediction. *Front. Appl. Math. Stat.* **2017**, *3*, 14. [CrossRef]
13. Islam, I.A.; Milon, M.I. Diabetes Prediction: A Deep Learning Approach. *Int. J. Inf. Eng. Electron. Bus.* **2019**, *11*, 21–27. [CrossRef]
14. Zhou, H.; Myrzashova, R.; Zheng, R. Diabetes Prediction Model Based on an Enhanced Deep Neural Network. *EURASIP J. Wirel. Commun. Netw.* **2020**, *2020*, 148. [CrossRef]
15. Pham, T.; Tran, T.; Phung, D.; Venkatesh, S. Predicting Healthcare Trajectories from Medical Records: A Deep Learning Approach. *J. Biomed. Inform.* **2017**, *69*, 218–229. [CrossRef]
16. Naz, H.; Ahuja, S. Deep learning approach for diabetes prediction using PIMA Indian dataset. *J. Diabetes Metab. Disord.* **2020**, *19*, 391–403. [CrossRef]
17. Kulkarni, A.; Chong, D.; Batarseh, F.A. 5—Foundations of data imbalance and solutions for a data democracy. In *Data Democracy*, 1st ed.; Batarseh, F.A., Yang, R., Eds.; Academic Press: Cambridge, MA, USA, 2020; pp. 83–106. [CrossRef]
18. Gupta, S.C.; Goel, N. Predictive Modeling and Analytics for Diabetes using Hyperparameter tuned Machine Learning Techniques. *Procedia Comput. Sci.* **2023**, *218*, 1257–1269. [CrossRef]
19. Chang, V.; Bailey, J.; Xu, Q.A.; Sun, Z. Pima Indians diabetes mellitus classification based on machine learning (ML) algorithms. *Neural Comput. Appl.* **2023**, *35*, 16157–16173. [CrossRef] [PubMed]
20. Pima-Indians-Diabetes. Available online: <https://www.openml.org/search?type=data&status=active&id=43582&sort=runs> (accessed on 30 December 2024).
21. Tigga, N.P.; Garg, S. Prediction of Type 2 Diabetes using Machine Learning Classification Methods. *Procedia Comput. Sci.* **2020**, *167*, 706–716. [CrossRef]
22. Ejayi, C.J.; Qin, Z.; Amos, J.; Ejayi, M.B.; Nnani, A.; Ejayi, T.U.; Agbesi, V.K.; Diokpo, C.; Okpara, C. A robust predictive diagnosis model for diabetes mellitus using Shapley-incorporated machine learning algorithms. *Healthc. Anal.* **2023**, *3*, 100166. [CrossRef]
23. Ivanov, I.; Toleva, B. An Algorithm to Predict Hepatitis Diagnosis. In Proceedings of the 11th International Scientific Conference on Computer Science, COMSCI 2023, Sofia, Bulgaria, 18 September 2023.
24. Agung, E.S.; Rifai, A.P.; Wijayanto, T. Image-based facial emotion recognition using convolutional neural network on emognition dataset. *Sci. Rep.* **2024**, *14*, 14429. [CrossRef]
25. Bhagat, M.; Bakariya, B. Implementation of Logistic Regression on Diabetic Dataset using Train-Test-Split, K-Fold and Stratified K-Fold Approach. *Natl. Acad. Sci. Lett.* **2022**, *45*, 401–404. [CrossRef]
26. Kolipaka, V.R.R.; Namburu, A. K-Fold Validation of Multi Models for Crop Yield Prediction with Improved Sparse Data Clustering Process. *Int. J. Intell. Syst. Appl. Eng.* **2023**, *11*, 454–463. Available online: <https://ijisae.org/index.php/IJISAE/article/view/3300> (accessed on 20 December 2024).
27. Prusty, S.; Patnaik, S.; Dash, S.K. SKCV: Stratified K-fold cross-validation on ML classifiers for predicting cervical cancer. *Front. Nanotechnol.* **2022**, *4*, N972421. [CrossRef]
28. Szeghalmy, S.; Fazekas, A.A. A Comparative Study of the Use of Stratified Cross-Validation and Distribution-Balanced Stratified Cross-Validation in Imbalanced Learning. *Sensors* **2023**, *23*, 2333. [CrossRef] [PubMed]
29. Al Sadi, K.; Balachandran, W. Leveraging a 7-Layer Long Short-Term Memory Model for Early Detection and Prevention of Diabetes in Oman: An Innovative Approach. *Bioengineering* **2024**, *11*, 379. [CrossRef] [PubMed]
30. Gragnaniello, M.; Marrazzo, V.R.; Borghese, A.; Maresca, L.; Breglio, G.; Riccio, M. Edge-AI Enabled Wearable Device for Non-Invasive Type 1 Diabetes Detection Using ECG Signals. *Bioengineering* **2025**, *12*, 4. [CrossRef]
31. Fuss, F.K.; Tan, A.M.; Weizman, Y. Advanced Dynamic Centre of Pressure Diagnostics with Smart Insoles: Comparison of Diabetic and Healthy Persons for Diagnosing Diabetic Peripheral Neuropathy. *Bioengineering* **2024**, *11*, 1241. [CrossRef]
32. Jiang, H.; Wang, H.; Pan, T.; Liu, Y.; Jing, P.; Liu, Y. Mobile Application and Machine Learning-Driven Scheme for Intelligent Diabetes Progression Analysis and Management Using Multiple Risk Factors. *Bioengineering* **2024**, *11*, 1053. [CrossRef] [PubMed]
33. Mohanty, P.K.; Francis, S.A.J.; Barik, R.K.; Roy, D.S.; Saikia, M.J. Leveraging Shapley Additive Explanations for Feature Selection in Ensemble Models for Diabetes Prediction. *Bioengineering* **2024**, *11*, 1215. [CrossRef]
34. Geantă, M.; Bădescu, D.; Chirca, N.; Nechita, O.C.; Radu, C.G.; Rascu, Ș.; Rădăvoi, D.; Sima, C.; Toma, C.; Jînga, V. The Emerging Role of Large Language Models in Improving Prostate Cancer Literacy. *Bioengineering* **2024**, *11*, 654. [CrossRef] [PubMed]

35. Bekbolatova, M.; Mayer, J.; Ong, C.W.; Toma, M. Transformative Potential of AI in Healthcare: Definitions, Applications, and Navigating the Ethical Landscape and Public Perspectives. *Healthcare* **2024**, *12*, 125. [CrossRef] [PubMed]
36. Maccaro, A.; Stokes, K.; Statham, L.; He, L.; Williams, A.; Pecchia, L.; Piaggio, D. Clearing the Fog: A Scoping Literature Review on the Ethical Issues Surrounding Artificial Intelligence-Based Medical Devices. *J. Pers. Med.* **2024**, *14*, 443. [CrossRef] [PubMed]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.

Review

Towards Transparent Healthcare: Advancing Local Explanation Methods in Explainable Artificial Intelligence

Carlo Metta ^{1,*}, Andrea Beretta ¹, Roberto Pellungrini ², Salvatore Rinzivillo ¹ and Fosca Giannotti ²

¹ Institute of Information Science and Technologies (ISTI-CNR), Via Moruzzi 1, 56127 Pisa, Italy; andrea.beretta@isti.cnr.it (A.B.); rinzivillo@isti.cnr.it (S.R.)

² Faculty of Sciences, Scuola Normale Superiore, P.za dei Cavalieri 7, 56126 Pisa, Italy; roberto.pellungrini@sns.it (R.P.); fosca.giannotti@sns.it (F.G.)

* Correspondence: carlo.metta@isti.cnr.it

Abstract: This paper focuses on the use of local Explainable Artificial Intelligence (XAI) methods, particularly the Local Rule-Based Explanations (LORE) technique, within healthcare and medical settings. It emphasizes the critical role of interpretability and transparency in AI systems for diagnosing diseases, predicting patient outcomes, and creating personalized treatment plans. While acknowledging the complexities and inherent trade-offs between interpretability and model performance, our work underscores the significance of local XAI methods in enhancing decision-making processes in healthcare. By providing granular, case-specific insights, local XAI methods like LORE enhance physicians' and patients' understanding of machine learning models and their outcome. Our paper reviews significant contributions to local XAI in healthcare, highlighting its potential to improve clinical decision making, ensure fairness, and comply with regulatory standards.

Keywords: artificial intelligence; explainable artificial intelligence; machine learning

1. Introduction

The advent of artificial intelligence (AI) technologies has been transformative for healthcare, offering unprecedented capabilities in disease diagnosis, patient outcome prediction, and the development of tailored treatment plans. However, the increasing complexity and opacity of AI algorithms have amplified the need for transparency and interpretability in their decision-making processes. This has led to the rise of Explainable Artificial Intelligence (XAI), particularly focusing on local interpretation methods such as the LORE (Local Rule-Based Explanations) [1], to make AI decisions in healthcare settings more transparent and comprehensible [2,3].

XAI seeks to equip AI algorithms with explanations, enabling healthcare professionals to gain insights into the rationale behind AI-generated decisions and predictions. XAI methodologies can be broadly divided in two families: *global* [4] and *local* [5] methods. While global methods aim at explaining the general reasoning of an AI model, local methods have the goal of explaining why an AI model gave a certain output for a particular instance, i.e., the data of a particular patient or the diagnostic images belonging to a specific individual. The application of local XAI methods addresses the demand for precise, case-by-case explanations, which are paramount for clinical decision making, enhancing patient care, and fostering trust in AI systems among healthcare providers and patients [6–13].

Significance of Local XAI in Healthcare. The integration of AI in healthcare has been both celebrated for its potential and scrutinized for its challenges, including issues of interpretability, potential biases, and ethical concerns [14–16]. Local XAI methods, such as LORE, offer nuanced, instance-specific explanations that are essential for understanding complex AI decisions in medical contexts [17]. These explanations not only build confidence in AI technologies but also aid in identifying and correcting biases, ensuring ethical usage, and complying with healthcare regulations [18].

Interpretability at the local level allows healthcare professionals to comprehend the reasoning behind specific AI decisions, facilitating their trust in and collaboration with AI systems. Moreover, local explanations play a crucial role in elucidating AI outcomes to patients, empowering them with knowledge about their care processes and decisions.

Challenges and Opportunities in Applying Local XAI. While local XAI methods present a promising approach for enhancing interpretability in healthcare AI, they also introduce challenges such as maintaining model performance and ensuring the relevance and comprehensibility of explanations to end-users. Balancing the complexity of healthcare data with the need for understandable, actionable insights requires innovative solutions and continuous advancements in XAI techniques.

To address these challenges, this paper explores a variety of local XAI approaches, emphasizing the contribution of the LORE method for its ability to generate detailed, rule-based explanations relevant to individual cases. Such techniques are pivotal in translating the intricate patterns recognized by AI models into intelligible information, thereby improving the clinical utility of AI and fostering a collaborative healthcare environment.

The Focused Approach of the XAI Project. The XAI Project G.A. 834756) (<https://xai-project.eu/>, accessed on 14 March 2024) is an ERC-funded project focused entirely on the development of Explainable AI models. In the XAI Project, one of the most prominent domains of research has been local explainability in healthcare AI [19], aiming to refine and promote the application of methods like LORE for better decision-making processes. By concentrating on the development of local XAI techniques, the project seeks to address the specific interpretability needs of healthcare professionals and patients, ensuring that AI systems are not only accurate but also transparent and trustworthy.

In summary, this paper presents a focused narrative on the role of local XAI methods in healthcare, illustrating how such approaches can surmount the interpretability challenges posed by complex AI models. Through detailed case studies and analysis of the LORE method, it aims to showcase the tangible benefits of local explainability in improving patient care, ensuring ethical AI use, and enhancing the acceptance of AI technologies in medical settings.

2. Related Work

The exploration of Explainable Artificial Intelligence (XAI) in healthcare settings, particularly through the lens of local interpretation methods like Local Rule-Based Explanations (LORE) [1], has garnered considerable attention. This section delves into advancements in the interpretability and transparency of AI models in healthcare, emphasizing the importance of local XAI methods and their contributions to the field.

Recent efforts have aimed at enhancing the interpretability of healthcare AI models, with Rajkomar et al. proposing an “Explainable AI Framework for Health” that integrates rule-based models, gradient-based methods, and attention mechanisms for generating interpretable healthcare predictions [20]. This framework’s application to patient mortality prediction has enabled healthcare professionals to derive actionable insights from the model’s decision-making process, showcasing the utility of comprehensive XAI approaches in clinical settings.

Ribeiro et al.’s Anchors method represents another noteworthy advancement, offering rule-based explanations tailored to individual predictions [21]. By concentrating on locally faithful explanations, the Anchors method has empowered healthcare practitioners with a clearer understanding of the factors influencing AI predictions in specific scenarios. Its application across various healthcare domains underscores the method’s effectiveness in improving interpretability at the local level.

The SHAP (SHapley Additive exPlanations) framework, introduced by Lundberg et al., utilizes game theory to allocate feature importance values for individual predictions, thus providing a detailed view of how each input affects the model’s output [22]. Applied in contexts such as hospital readmission prediction, disease progression modeling, and

electronic health records analysis, SHAP has been instrumental in enhancing transparency and interpretability in healthcare AI.

Furthermore, interpretability techniques like LIME (Local Interpretable Model-agnostic Explanations) and LORE (Local Rule-Based Explanations) have seen wide adoption in the healthcare sector [1,23]. LIME's approach to generating local explanations by approximating complex model decision boundaries complements LORE's use of a genetic algorithm to create a synthetic neighborhood for a local interpretable predictor. This predictor, in turn, facilitates the generation of meaningful explanations that include decision rules and counterfactual scenarios, thereby illuminating the influence of specific factors on outcomes [1]. These techniques have been applied to a range of healthcare domains, from disease prediction to medical imaging and clinical decision support, demonstrating their versatility and impact.

Caruana et al.'s work on developing intelligible models for healthcare contexts, such as pneumonia risk prediction and hospital readmission, further highlights the progress in creating interpretable AI systems for clinical use [24]. By employing decision trees and rule-based models, their research has contributed significantly to the field, enhancing both the transparency and the adoption of AI in clinical practice.

The integration of domain knowledge and expert input into XAI approaches marks an evolving research direction, promising to enrich interpretability and align AI decision-making processes with established medical practices. This blend of technical innovation and domain expertise is crucial for advancing the application of local XAI methods in healthcare, ensuring that AI-assisted decision making is both transparent and grounded in clinical realities [25].

In conclusion, advancements in XAI for healthcare, particularly the focus on local methods, highlight a growing commitment to enhancing AI model transparency and interpretability. These efforts underscore the field's progress towards developing AI systems that are not only technically proficient but also understandable and trustworthy for healthcare professionals and patients alike, fostering improved decision making, patient care, and adherence to regulatory standards.

3. Methodology

Before discussing the specific methodologies supporting our research, it is crucial to contextualize our work within the broader landscape of artificial intelligence technologies, particularly deep learning. Deep learning [26], a subset of machine learning, has emerged as a transformative force in various domains, including healthcare. It refers to the development of algorithms that can learn and make decisions or predictions based on data. These algorithms, known as neural networks, are designed to mimic the human brain's architecture and function, processing vast numbers of data to identify patterns and insights that are not immediately apparent to human observers.

As outlined in the seminal work [27], deep learning involves multi-layered neural networks that learn and make inferences from data in a way that captures the complexity and subtlety of the information being processed. This capability makes deep learning particularly valuable in healthcare, where the ability to analyze and interpret complex medical data can lead to more accurate diagnoses, personalized treatment plans, and, ultimately, better patient outcomes.

The key methodologies used in the project are related to the LORE method, introduced by Guidotti et al. [1]. LORE is a powerful framework for generating local and interpretable explanations for machine learning models. LORE utilizes a genetic algorithm to create a synthetic neighborhood, which serves as the basis for training a local interpretable predictor. This predictor captures the underlying logic of the model's decision-making process, enabling the derivation of meaningful explanations.

One of the key characteristics of LORE is its ability to provide transparent and understandable explanations for individual predictions. By focusing on local interpretability, LORE aims to explain the reasoning behind a specific prediction rather than the overall

behavior of the model. This makes it particularly useful in situations where interpretability at the instance level is crucial, such as in healthcare and finance.

The explanations consist of two main components. First, a decision rule is derived from the logic of the local interpretable predictor. This decision rule sheds light on the factors that influenced the model's decision, providing insights into the important features and their corresponding weights. This information helps in understanding the key drivers behind the prediction. Additionally, LORE produces a set of counterfactual rules as part of the explanation. These counterfactual rules suggest modifications to the instance's features that would lead to a different outcome. By providing actionable suggestions for changing the input variables, LORE enables users to explore what-if scenarios and understand how small changes can influence the model's predictions.

The availability of the LORE framework, along with the accompanying code (<https://github.com/riccotti/LORE>, accessed on 14 March 2024), facilitates its adoption and implementation in various domains. In the next sections, different research projects are described. They leverage over the LORE methodology from different points of view.

Detailed LORE Framework

LORE operates on the principle of providing instance-specific explanations by creating a local, interpretable model around a prediction of interest. It begins by selecting an instance for which an explanation is desired. Then, it generates a synthetic dataset that mimics the locality of the original instance through a genetic algorithm. This local dataset is used to train a simple, interpretable model, such as a decision tree, which serves to approximate the behavior of the complex model near the instance. The explanation is then derived from this interpretable model in the form of rules, which highlight the decision-making process for the specific instance.

Implementation Steps: Selection of Target Instance: Choose the specific prediction or instance that requires explanation.

Synthetic Neighborhood Generation: Utilize a genetic algorithm to generate a synthetic dataset that represents the local decision boundary around the target instance.

Training of Interpretable Model: Train a simple model, like a decision tree, on this synthetic dataset to capture the local decision logic of the complex model.

Derivation of Explanation: Extract rules from the interpretable model that explain the prediction of the target instance. These rules offer insights into which features and conditions influence the decision.

LORE's ability to provide clear, case-specific explanations makes it highly valuable in healthcare settings, where understanding the rationale behind AI-driven diagnostic or prognostic predictions is crucial. For instance, LORE has been applied to interpret AI decisions in predicting patient outcomes, understanding disease progression, and personalizing treatment plans. Its interpretability supports clinical decision making, enhances trust among medical practitioners, and facilitates patient communication.

LORE distinguishes itself from other XAI methods like LIME or SHAP primarily through its emphasis on generating a synthetic neighborhood around an instance. This approach allows LORE to provide highly localized explanations that are directly relevant to the specific case at hand. While LIME also focuses on local interpretability, it approximates the model's decision boundary linearly, which might not capture complex nonlinear relationships as effectively as LORE's method. SHAP, on the other hand, provides a global interpretation by assigning importance values to features based on their contribution to the model's output. LORE's advantage lies in its detailed, rule-based explanations that can be more intuitively understood by healthcare professionals for specific patient cases.

4. XAI Frameworks for Healthcare

4.1. DoctorXAI

An ontology-based approach, as described in "Doctor XAI: an ontology-based approach to black-box sequential data classification explanations" [28], aims to provide

explanations for the black-box predicting of multi-labeled, sequential, ontology-linked data [29]. The methodology involves the use of ontologies, which are formal representations of knowledge, to capture domain-specific concepts and relationships [30]. This paper focuses on explaining Doctor AI [31], a multi-label classifier which takes as input the clinical history of a patient in order to predict the next visit.

In greater detail, the methodology begins by selecting real neighbors, which are data points closest to the instance to be explained, either through a standard distance metric or ontology-based similarities. A synthetic neighborhood is then generated by perturbing the real neighbors to maintain locality. The challenge lies in generating meaningful synthetic instances, and here the authors leverage the ICD-9 ontology to ensure the expressiveness of the neighborhood. Unlike other techniques, the perturbations are not applied directly to the instance to be explained to prevent homogeneity in the neighborhood. Two alternative paths are followed; see Figure 1: the red path involves normal perturbation and encoding/decoding steps to transform the data for interpretable models, while the blue path involves ontological perturbation directly on sequential data. In both paths, the synthetic neighborhood is labeled by the black-box model and used to train an interpretable model, such as a multi-label decision tree. Rule-based explanations are then extracted from the decision tree. The methodology extends the general framework with novel contributions for dealing with structured and sequential data. These components can be independently incorporated into the explanation pipeline based on the nature of the data point to be explained.

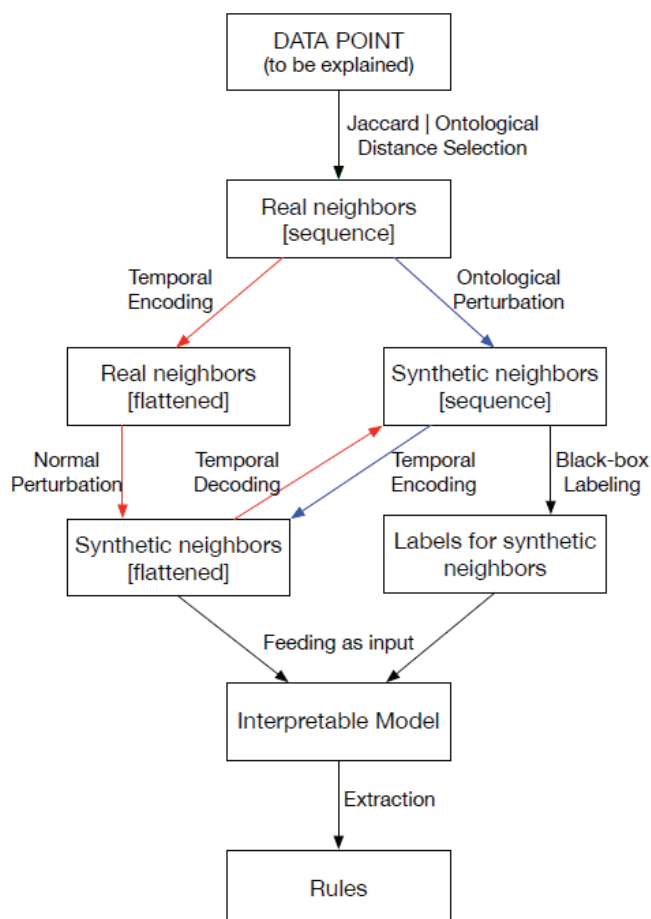


Figure 1. DoctorXAI explanation pipeline from [28].

The authors use the MIMIC-III dataset [32] which contains de-identified health-related data associated with over 40,000 patients who stayed in critical care units of the Beth Israel Deaconess Medical Center between 2001 and 2012. In the experiments conducted,

the ontology-based approach is compared against other existing methods for generating explanations. Various evaluation metrics, such as accuracy, coverage, and coherence, may be employed to assess the quality and comprehensibility of the explanations provided by the approach. The results of the experiments highlight the efficacy of the ontology-based approach in generating meaningful explanations for black-box sequential data classification models. The approach demonstrates its ability to capture domain-specific knowledge, extract relevant features, and provide interpretable explanations that enhance the understanding of the underlying decision-making process (DoctorXAI demo: <https://kdd.isti.cnr.it/DrXAI-viz/>, accessed on 14 March 2024); see Figure 2.

This paper suggests future work in exploring alternative synthetic neighbor generation for sequential data and assessing the impact of random components. Additionally, the authors plan to extend the technique to explain how black-box regressors predict continuous outcomes, which is relevant in healthcare for risk stratification prediction tasks.



Figure 2. Doctor XAI explanation presented to the participants of the experiment. In the image, distinct dots represent a single visit of a patient, and distinct colors represent the relevance of each dot to the algorithmic decision. Dots associated with irrelevant conditions remain gray, whereas those deemed relevant are depicted in blue. Additionally, DoctorXAI highlights, as yellow dots, any conditions absent from the patient’s clinical history that could have altered the algorithmic suggestion.

DoctorXAI has been proven to enhance physicians’ interactions with machine learning models. In the work of Panigutti et al. [33], the authors conduct a rigorous, survey-based analysis of physicians’ interactions with an AI-based Clinical Decision Support System equipped with DoctorXAI. The results indicate that the explanations provided by DoctorXAI enhance the trust between physicians and the AI-based system.

4.2. FairLens

The pervasive application of AI in critical areas, especially healthcare, has brought to the forefront the challenges associated with unintended biases. These biases, if unchecked, can have profound implications, especially when decisions impact patient care. Recognizing this, in [34], the authors present FairLens, a tool designed to audit, discover, and explain biases in AI systems, particularly those deployed in clinical settings. A general overview of FairLens is presented in Figure 3.

FairLens is rooted in a multi-step approach:

- **Stratification of Patient Data:** Before any analysis, the tool stratifies available patient data based on various attributes, including age, ethnicity, gender, and insurance type. This stratification allows for a more granular analysis of how the AI model performs across different patient subgroups.
- **Performance Assessment:** Once stratified, FairLens evaluates the model's performance on these subgroups. It identifies areas where the model might be underperforming or showing biases. This step is crucial as it pinpoints specific patient groups that might be adversely affected by the model's decisions.
- **Explanation of Model Errors:** Going beyond mere identification, FairLens delves into the reasons behind the model's errors. Using advanced XAI techniques, the tool determines which elements in a patient's clinical history contribute to the model's inaccuracies. This step is pivotal as it not only highlights the errors but also provides insights into why they occur.

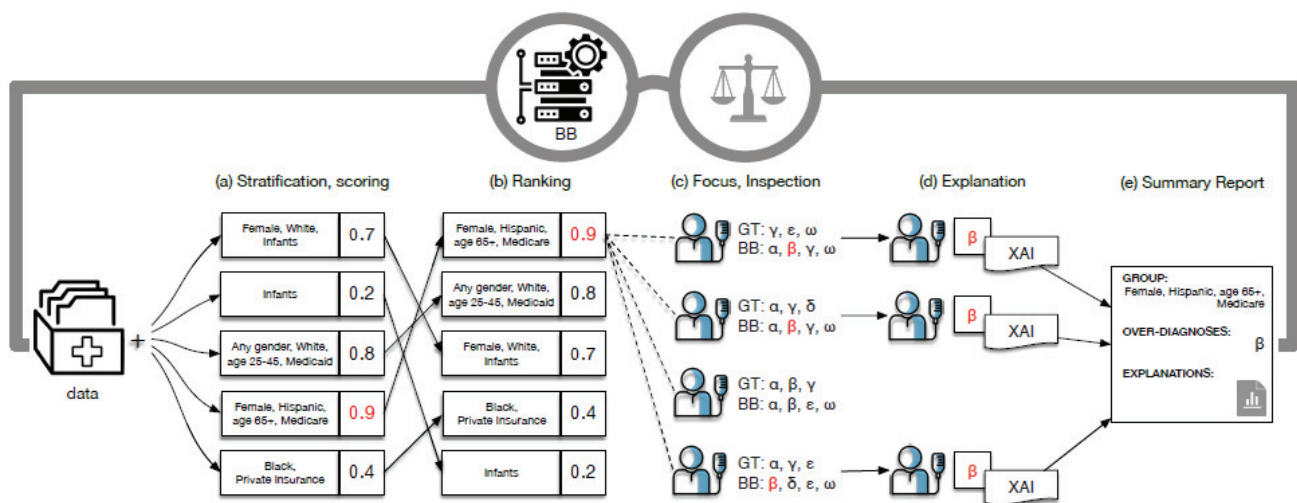


Figure 3. FairLens pipeline from [34].

The FairLens pipeline is a systematic process that ensures a comprehensive audit of the AI model. At first, the patient data are divided based on predefined conditions, creating various groups. Each group is then scored based on the model's performance, providing a quantitative measure of the model's accuracy for that subgroup.

Post-scoring, the groups are ranked. This ranking serves as an indicator, highlighting groups where the model's performance is low-grade. Selected groups (based on ranking or expert input) undergo a detailed inspection. Here, the model's predictions are compared against actual data to identify over-represented or under-represented conditions.

For mislabeled conditions, FairLens provides explanations. It identifies clinical conditions that are frequently misclassified and elucidates the elements in patients' histories that influence these misclassifications. The entire analysis culminates in a comprehensive report that details the findings, providing both a bird's-eye view and in-depth insights.

FairLens represents a significant step forward in ensuring that AI models, especially those in healthcare, are free from detrimental biases. By providing a systematic methodology and a clear pipeline, it offers a robust framework for auditing black-box clinical decision support systems. The tool's ability to not just identify but also explain biases makes it invaluable for healthcare professionals, ensuring that AI-driven decisions are both accurate and fair.

4.3. MARLENA

Machine learning models, especially deep learning ones, have become central to many decision-making systems in healthcare. They assist in diagnosis, predict disease spread, and

help in identifying high-risk patient groups. However, the inherent lack of transparency in these models can lead to mistrust, potential biases, and even legal implications. MARLENA (Multi-label Rule-based EXplANAtions) [35] is introduced as a solution to the interpretability challenge. It is designed to provide explanations for decisions made by multi-label black-box classifiers. The main idea of miming the local behavior of a black-box is common with other approaches such as LIME [23] and LORE [1]. However, none of these approaches is applicable to explain multi-label black-box classifiers. An overview of MARLENA is presented in Figure 4. The novel methodology is broken down into three primary steps:

- **Synthetic Neighborhood Generation:** Before explaining a decision, MARLENA first creates a synthetic neighborhood around the instance in question. This neighborhood is populated with data points that are similar to the instance, ensuring that the explanation is localized and relevant.
- **Learning a Decision Tree:** Using the synthetic neighborhood, MARLENA constructs a decision tree. Decision trees are inherently interpretable, making them suitable for this purpose.
- **Deriving Decision Rules:** From the constructed decision tree, MARLENA extracts decision rules that provide a clear and concise explanation for the black-box decision concerning the instance.

The core methodology revolves around the generation of a neighborhood around the instance that needs elucidation. This is crucial because the explanation is intended to be local, focusing on the behavior of the black-box classifier concerning that specific instance.

To generate this neighborhood, MARLENA employs two strategies. **Constructing a Core Real Neighborhood:** This involves identifying real instances from the dataset that are close to the instance in both the feature space and decision space. This real neighborhood provides a foundation upon which synthetic neighbors can be generated. **Generating Synthetic Neighbors:** Based on the empirical distributions of the instance's features derived from the real neighborhood, MARLENA generates synthetic neighbors. These neighbors are designed to mimic the behavior of the black-box classifier in the vicinity of the instance. Once the neighborhood is established, MARLENA proceeds with the construction of the decision tree and the extraction of decision rules.

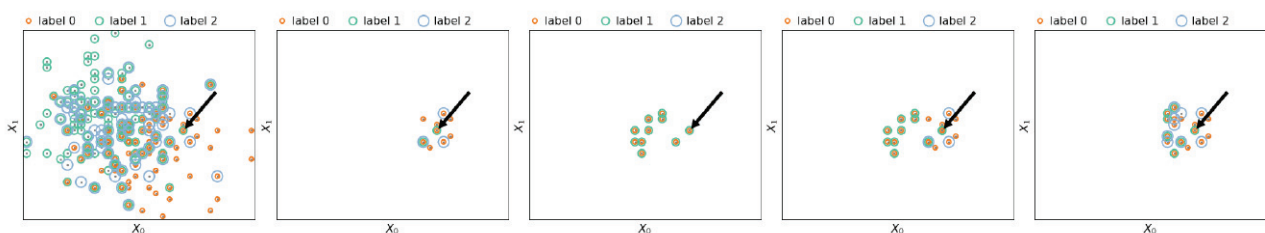


Figure 4. From [35]. A graphical representation of mixed neighborhood generation starting from a sample dataset with three different labels. (1st) a dataset sample, and the arrow points out the instance to explain x ; mixed neighborhood generation: (2nd) real instances close to x with respect to the feature space; (3rd) real instances close to x with respect to the target space; (4th) a merge of the previous sets of instances. Unified core real neighborhood: (5th) real instances close to x with respect to feature and target spaces, i.e., the real core neighborhood.

MARLENA offers a structured approach to demystifying decisions made by multi-label black-box classifiers. By focusing on local explanations and leveraging interpretable models like decision trees, the method ensures that the explanations provided are both meaningful and relevant.

4.4. The International Skin Imaging Collaboration

ABELE (Adversarial Black-box Explainer generating Latent Exemplars) [36] is a local model-agnostic explainer that takes an image and a black-box classifier as input and returns a set of exemplar and counter-exemplar images, as well as a saliency map.

Exemplars and counter-exemplars are synthetically generated images classified with the same outcome as the input image and with an outcome other than the input image, respectively. They can be visually analyzed to understand the reasons for the decision. The saliency map highlights the areas of the input image that contribute to its classification and areas that push it into another class.

ABELE works by generating a neighborhood in the latent feature space using an Adversarial Autoencoder (AAE [37]). The image to be explained is passed as input to the AAE where the encoder returns the latent representation using latent features. A genetic approach maximizing a fitness function was adopted to accomplish the neighborhood generation. In this respect, ABLE takes advantage of a latent version of LORE.

After the generation process, for any instance in the neighborhood, ABLE checks the validity of the instance by querying the discriminator and decoding it into an image. Then, it queries the black-box classifier with the image to obtain the class. Given the local neighborhood, ABLE builds a decision tree classifier trained on the neighborhood labeled with the black-box classifier. The surrogate tree is intended to locally mimic the behavior of the black-box classifier in the neighborhood. It extracts the decision rule and counter-factual rules enabling the generation of exemplars and counter-exemplars.

The overall effectiveness of ABLE lies in the goodness of the encoder and decoder function adopted. The better the AAE, the more realistic and useful the explanations will be.

In recent years, deep learning, particularly through convolutional neural networks (CNNs), has significantly advanced the detection and diagnosis of skin cancer lesions [38–41], achieving diagnostic accuracies comparable to dermatologists. This progress promises improved early detection rates and broader access to high-quality diagnostic services. However, the effectiveness of these models in clinical settings hinges on their interpretability and the transparency of their decision making, ensuring healthcare professionals can integrate AI insights confidently into patient care. In [42–45], a case study on skin lesion diagnosis using a ResNet classifier trained on the ISIC (International Skin Imaging Collaboration) dataset is presented. The classifier's decisions are explained using ABLE.

A user interaction module was implemented as a web application to present the results of the classification and the corresponding explanation. The module communicates with a backend that exposes the functionalities of the black-box and ABLE via a RESTful interface (https://kdd.isti.cnr.it/isic_viz/, accessed on 14 March 2024). The visual space of the application is organized into two sections (see Figure 5). The upper part shows the instance under analysis with the classification returned by the ResNet on the left and a synthetic counter-exemplar image returned by ABLE on the right. The lower part of the module shows four exemplars, i.e., a set of images returned by ABLE that have the same label assigned by the ResNet to the instance under analysis.

The customization of the autoencoder, specifically an Adversarial Autoencoder (AAE), is crucial in this case study due to the complexity of the image classification task and the limitations of the dataset. The ISIC dataset, which is used for training the ResNet classifier [46], presents challenges such as fragmentation, imbalance, lack of uniform digitization, and shortage of data. Training an AAE in a standard fashion without addressing these issues results in poor performance, mainly due to a persistent mode collapse.

To overcome these challenges, a collection of cutting-edge techniques were implemented, including Mini Batch Discrimination and Denoising autoencoders. The model of AAE adopted is a Progressive Growing AAE, which helps achieve more stable training of generative models for high-resolution images. The main idea is to start with a very low-resolution image and, step by step, add blocks of layers that simultaneously increase

the output size of the generator model and the input size of the discriminator model until the desired size is achieved. In this case, the desired size is 224×224 pixels.

The latent space dimension is kept fixed, so the discriminator always takes as input tensors of the same size. The incremental addition of the layers allows the Progressive Growing AAE to first learn large-scale structure and progressively shift the attention to finer detail. This approach greatly reduces mode collapse and enables the generation of varied and high-quality synthetic skin lesion images.

The customization of the AAE is necessary to make it usable for the complex image classification task addressed by the ResNet classifier. After a thorough fine-tuning of all three network structures (encoder, decoder, and discriminator), the Progressive Growing AAE with 256 latent features achieves a reconstruction error measure through RMSE that ranges from 0.08 to 0.24 depending on whether the most common or the rarest skin lesion class is considered. This customization allows ABELE to generate meaningful explanations and can be tested in a survey involving real participants.

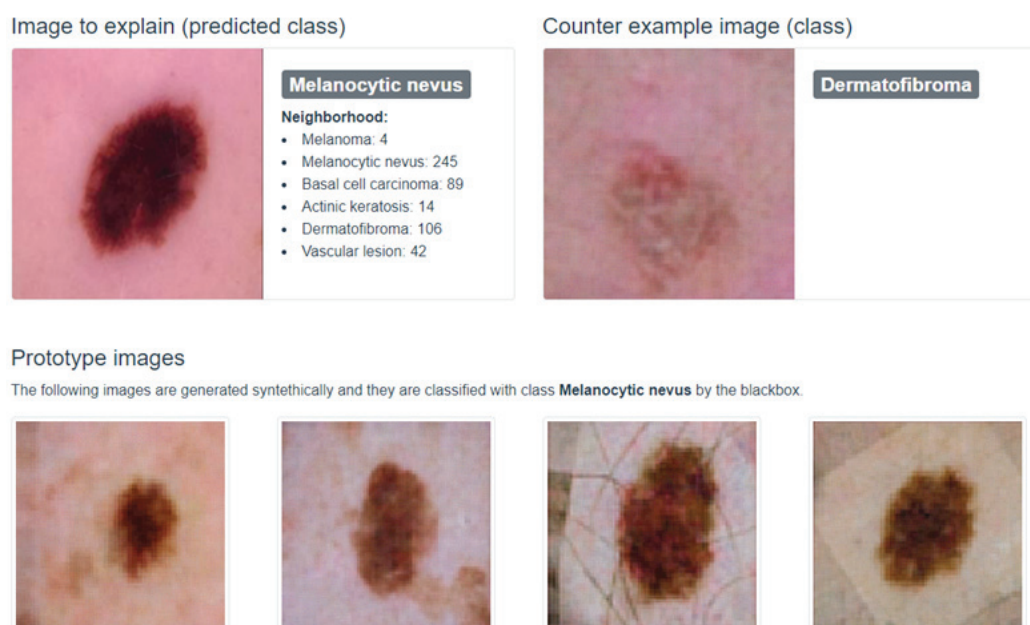


Figure 5. A user visualization module to present the classification and the corresponding explanation. The upper part presents the input instance and a counter-exemplar. The lower part shows four exemplars that share the same class as the input.

A survey was conducted involving domain experts, beginners, and unskilled people to assess the effectiveness of the explanations provided by ABELE. The results of the survey show that the usage of explanations increases trust and confidence in the automatic decision system. This phenomenon is more evident among domain experts and people with the highest level of education. After receiving wrong advice from an AI model, domain experts tend to decrease their trust in the same model for future analysis.

The survey was designed to validate the effectiveness of the explanations returned by ABELE for skin lesion diagnosis. The main purpose was to validate the effectiveness of the explanations in assisting doctors and medical experts in the diagnosis and treatment of skin cancers, as well as to investigate their confidence in automatic diagnosis models based on black-boxes and on the explanations provided by the explainer.

The survey was organized into ten questions composed of various points. Participants were presented with an unlabeled skin lesion image randomly chosen from the dataset and its explanation as generated by ABELE. They were asked to classify the given image among two different given classes exploiting the explanation. Participants were also presented with a labeled image and they were asked to quantify their level of confidence in the black-box classification. The same labeled image was then presented with the visual aid

of the explanation returned by ABELE, and they were asked to quantify their confidence once more after looking at the explanations. Participants were also asked to quantify how much the exemplars and counter-exemplars helped them to classify skin lesion images in accordance with the AI and how much they trust the explanations.

The survey results support the hypothesis that explanation methods without a consistent validation are not useful. The results also highlight the analysis of the latent space of the autoencoder made available by ABELE. The latent space analysis suggests an interesting separation of the images that can hopefully be helpful in separating similar classes of skin lesions that are frequently misclassified by humans.

5. Discussion and Conclusions

This paper has discussed the integration and application of Explainable Artificial Intelligence (XAI) within healthcare, focusing on the challenges, developments, and future directions of XAI in medical diagnostics and patient care. Throughout the ERC XAI project, significant progress has been made in understanding the complex dynamics of XAI in healthcare, its potential benefits, and the inherent challenges encountered.

Challenges and Overcoming Strategies. One of the primary challenges encountered during the project was the complexity of medical data and the difficulty of generating accurate, comprehensible explanations for AI-based decisions. The heterogeneity of healthcare data, along with the high stakes involved in medical decision making, necessitates explanations that are not only technically accurate but also easily understandable by healthcare professionals. To address this, we adopted a multi-faceted approach integrating local explanation generation with formal verification methods. This approach ensured that explanations were both locally relevant and globally consistent with the classifier's logic, thereby enhancing the trustworthiness and explainability of AI systems in healthcare.

Importance and Challenges in Healthcare. In the medical field, the adoption of XAI methods faces unique challenges, including ensuring patient privacy, dealing with high-dimensional data, and the critical need for accuracy. Despite these challenges, the importance of XAI in healthcare cannot be overstated. Detailed, understandable AI explanations empower clinicians to make informed decisions, foster patient trust, and enhance the overall effectiveness of medical treatments. Our work, through projects like DoctorXAI and MARLENA, demonstrates the feasibility and value of applying XAI to a range of healthcare applications, from diagnosing skin lesions to evaluating cardiac risk.

Future Directions and Methodologies. Looking forward, the field of XAI in healthcare is poised for rapid growth and innovation. Future methodologies should focus on improving the robustness and versatility of explanation models, incorporating more diverse data types (e.g., genomic data, electronic health records), and exploring new forms of explanations (e.g., visual explanations, interactive models). Additionally, there is a need for more interdisciplinary research that combines insights from data science, medicine, psychology, and ethics to develop XAI systems that are not only technically proficient but also ethically sound and aligned with patient care goals.

Integrating XAI with Medicine. The future of XAI in healthcare lies in its seamless integration as a decision support system, complementing, not replacing, human expertise. For this to be realized, XAI systems must be designed with openness, transparency, and interpretability at their core. This approach will ensure that healthcare professionals can trust and effectively use AI recommendations, leading to improved patient outcomes.

Building Trust among Healthcare Professionals. To build and maintain trust in XAI systems among medical practitioners, it is essential to focus on user-centered design principles, ensuring that explanations are relevant, actionable, and tailored to the user's expertise level. Avoiding overly complex or opaque AI models and instead emphasizing the transparency and reliability of explanations will be key. Additionally, ongoing education and training for healthcare professionals on the capabilities and limitations of AI will play a critical role in fostering a collaborative environment where AI and human expertise work hand in hand.

In our exploration of Explainable Artificial Intelligence within healthcare, a significant aspect that emerges is the imperative of human–machine collaborative decision making. This symbiotic interaction underscores the philosophy that the greater involvement of human judgment alongside AI can significantly enhance the explainability and ethical dimensions of healthcare decisions. The interplay between human insight and AI’s analytical skills promises to elevate clinical decision making to new heights, fostering a deeper trust and understanding between healthcare providers and the technology they leverage. Moreover, this collaboration can serve as a cornerstone for ethical AI use, ensuring decisions are not only accurate but also transparent and aligned with patient values and needs. As we look towards the future, the integration of human expertise with sophisticated AI algorithms will be crucial in navigating the complex ethical landscape of healthcare, ensuring that AI-assisted decisions are made with a comprehensive understanding of patient care, thereby reinforcing the essence of medicine and underscoring its profound human-centric nature.

In conclusion, the integration of XAI into healthcare holds tremendous promise for enhancing medical diagnostics, patient care, and treatment outcomes. By continuing to address the challenges, leveraging the strengths of AI and human expertise, and focusing on patient-centered outcomes, the future of XAI in healthcare is bright. With sustained research and development, XAI can become an indispensable tool in the medical field, offering insights and explanations that support clinical decision making and contribute to the advancement of personalized medicine.

6. Future and Ongoing Work

As we venture into the future of XAI in healthcare, the emphasis on human–machine collaborative decision making will play a pivotal role in shaping research and development directions. Our forthcoming projects aim to delve deeper into models and frameworks that not only advance the technical capabilities of XAI but also enhance its alignment with human expertise and ethical considerations in clinical settings. This will involve developing systems that are capable of incorporating feedback from healthcare professionals directly into the AI learning process, thus refining the accuracy and applicability of AI outputs in a real-world context. Moreover, the exploration of the ethical and regulatory implications of these collaborative systems will be fundamental. By fostering a more integrated approach to AI in healthcare, where technology and human expertise complement each other, we anticipate not only bridging the gap between AI’s potential and its practical application but also contributing to the development of AI systems that are both ethically responsible and highly effective in enhancing patient care outcomes.

We present a sketch of our ongoing projects that aim to further develop and expand this promising area.

6.1. Cardiac Risk Evaluator

In an upcoming work we present VERIFAI-LORE, a framework designed to enhance the trustworthiness and explainability of AI-based classifiers. This is achieved through a unique integration of search-based approaches, machine-learned explanations, satisfiability solving, and theorem proving. Specifically, it utilizes LORE (Local Rule-Based Explanations) to generate local explanations for classifications by sampling around an instance and constructing a decision tree. These explanations comprise logic rules and counter-rules, indicating the attributes that contributed to a classification and conditions for a different classification, respectively.

However, recognizing that these explanations may be locally valid but underconstrained for certain instances, VERIFAI-LORE introduces a formal verification step. This step involves translating the model into Java and writing a JML (Java Modeling Language) contract in the form of precondition–postcondition pairs to verify the consistency of an explanation with the classifier through theorem proving. This allows the framework to

ensure that explanations are not only statistically valid but also logically consistent across all possible inputs, addressing the challenge of underconstrained explanations.

The integration of explanation generation and formal verification in the VERIFAI-LORE framework aims to provide globally consistent and locally valid explanations for each classification, thereby enhancing classifier trustworthiness. The framework is evaluated in a case study on the prognosis of Acute Coronary Syndrome (ACS) [47], demonstrating its capability to provide classifications with associated confidence levels and explanations that are formally verified for consistency. The evaluation shows that checking the JML contract for explanations takes on average 12.6 s and 304.5 MB for consistent explanations and 68.7 s and 326.5 MB for underconstrained explanations, indicating that underconstrained explanations occur infrequently. This contributes to advancing trustworthy and explainable classification. A visual interface (https://kdd.isti.cnr.it/cre_vue/#/, accessed on 14 March 2024) has been made public and allows for the appreciation of the quality of explanations based on custom inputs entered by the user. Figure 6 demonstrates the explanation of the outcome related to test data.

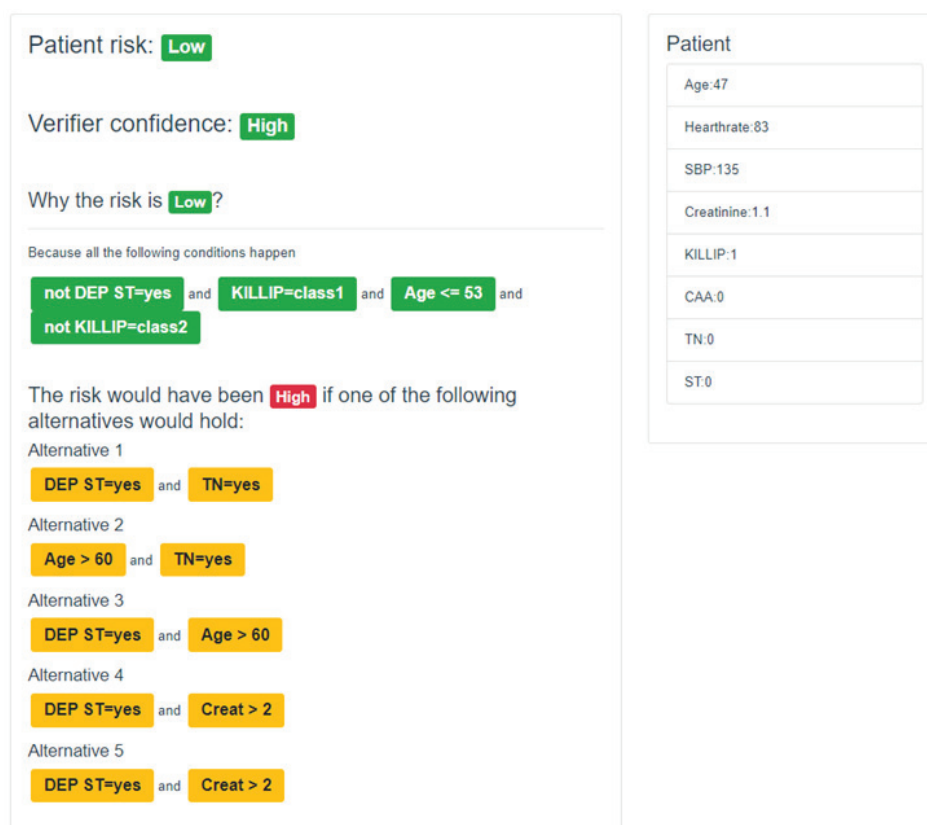


Figure 6. User visualization module to present the classification and the corresponding explanation of the outcome related to test data consisting of age: 47, heart rate: 83, Systolic Blood Pressure: 135, creatinine: 1.1, cardiac arrest at admission: no, ST segment deviation on EKG: no, TN (abnormal cardiac enzymes): no, and KILLIP class: no CHF.

6.2. Prostate Imaging Cancer AI

This ongoing project involves the application of local explanation algorithms to a different context—a prostate cancer MRI dataset [48–51]. This dataset, collected in collaboration with the Prostate Cancer Unit at Ospedale Careggi of Florence and the PI-CAI Grand Challenge (<https://pi-cai.grand-challenge.org/>, accessed on 14 March 2024), is composed of T2-weighted, Apparent Diffusion Coefficient (ADC), and DWI magnetic resonance images. Our goal is to harness the power of local methods in generating meaningful explanations for complex imaging analyses to improve our understanding of prostate cancer diagnostics (Figure 7). The project will not only focus on enhancing the explainability of

local methods but will also delve into an innovative realm of cross-domain explanations between different modalities, i.e., image and tabular data. By doing so, we plan to bridge the gap between different imaging modalities and foster a more integrated, comprehensive understanding of prostate cancer diagnosis, thereby contributing to more effective patient management and treatment outcomes.

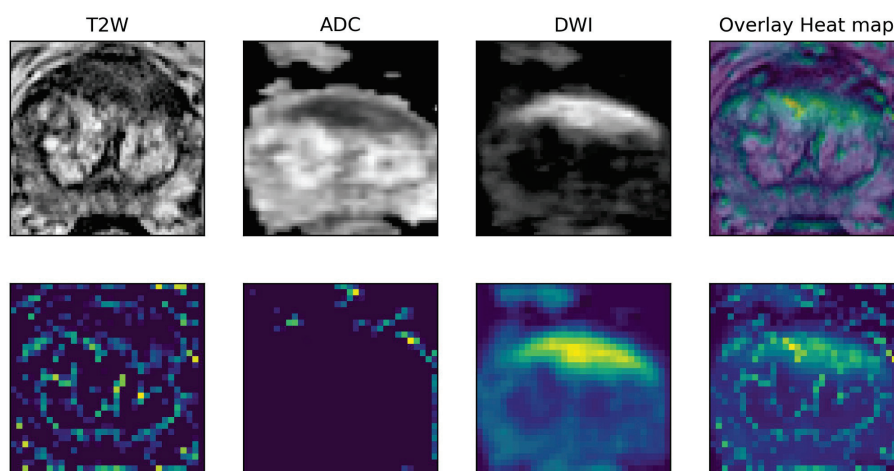


Figure 7. A saliency map extracted from a 3-channel multimodal classifier comprising T2W, ADC, and DWI images of a prostatic gland with a high-grade prostate lesion. Each activation map sizes the contribution of each modality.

6.3. Diabetology

An upcoming project in collaboration with the Diabetology Department at Cisanello University Hospital (Pisa, Italy) explores the application of Explainable AI data describing patients' conditions before and after liver transplantation. Leveraging a fairly small dataset of roughly 450 patients including pre- and post-operative factors, we aim to develop Explainable AI models with the objective of answering a number of questions regarding the effect of liver transplantation on patients [52]. Using explanations as a lens to inspect the models, we are going to investigate possible effects on the onset of diabetes, patient survival, and graft survival. The key challenges involve the highly imbalanced dataset, with the survival rate significantly higher than mortality, and the low number of samples. We plan to apply data balancing techniques cautiously to maintain the integrity of real-world examples so that Explainable AI techniques can still extract meaningful insights from the predictive models. The ultimate goal is to identify key factors contributing to survival rates and understand the relationship between a patient's diabetic condition and survival outcome, thereby offering physicians a way to find new and actionable insights.

6.4. DoctorXAI++

DoctorXAI [28] has been proven to be highly beneficial to clinicians for understanding a machine learning model's decision and for improving trust in the model output [33]. The various components of the DoctorXAI architecture, however, can be improved, in light of recent advancements in the model's performance and in generative AI techniques. This project has the objective of improving DoctorXAI by applying more modern deep learning models to the updated MIMICIV [53] dataset with the ICD10 ontology. The initial experiments we have performed in this direction have the objective of investigating the synthetic neighborhood generation mechanism of DoctorXAI, improving it with the use of Large Language Models (LLMs), trained to approximate the original data distribution. LLMs show the potential of learning the original ontology directly from the data. This may indicate that LLMs can be used both to perform predictions and to replicate ontology-based

explanations using only the raw ICD data. A general schema of our approach is shown in Figure 8.

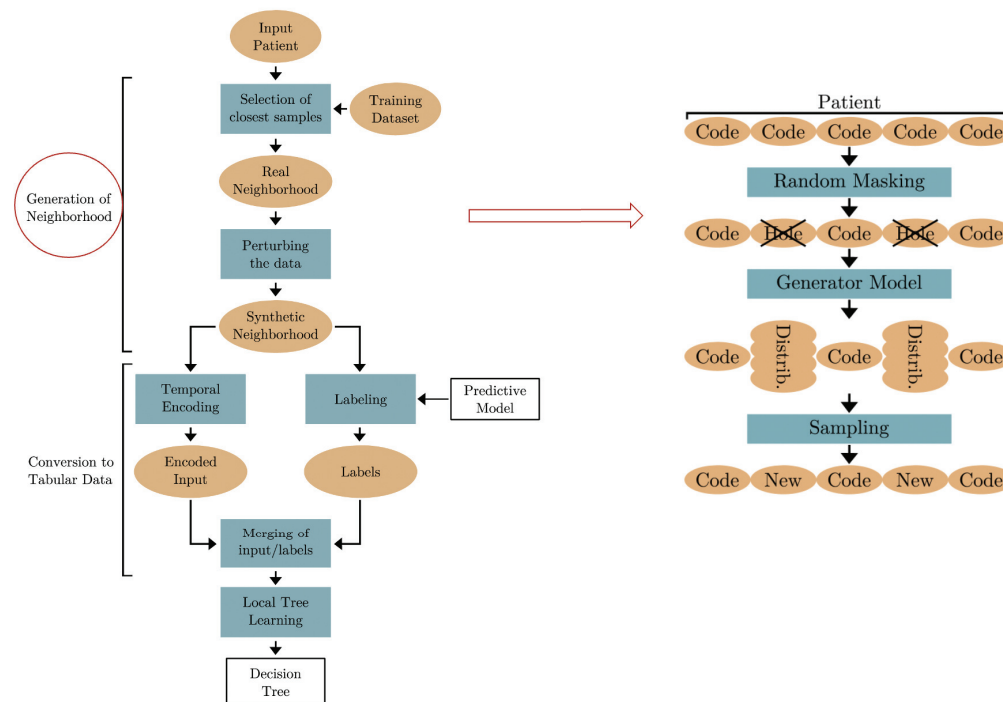


Figure 8. General schema of DoctorXAI++.

Author Contributions: Conceptualization, C.M., F.G., R.P., S.R. and A.B.; Data curation, A.B.; Funding acquisition, S.R. and F.G.; Methodology, C.M., A.B., R.P. and S.R.; Software, C.M., R.P. and S.R.; Validation, C.M., A.B. and R.P.; Visualization, C.M. and S.R. All authors have read and agreed to the published version of the manuscript.

Funding: This work is supported by the European Community under the Horizon 2020 programme, G.A. 871042 *SoBigData++*, G.A. 952026 *HumanE AI Net*, G.A. 101092749 *CREXDATA*, ERC-2018-ADG G.A. 834756 *XAI*, G.A. 952215 *TAILOR*, and the NextGenerationEU programme under the funding schemes PNRR-PE-AI scheme (M4C2, investment 1.3, line on AI), FAIR (Future Artificial Intelligence Research), and “SoBigData.it—Strengthening the Italian RI for Social Mining and Big Data Analytics”—Prot. IR0000013.

Conflicts of Interest: The authors declare no conflicts of interest.

References

- Guidotti, R.; Monreale, A.; Ruggieri, S.; Pedreschi, D.; Turini, F.; Giannotti, F. Factual and counterfactual explanations for black box decision making. *IEEE Intell. Syst.* **2019**, *34*, 14–23. [CrossRef]
- Guidotti, R.; Monreale, A.; Ruggieri, S.; Turini, F.; Giannotti, F.; Pedreschi, D. A Survey of Methods for Explaining Black Box Models. *ACM Comput. Surv.* **2018**, *51*, 1–42. [CrossRef]
- Molnar, C. *Interpretable Machine Learning*; Leanpub: Victoria, BC, Canada, 2020.
- Saleem, R.; Yuan, B.; Kurugollu, F.; Anjum, A.; Liu, L. Explaining deep neural networks: A survey on the global interpretation methods. *Neurocomputing* **2022**, *513*, 165–180. [CrossRef]
- Barredo Arrieta, A.; Díaz-Rodríguez, N.; Del Ser, J.; Bennetot, A.; Tabik, S.; Barbado, A.; Garcia, S.; Gil-Lopez, S.; Molina, D.; Benjamins, R.; et al. Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. *Inf. Fusion* **2020**, *58*, 82–115. [CrossRef]
- Chaddad, A.; Peng, J.; Xu, J.; Bouridane, A. Survey of Explainable AI Techniques in Healthcare. *Sensors* **2023**, *23*, 634. [CrossRef] [PubMed]
- Wang, C.; Liu, Y.; Wang, F.; Zhang, C.; Wang, Y.; Yuan, M.; Yang, G. Towards Reliable and Explainable AI Model for Solid Pulmonary Nodule Diagnosis. *arXiv* **2022**, arXiv:2204.04219.

8. Boutorh, A.; Rahim, H.; Bendoumia, Y. Explainable AI Models for COVID-19 Diagnosis Using CT-Scan Images and Clinical Data. In Proceedings of the International Meeting on Computational Intelligence Methods for Bioinformatics and Biostatistics, Virtual, 15–17 November 2022; pp. 185–199.
9. Papanastasiopoulos, Z.; Samala, R.K.; Chan, H.P.; Hadjiiski, L.; Paramagul, C.; Helvie, M.A.; Neal, C.H. Explainable AI for medical imaging: deep-learning CNN ensemble for classification of estrogen receptor status from breast MRI. In Proceedings of the Medical Imaging 2012: Computer-Aided Diagnosis, Houston, TX, USA, 16–21 June 2012; Volume 11314.
10. Jampani, V.; Ujjwal, Sivaswamy, J.; Vaidya, V. Assessment of computational visual attention models on medical images. In Proceedings of the Eighth Indian Conference on Computer Vision, Graphics and Image Processing (ICVGIP), Mumbai, India, 16–19 December 2012; Volume 80, pp. 1–8.
11. Farahani, F.V.; Fiok, K.; Lahijanian, B.; Karwowski, W.; Douglas, P.K. Explainable AI: A review of applications to neuroimaging data. *Front. Neurosci.* **2022**, *16*, 906290. [CrossRef] [PubMed]
12. Yoo, S.H.; Geng, H.; Chiu, T.L.; Yu, S.K.; Cho, D.C.; Heo, J.; Choi, M.S.; Choi, I.H.; Cung, Van C.; Nhung, N.V.; et al. Deep Learning-Based Decision-Tree Classifier for COVID-19 Diagnosis From Chest X-ray Imaging. *Front. Med.* **2020**, *7*, 427. [CrossRef]
13. Wang, C.; Liu, Y.; Wang, F.; Zhang, C.; Wang, Y.; Yuan, M.; Yang, G. Explainability of deep neural networks for MRI analysis of brain tumors. *Int. J. Comput. Assist. Radiol. Surg.* **2022**, *17*, 1673–1683.
14. Zhou, J.; Chen, F.; Holzinger, A. Towards Explainability for AI Fairness. In *xxAI—Beyond Explainable AI: International Workshop, Held in Conjunction with ICML 2020, July 18, 2020, Vienna, Austria, Revised and Extended Papers*; Holzinger, A., Goebel, R., Fong, R., Moon, T., Müller, K.R., Samek, W., Eds.; Springer International Publishing: Cham, Switzerland, 2022; pp. 375–386. [CrossRef]
15. Mitchell, M.; Wu, S.; Zaldivar, A.; Barnes, P.; Vasserman, L.; Hutchinson, B.; Spitzer, E.; Raji, I.D.; Gebru, T. Model Cards for Model Reporting. In Proceedings of the Conference on Fairness, Accountability, and Transparency (FAT*19), New York, NY, USA, 29–31 January 2019; pp. 220–229. [CrossRef]
16. Tonekaboni, S.; Joshi, S.; Mccradden, M.; Goldenberg, A. Do no harm: A roadmap for responsible machine learning for health care. *Nat. Med.* **2019**, *25*, 1337–1340. [CrossRef]
17. Tonekaboni, S.; Joshi, S.; Mccradden, M.; Goldenberg, A. What Clinicians Want: Contextualizing Explainable Machine Learning for Clinical End Use. *arXiv* **2019**, arXiv:1905.05134.
18. Antoniadi, A.M.; Du, Y.; Guendouz, Y.; Wei, L.; Mazo, C.; Becker, B.; Mooney, C. Current Challenges and Future Opportunities for XAI in Machine Learning-Based Clinical Decision Support Systems: A Systematic Review. *Appl. Sci.* **2021**, *11*, 5088. [CrossRef]
19. Tjoa, E.; Guan, C. A Survey on Explainable Artificial Intelligence (XAI): Toward Medical XAI. *IEEE Trans. Neural Netw. Learn. Syst.* **2021**, *32*, 4793–4813. [CrossRef] [PubMed]
20. Rajkomar, A. Scalable and accurate deep learning with electronic health records. *Npj Digit. Med.* **2018**, *1*, 18. [CrossRef] [PubMed]
21. Ribeiro, M.T.; Singh, S.; Guestrin, C. Anchors: High-Precision Model-Agnostic Explanations. In Proceedings of the AAAI Conference on Artificial Intelligence, New Orleans, LA, USA, 2–7 February 2018; Volume 32.
22. Lundberg, S.M.; Lee, S.I. A unified approach to interpreting model predictions. *Adv. Neural Inf. Process. Syst.* **2017**, *30*, 4765–4774.
23. Ribeiro, M.T.; Singh, S.; Guestrin, C. Why should I trust you? Explaining the predictions of any classifier. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and data Mining, San Francisco, CA, USA, 13–17 August 2016; pp. 1135–1144.
24. Caruana, R.; Lou, Y.; Gehrke, J.; Koch, P.; Sturm, M.; Elhadad, N. Intelligible Models for HealthCare: Predicting Pneumonia Risk and Hospital 30-day Readmission. In Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Sydney, NSW, Australia, 10–13 August 2015; pp. 1721–1730.
25. Larasati, R.; De Liddo, A.; Motta, E. Meaningful Explanation Effect on User’s Trust in an AI Medical System: Designing Explanations for Non-Expert Users. *ACM Trans. Interact. Intell. Syst.* **2023**, *13*, 30. [CrossRef]
26. Goodfellow, I.J.; Bengio, Y.; Courville, A. *Deep Learning*; MIT Press: Cambridge, MA, USA, 2016. Available online: <http://www.deeplearningbook.org> (accessed on 14 March 2024).
27. Zhang, W.; Yang, G.; Lin, Y.; Ji, C.; Gupta, M.M. On Definition of Deep Learning. In Proceedings of the 2018 World Automation Congress (WAC), Washington, DC, USA, 3–6 June 2018; pp. 1–5. [CrossRef]
28. Panigutti, C.; Perotti, A.; Pedreschi, D. Doctor XAI: An ontology-based approach to black-box sequential data classification explanations. In Proceedings of the Conference on Fairness, Accountability, and Transparency, Barcelona, Spain, 27–30 January 2020; pp. 629–639.
29. Poggi, A.; Lembo, D.; Calvanese, D.; De Giacomo, G.; Lenzerini, M.; Rosati, R. Linking Data to Ontologies. In *Proceedings of the Journal on Data Semantics X*; Spaccapietra, S., Ed.; Springer: Berlin/Heidelberg, Germany, 2008; pp. 133–173.
30. Chen, J.; Guo, C.; Lu, M.; Ding, S. Unifying Diagnosis Identification and Prediction Method Embedding the Disease Ontology Structure From Electronic Medical Records. *Front. Public Health* **2022**, *9*, 793801. [CrossRef]
31. Choi, E.; Bahadori, M.T.; Schuetz, A.; Stewart, W.F.; Sun, J. Doctor AI: Predicting Clinical Events via Recurrent Neural Networks. In Proceedings of the Machine Learning for Healthcare Conference, Los Angeles, CA, USA, 19–20 August 2016; pp. 301–318.
32. Johnson, A.E.W. MIMIC-III, a freely accessible critical care database. *Sci. Data* **2016**, *3*, 160035. [CrossRef] [PubMed]
33. Panigutti, C.; Beretta, A.; Giannotti, F.; Pedreschi, D. Understanding the impact of explanations on advice-taking: A user study for AI-based clinical Decision Support Systems. In Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems (CHI’22), New York, NY, USA, 29 April 2022. [CrossRef]

34. Panigutti, C.; Perotti, A.; Panisson, A.; Bajardi, P.; Pedreschi, D. FairLens: Auditing black-box clinical decision support systems. *Inf. Process. Manag.* **2021**, *58*, 102657. [CrossRef]
35. Panigutti, C.; Guidotti, R.; Monreale, A.; Pedreschi, D. Explaining Multi-label Black-Box Classifiers for Health Applications. In Proceedings of the International Workshop on Health Intelligence, Nashville, TN, USA, 4 December 2019; Springer: Berlin/Heidelberg, Germany, 2019.
36. Guidotti, R.; Monreale, A.; Matwin, S.; Pedreschi, D. Black Box Explanation by Learning Image Exemplars in the Latent Feature Space. In Proceedings of the Machine Learning and Knowledge Discovery in Databases, Ghent, Belgium, 14–18 September 2020; Brefeld, U., Fromont, E., Hotho, A., Knobbe, A., Maathuis, M., Robardet, C., Eds.; Springer: Cham, Switzerland, 2020; pp. 189–205.
37. Makhzani, A.; Shlens, J.; Jaitly, N.; Goodfellow, I.J. Adversarial Autoencoders. *arXiv* **2015**, arXiv:abs/1511.05644.
38. Mirikharaji, Z.; Abhishek, K.; Bissoto, A.; Barata, C.; Avila, S.; Valle, E.; Celebi, M.E.; Hamarneh, G. A survey on deep learning for skin lesion segmentation. *Med. Image Anal.* **2023**, *88*, 102863. [CrossRef] [PubMed]
39. Acosta, M.F.J.; Tovar, L.Y.C.; Garcia-Zapirain, M.B.; Percybrooks, W.S. Melanoma diagnosis using deep learning techniques on dermoscopic images. *BMC Med. Imaging* **2021**, *21*, 6. [CrossRef] [PubMed]
40. Esteva, A.; Kuprel, B.; Novoa, R.A.; Ko, J.; Swetter, S.M.; Blau, H.M.; Thrun, S. Dermatologist-level classification of skin cancer with deep neural networks. *Nature* **2017**, *542*, 115–118. [CrossRef] [PubMed]
41. Gouda, W.; Sama, N.U.; Al-Waakid, G.; Humayun, M.; Jhanjhi, N.Z. Detection of Skin Cancer Based on Skin Lesion Images Using Deep Learning. *Healthcare* **2022**, *10*, 1183. [CrossRef] [PubMed]
42. Metta, C.; Guidotti, R.; Yin, Y.; Gallinari, P.; Rinzivillo, S. Exemplars and Counterexemplars Explanations for image classifiers, targeting skin lesion labeling. In Proceedings of the IEEE Symposium on Computers and Communications, Athens, Greece, 5–8 September 2021.
43. Metta, C.; Guidotti, R.; Yin, Y.; Gallinari, P.; Rinzivillo, S. Exemplars and Counterexemplars Explanations for Skin Lesion Classifiers. In Proceedings of the HHAI2022: Augmenting Human Intellect, Munich, Germany, 13–17 June 2022; pp. 258–260.
44. Metta, C.; Beretta, A.; Guidotti, R.; Yin, Y.; Gallinari, P.; Rinzivillo, S.; Giannotti, F. Improving Trust and Confidence in Medical Skin Lesion Diagnosis through Explainable Deep Learning. *Int. J. Data Sci. Anal.* **2023**. [CrossRef]
45. Metta, C.; Beretta, A.; Guidotti, R.; Yin, Y.; Gallinari, P.; Rinzivillo, S.; Giannotti, F. Advancing Dermatological Diagnostics: Interpretable AI for Enhanced Skin Lesion Classification. *Diagnostics* **2024**, *14*, 753. [CrossRef]
46. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep Residual Learning for Image Recognition. In Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 27–30 June 2016; pp. 770–778. [CrossRef]
47. Al-Zaiti, S.; Besomi, L.; Bouzid, Z.; Faramand, Z.; Frisch, S.; Martin-Gill, C.; Gregg, R.; Saba, S.; Callaway, C.; Sejdić, E. Machine learning-based prediction of acute coronary syndrome using only the pre-hospital 12-lead electrocardiogram. *Nat. Commun.* **2020**, *11*, 3966. [CrossRef] [PubMed]
48. Suh, J.; Yoo, S.; Park, J.; Cho, S.Y.; Cho, M.C.; Son, H.; Jeong, H. Development and validation of an explainable artificial intelligence-based decision-supporting tool for prostate biopsy. *BJU Int.* **2020**, *126*, 694–703. [CrossRef] [PubMed]
49. Hassan, M.R.; Islam, M.F.; Uddin, M.Z.; Ghoshal, G.; Hassan, M.M.; Huda, S.; Fortino, G. Prostate cancer classification from ultrasound and MRI images using deep learning based Explainable Artificial Intelligence. *Future Gener. Comput. Syst.* **2022**, *127*, 462–472. [CrossRef]
50. Ramírez-Mena, A.; Andrés-León, E.; Alvarez-Cubero, M.J.; Anguita-Ruiz, A.; Martinez-Gonzalez, L.J.; Alcalá-Fdez, J. Explainable artificial intelligence to predict and identify prostate cancer tissue by gene expression. *Comput. Methods Programs Biomed.* **2023**, *240*, 107719. [CrossRef]
51. Hamm, C.A.; Baumgärtner, G.L.; Biessmann, F.; Beetz, N.L.; Hartenstein, A.; Savic, L.J.; Froböse, K.; Dräger, F.; Schallenberg, S.; Rudolph, M.; et al. Interactive Explainable Deep Learning Model Informs Prostate Cancer Diagnosis at MRI. *Radiology* **2023**, 307. [CrossRef] [PubMed]
52. Bhat, M.; Rabindranath, M.; Chara, B.S.; Simonetto, D.A. Artificial intelligence, machine learning, and deep learning in liver transplantation. *J. Hepatol.* **2023**, *78*, 1216–1233. [CrossRef] [PubMed]
53. Johnson, A.E.W.; Bulgarelli, L.; Shen, L.; Gayles, A.; Shammout, A.; Horng, S.; Pollard, T.J.; Hao, S.; Moody, B.; Gow, B.; et al. MIMIC-IV, a freely accessible electronic health record dataset. *Sci. Data* **2023**, *10*, 1. [CrossRef] [PubMed]

Disclaimer/Publisher’s Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.

Article

Lightweight Super-Resolution Techniques in Medical Imaging: Bridging Quality and Computational Efficiency

Akmalbek Abdusalomov ¹, Sanjar Mirzakhililov ², Zaripova Dilnoza ², Kudratjon Zohirov ³, Rashid Nasimov ⁴, Sabina Umirzakova ^{1,*} and Young-Im Cho ^{1,*}

¹ Department of Computer Engineering, Gachon University Sujeong-Gu, Seongnam-Si 13120, Gyeonggi-Do, Republic of Korea; bobomirzaevich@gmail.com

² Department of Computer Systems/Information and Educational Technologies, Tashkent University of Information Technologies Named After Muhammad Al-Khwarizmi, Tashkent 100200, Uzbekistan; mirzaxililov86@tuit.uz (S.M.); zaripovada85@gmail.com (Z.D.)

³ Department of Computer Systems, Karshi Branch of the Tashkent University of Information Technologies Named After Muhammad al-Khwarizmi, Tashkent 100200, Uzbekistan; qzohirov@gmail.com

⁴ Department of Artificial Intelligence, Tashkent State University of Economics, Tashkent 100066, Uzbekistan; rashid.nasimov@tsue.uz

* Correspondence: sabinatuit@gachon.ac.kr (S.U.); yicho@gachon.ac.kr (Y.-I.C.)

Abstract: Medical imaging plays an essential role in modern healthcare, providing non-invasive tools for diagnosing and monitoring various medical conditions. However, the resolution limitations of imaging hardware often result in suboptimal images, which can hinder the precision of clinical decision-making. Single image super-resolution (SISR) techniques offer a solution by reconstructing high-resolution (HR) images from low-resolution (LR) counterparts, enhancing the visual quality of medical images. In this paper, we propose an enhanced Residual Feature Learning Network (RFLN) tailored specifically for medical imaging. Our contributions include replacing the residual local feature blocks with standard residual blocks, increasing the model depth for improved feature extraction, and incorporating enhanced spatial attention (ESA) mechanisms to refine the feature selection. Extensive experiments on medical imaging datasets demonstrate that the proposed model achieves superior performance in terms of both quantitative metrics, such as PSNR and SSIM, and qualitative visual quality compared to existing state-of-the-art models. The enhanced RFLN not only effectively mitigates noise but also preserves critical anatomical details, making it a promising solution for high-precision medical imaging applications.

Keywords: medical imaging; super-resolution; lightweight model; residual learning

1. Introduction

Medical imaging plays a crucial role in modern healthcare, providing non-invasive methods for diagnosing, monitoring, and treating various medical conditions. With the ever-increasing demand for higher precision in medical image interpretation, enhancing the resolution of medical images has become a fundamental task [1]. High-resolution medical images provide better visualization of intricate anatomical details, aiding healthcare professionals in making accurate diagnoses and planning effective treatments [2]. However, the limitations of imaging hardware often lead to sub-optimal image resolution, making the development of advanced super-resolution techniques imperative.

Single image super-resolution (SISR) aims to reconstruct a high-resolution (HR) image from a given low-resolution (LR) image, recovering fine details that are essential for accurate analysis [3]. In recent years, deep learning-based approaches have demonstrated significant improvements in SISR tasks, surpassing traditional methods in both visual quality and computational efficiency [4]. This paper focuses on enhancing the Residual Feature Learning Network (RFLN) [5] for medical imaging by proposing modifications

tailored specifically for the unique challenges posed by medical datasets [6]. The proposed model, an enhanced version of the RFLN, replaces the original residual local feature block with a standard residual block and introduces deeper layers for improved feature extraction, which is particularly suited for medical images, that often have only a single input channel. Our modifications aim to retain essential features while reducing irrelevant noise, thereby providing high-fidelity reconstructions, which are critical for medical applications. This paper details the architectural changes, the training methodology, and a comprehensive comparison with other state-of-the-art super-resolution models, demonstrating the superiority of the proposed model, both in terms of performance metrics and the visual results of medical imaging tasks.

The main contributions of this work are summarized as follows:

- We propose an improved version of the RFLN architecture tailored specifically for medical imaging. The modifications include replacing the residual local feature blocks with standard residual blocks and increasing the model depth to improve feature extraction and resolution quality.
- To further improve the feature refinement process, we integrate an enhanced spatial attention mechanism into the model. This helps focus on the most relevant areas of the input image, enhancing the overall quality of the super-resolved output, which is crucial for medical image interpretation.
- We conduct extensive experiments using specialized medical imaging datasets, demonstrating the efficacy of the proposed model in terms of both quantitative metrics and qualitative results. The proposed model outperforms existing state-of-the-art models, showcasing its potential for real-world medical applications.
- We provide a detailed analysis of the proposed model's performance compared to other leading SISR models. Our results highlight the advantages of our modifications in handling the unique challenges presented by medical images, such as the preservation of subtle anatomical details and noise reduction.

These contributions collectively advance the field of super-resolution for medical imaging, providing a robust framework that can significantly enhance the quality of medical images, ultimately aiding healthcare professionals in delivering accurate and effective diagnoses. In the following sections, we outline the structure of this paper: Section 2 reviews related works in the field of single image super-resolution (SISR). Section 3 presents the methodology, detailing the proposed enhancements to the RFLN. Section 4 describes the experiments and results, while Section 5 discusses the findings and implications. Finally, Section 6 concludes the paper by summarizing the main contributions and future directions.

2. Related Works

The field of SISR has seen considerable advancements in recent years, primarily due to the advent of deep learning-based methods [7]. Early approaches, such as bicubic interpolation [8] and sparse representation-based methods [9], laid the groundwork for SISR but were limited in their ability to capture complex textures and fine details. Traditional machine learning techniques, such as Sparse Coding-based Super-Resolution (SC-SR) [10] and Neighbor Embedding-based Super-Resolution (NE-SR) [11], provided some improvements over basic interpolation techniques but still struggled with high-frequency detail restoration. The introduction of convolutional neural networks (CNNs) marked a turning point for SISR. Dong et al. introduced the Super-Resolution Convolutional Neural Network (SRCNN) [3], which was among the first to demonstrate the potential of deep learning for super-resolution tasks. SRCNN utilized a straightforward architecture, achieving notable improvements in visual quality over traditional methods. Subsequent models, such as Very Deep Super-Resolution (VDSR) [12] and the Deep Recursive Convolutional Network (DRCN) [13], leveraged deeper architectures and recursive structures to further improve super-resolution performance. Generative adversarial networks (GANs) have also been employed to enhance the perceptual quality of super-resolved images [14]. The SRGAN [15] model was a pioneering effort that introduced adversarial loss to encourage the genera-

tion of high-frequency details [16], producing images that were perceptually closer to the ground truth. However, GAN-based models often suffer from instability during training and may introduce artifacts that compromise the accuracy of medical images, where precise detail is crucial.

More recent works have focused on residual learning and attention mechanisms [17] to address the challenges of high-fidelity image reconstruction. The authors of [18] introduced the Enhanced Deep Residual Network for Single Image Super-Resolution (EDSR), which discarded unnecessary layers and batch normalization to achieve better performance. The authors of [19] proposed the Residual Channel Attention Network (RCAN) [20], which leveraged channel-wise attention to adaptively refine features, leading to significant improvements in image quality [21]. These models have demonstrated the importance of focusing on key features while suppressing irrelevant information, an approach that aligns closely with the goals of our proposed method. The RFLN, which serves as the baseline for our work, utilizes residual local feature blocks to extract both local and global features [22]. The RFLN has shown promise in maintaining a stable gradient flow and capturing intricate details, making it effective for general SISR tasks [23]. However, medical imaging presents unique challenges, such as the need for high precision in areas with subtle anatomical details, which necessitates specialized modifications [24].

In the domain of medical imaging, several models have been proposed to enhance the quality of medical images specifically. The authors of [25] proposed a deep learning framework for MRI super-resolution [26], demonstrating the effectiveness of tailored loss functions for medical data. The authors of [27] introduced a semi-supervised framework, Mine your own Anatomy (MONA), which strategically utilizes dataset characteristics for improved segmentation. The authors of [28] presented a fuzzy neural block that converts images into a fuzzy domain, processes pixel uncertainty with fuzzy rules, and fuses these results with standard convolutional outputs. The authors of [29] proposed a novel Multimodal Multi-Head Convolutional Attention (MMHCA) module to enhance super-resolution for these scans. The module jointly applies spatial-channel attention via convolutions on concatenated input tensors, where the kernel size controls the spatial attention reduction and the number of filters manages the channel attention reduction. The authors of [30] proposed a novel UMIE approach that encodes HQ features directly into the enhancement process using a variation normalization module. This joint modeling of LQ and HQ domains ensures better guidance. The network is trained adversarial with a discriminator to ensure the output belongs to the HQ domain. The work outlined in [31] enhances traditional SR methods by incorporating a channel attention block specifically designed for high-frequency features, which are critical for detailed medical diagnostics. DRFDCAN utilizes a residual-within-residual architecture to improve inference speed and reduce memory usage without compromising image quality. The problem addressed in this study is the ability to enhance low-resolution medical images to high-resolution quality while maintaining computational efficiency. The limitations of existing medical imaging hardware often lead to images that lack sufficient resolution for precise clinical diagnosis. While recent advancements in SISR using deep learning have shown promise, many state-of-the-art models are computationally intensive and not suitable for practical deployment in medical environments. Therefore, the challenge lies in developing a lightweight, efficient SISR model that can effectively enhance image resolution without compromising quality or requiring extensive computational resources. This study aims to address these challenges by proposing an enhanced RFLN specifically tailored for medical imaging.

Previous methods in medical image enhancement face several limitations that are addressed by our model [32]. Many, like MRI super-resolution and MONA, struggle with generalization across different modalities and rely heavily on specific datasets [33]. Our model, with its deeper architecture and standard residual blocks, adapts better across various medical images. Additionally, noise mitigation in earlier approaches, such as fuzzy neural blocks, is less effective, while our model targets noise reduction without sacrificing important features. Computational complexity is another drawback, particularly in models

like MMHCA and DRFDCAN, which are resource-intensive [34]. Our model streamlines this by using enhanced spatial attention for efficient processing. Moreover, methods like DRFDCAN can overemphasize details at the cost of larger structural integrity, which our approach balances. Finally, unlike adversarial-based techniques like UMIE, which can be unstable, our model offers stable, consistent results, without introducing artifacts. Our model provides a more efficient, adaptable, and reliable solution for medical imaging.

Our proposed model builds on these advancements by enhancing the RFLN architecture to better suit medical imaging applications. By replacing the residual local feature blocks with standard residual blocks and increasing the model depth, our approach aims to retain critical features while effectively mitigating noise. Additionally, the incorporation of enhanced spatial attention (ESA) mechanisms ensures that the model can focus on the most relevant features, thereby improving both the visual quality and diagnostic utility of the reconstructed images.

3. Methodology

In this section, we present the enhanced RFLN model tailored for single image super-resolution. In our work, we specifically adapted the model for medical imaging by replacing its residual local feature block with a standard residual block. Section 3.1 provides a comprehensive overview of the baseline model, while Section 3.2 offers a detailed structural analysis of the proposed modifications and loss functions.

3.1. Residual Local Feature Network

The RFLN is a deep learning model primarily designed to enhance single image super-resolution tasks (Figure 1). It leverages the concept of residual learning, which facilitates the ability to focus on learning the residual (or difference) between the LR input and the HR output, rather than attempting to directly reconstruct the high-resolution image. This technique has proven effective in mitigating vanishing gradient issues, allowing for more efficient and accurate deep network training. At the core of the RFLN architecture are residual local feature blocks, which are designed to extract and preserve both local and global features from the input image. These blocks work in synergy to capture intricate details across various scales of the image, ensuring that the reconstructed high-resolution image maintains sharpness and fine texture details. The architecture also incorporates several layers of convolutional operations, each followed by non-linear activation functions, which together contribute to the progressive refinement of the image resolution.

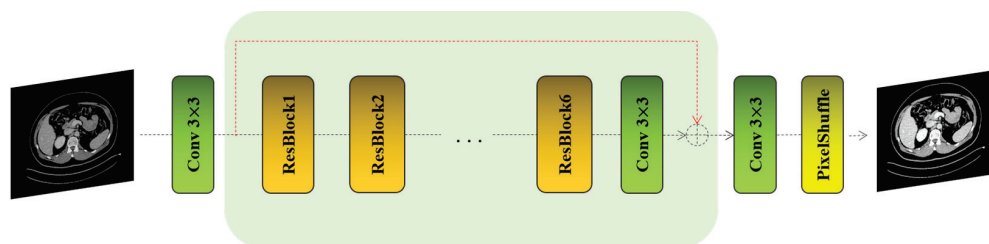


Figure 1. The architecture of the modified RFLN.

The residual local feature blocks within the network are particularly adept at retaining essential image features while suppressing irrelevant noise, which is critical for high-quality image super-resolution. In essence, RFLN focuses on learning local dependencies within the image, while the residual connections across layers ensure that the network can maintain stable gradient flow during training, preventing performance degradation in deeper layers. Furthermore, the integration of upsampling techniques toward the later stages of the network enables the final generation of the high-resolution output. These methods, often involving pixel shuffling or deconvolution, ensure that the output resolution is increased efficiently without introducing significant artifacts. By leveraging these techniques, RFLN

can achieve precise and visually appealing super-resolution, making it particularly effective for applications that require high-fidelity image reconstruction.

3.2. The Proposed Model

where each block consists of three convolutional layers with 3×3 kernel sizes, each paired with ReLU activation functions to introduce non-linearity to the feature maps (Figure 2). Following this, a concatenation layer combines the output of the block with the input feature map.

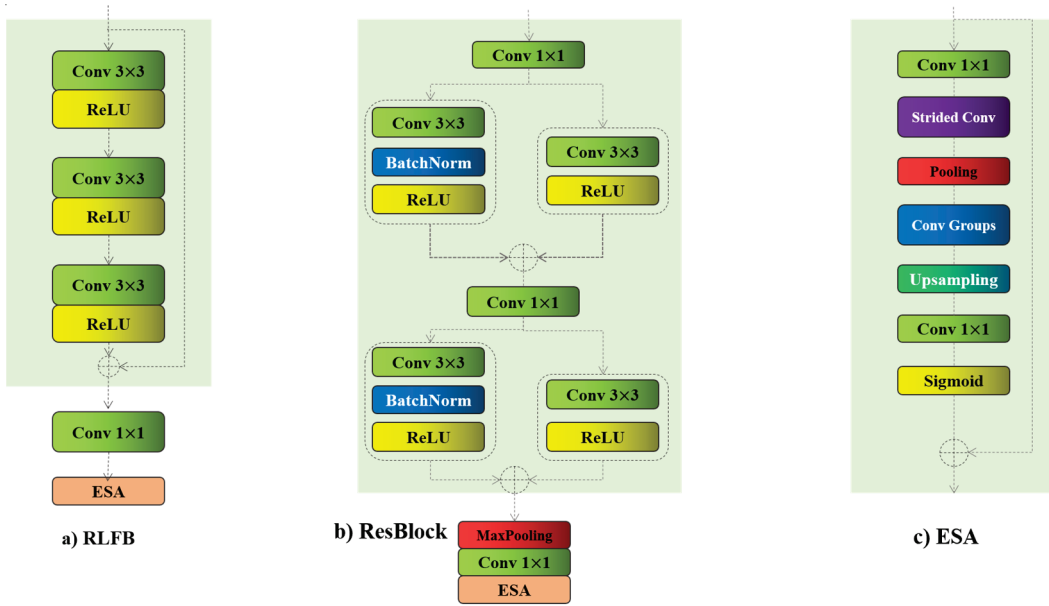


Figure 2. (a) RLFB: The residual local feature block; (b) ResBlock: Modified RLFB; (c) ESA: Enhanced Spatial Attention.

These steps are followed by the addition of another convolutional layer and an ESA block, which contribute to the enhancement of inference time. The incorporation of the ESA block in the proposed architecture is essential for effectively refining feature selection and improving the quality of the reconstructed images. In medical imaging, preserving subtle anatomical details is critical for accurate diagnosis, and the ESA block helps the model focus on these crucial regions of the input images. The ESA block operates by selectively emphasizing significant spatial areas, which enhances the model's ability to retain important features while mitigating irrelevant noise. By capturing both local and global dependencies, the ESA block allows the model to distinguish between essential anatomical information and less relevant background features. This targeted attention is particularly beneficial for medical images, where precise feature extraction can greatly influence clinical outcomes. The use of the ESA block ultimately results in higher-quality super-resolved images with superior visual fidelity, making it an integral component of the model architecture for medical applications.

In our modified ResBlocks, we restructure the entire model, slightly increasing its depth to capture more comprehensive information, as we are working with medical images that typically have only a single input channel. This adjustment allows for more effective feature extraction, ensuring that critical details are retained, which is essential for the precision required in medical image processing. The input image, $X_{input} \in R^{W \times H \times C}$, goes as the first layer into ResBlock1, as shown in Equation (1):

$$F_{layer1} = F_{1 \times 1}(X_{input}) \quad (1)$$

F_{layer1} constitutes the initial layer, which comprises a single convolutional layer with a 3×3 kernel size, designed to extract low-level features:

$$F_{layer2} = \max\left(0, x \cdot \left(\text{BatchNorm}\left(F_{3 \times 3}\left(F_{layer1}\right)\right)\right)\right) \quad (2)$$

Equation (2), where F_{layer2} denotes the second layer, incorporates a 3×3 convolution layer that enhances feature mapping. This is succeeded by batch normalization, which stabilizes the learning process by normalizing the input layer by re-centering and re-scaling. Following batch normalization, the ReLU activation function is applied to introduce non-linearity, facilitating the ability of the model to learn complex patterns in the data:

$$F_{layer3} = \max\left(0, x \cdot \left(F_{3 \times 3}\left(F_{layer1}\right)\right)\right) \quad (3)$$

$$F_{layer4} = F_{1 \times 1}\left(F_{concat}\left(F_{layer2}, F_{layer3}\right)\right) \quad (4)$$

Equation (3) delineates the layers equipped with a ReLU activation function and a 3×3 convolution layer, tailored for extracting coarser features. Concurrently, Equation (4) illustrates the process of element-wise concatenation, coupled with a layer dedicated to the extraction of low-level features:

$$F_{layer5} = \max\left(0, x \cdot \left(\text{BatchNorm}\left(F_{3 \times 3}\left(F_{layer4}\right)\right)\right)\right) \quad (5)$$

$$F_{layer6} = \max\left(0, x \cdot \left(F_{3 \times 3}\left(F_{layer4}\right)\right)\right) \quad (6)$$

Equations (5) and (6) replicate the same blocks as those in Equations (2) and (3). In these instances, the input feature map, following concatenation, restores some information and possesses more complex features for subsequent feature extraction layers. Additionally, element-wise concatenation aids the feature map in preserving essential information, preventing the loss of crucial details:

$$F_{layer7} = F_{concat}\left(F_{layer6}, F_{layer5}\right) \quad (7)$$

$$F_{layer8} = F_{1 \times 1}\left(\text{MaxPooling}\left(F_{layer7}\right)\right) \quad (8)$$

$$F_{layer9} = F_{ESA}\left(F_{layer8}\right) \quad (9)$$

In Equation (7), we employ the second concatenation layer, followed by the application of a pooling layer. This step ensures that the model captures the most essential or representative features from the input feature maps. Specifically, we utilize max pooling, which selects the maximum value from a group of pixels within a feature map. This approach is effective because the maximum value typically represents a distinct feature or shape characteristic, aiding in the robust recognition and representation of important spatial hierarchies in the data. In the refinement stages, we employ the Mean Squared Error (MSE) loss. This metric is pivotal in diminishing the squared discrepancies between the predicted and actual pixel values, which is integral for augmenting the precision of super-resolution outcomes. The MSE loss quantifies the average squared variances between the true and forecasted values. It is conventionally articulated as follows:

$$L_{MSE} = \frac{1}{N} \sum_i (y_i - y'_i)^2 \quad (10)$$

This expression underscores the aim to minimize the mean of the squared errors, thereby enhancing the fidelity and quality of the super-resolved images. In Equation (10), y_i represents the ground truth high-resolution images, while y'_i denotes the predicted high-resolution images. These variables are crucial for assessing the performance of

super-resolution models, focusing on reducing the discrepancies between the actual and computed outputs to enhance image quality.

4. Experiments and Results

4.1. The Dataset

In our research, we harness specialized datasets that are designed specifically for super-resolution tasks, such as Figshare and the Kidney Stone collections. These datasets are expertly structured to facilitate the training and evaluation of sophisticated super-resolution models. As integral components of our study, they include a comprehensive array of images meticulously prepared and annotated to support the development of cutting-edge image enhancement technologies. These datasets are instrumental in advancing techniques that precisely enhance image resolution, catering to the unique demands of super-resolution applications. Each dataset includes a diverse set of images from various scenarios, enriched with detailed annotations that delineate critical areas within the images, thereby providing high-quality data crucial for refining model accuracy and performance. The characteristics of the datasets used for testing, namely Figshare and Kidney Stone, are presented below. These datasets were carefully selected to provide a diverse range of medical images, ensuring that the proposed model could be effectively evaluated across different medical imaging modalities (Table 1).

Table 1. Characteristics of the Figshare and Kidney Stone datasets used in our experiments.

Dataset	Number of Images	Image Resolution	Modality	Description
Figshare	500	512 × 512	CT and MRI	A diverse set of medical images annotated for super-resolution tasks, covering multiple anatomical regions.
Kidney Stone	300	256 × 256	Ultrasound	A specialized dataset focusing on kidney stone images, annotated for improved clarity in super-resolution tasks.

Table 1 provides a clear overview of the datasets used, detailing the number of images, their resolution, the imaging modality, and a brief description of the dataset content. Including this information allows for better reproducibility of this study and transparency regarding the data used.

4.2. Data Preprocessing

The data preprocessing phase for the Figshare and Kidney Stone datasets, which are tailored for medical super-resolution tasks, involves a pipeline rigorously designed to enhance the robustness and accuracy of the proposed model. Initially, all images are resized to ensure uniformity, meeting the specific input requirements of the proposed model and maintaining consistency across the dataset. Considering the variability in medical imaging, such as differing modalities and scan qualities, we introduce a series of augmentations. These adjustments include random rotations, flips, and variations in brightness and contrast, which help to mimic the diverse conditions found in real-world medical settings. Additionally, to replicate common imaging challenges like noise and slight blurring, which may occur due to machine imperfections or patient movement, Gaussian noise and blurring are applied. Through this preparation process we aim to acclimate the model to potential real-world imperfections it might encounter.

Normalization of each image follows, standardizing pixel values to align with the neural expectations of the proposed model, thus promoting stable and efficient learning. If the datasets include specific annotations, such as regions of interest around kidney stones, these are meticulously adjusted to maintain accuracy after image transformations. Lastly, to ensure comprehensive training and prevent model bias, we balance the dataset by

employing techniques like oversampling or undersampling to represent various medical conditions adequately (Figure 3).

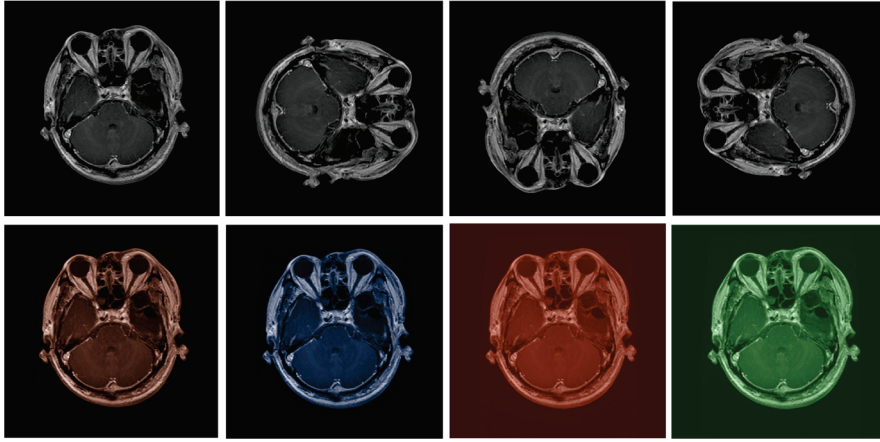


Figure 3. Data preprocessing.

4.3. Metrics

In this paper, the primary metrics we use to evaluate the model's performance are the Peak Signal-to-Noise Ratio (PSNR) and the Structural Similarity Index Measure (SSIM). These metrics are crucial for assessing the quality of the images that have been enhanced by the super-resolution process. PSNR is used to measure the ratio between the maximum possible power of a signal and the power of any corrupting noise that affects the fidelity of its representation, which, in image processing, translates to how much detail can be perceived in the super-resolved image:

$$\text{PSNR} = 20 \cdot \log_{10} \left(\frac{\text{MAX}_I}{\sqrt{\text{MSE}}} \right) \quad (11)$$

where MAX_I is the maximum possible pixel value of the image (e.g., 255 for 8-bit images) and MSE is the mean squared error between the original and the reconstructed images. SSIM, on the other hand, evaluates the visual impact of three characteristics of an image (luminance, contrast, and structure), thereby providing a more comprehensive measure of image quality and perceived changes in structural information. These metrics are instrumental in demonstrating the effectiveness of the proposed model in improving the resolution of medical images while maintaining a balance between computational efficiency and enhancement quality:

$$\text{SSIM}(x, y) = \frac{(2\mu_x\mu_y + c_1)(2\sigma_{xy} + c_2)}{(\mu_x^2 + \mu_y^2 + c_1)(\sigma_x^2 + \sigma_y^2 + c_2)} \quad (12)$$

where μ_x and μ_y are the averages of x and y , respectively, σ_x , and σ_y are the variance of x and y , respectively, σ_{xy} is the covariance of x and y , and c_1 and c_2 are two variables to stabilize division with a weak denominator.

4.4. Experimental Results

Figures 4 and 5 illustrate a meticulous evaluation of super-resolution techniques applied to a medical MRI image, focusing on a patch region within the brain. These figures display a sequence of images showing the original scan and subsequent enhancements using various super-resolution models, including SRCNN, SRGAN, VDSR [12], a baseline model, and the proposed model.

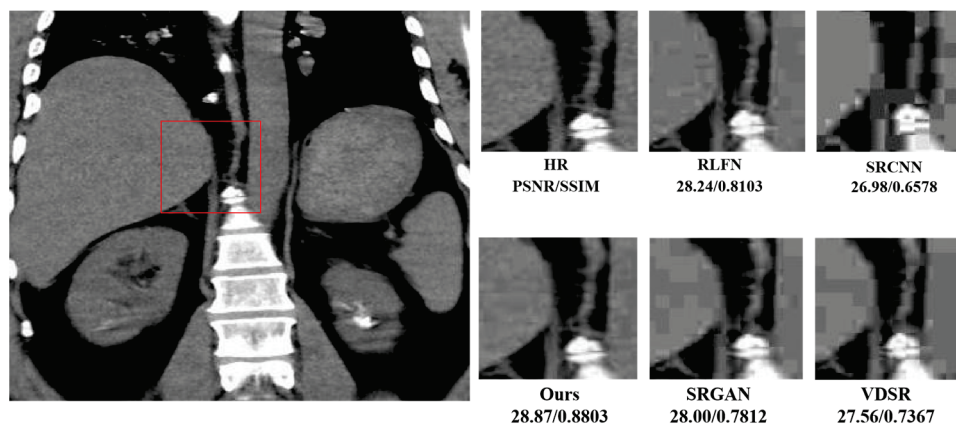


Figure 4. MRI images.

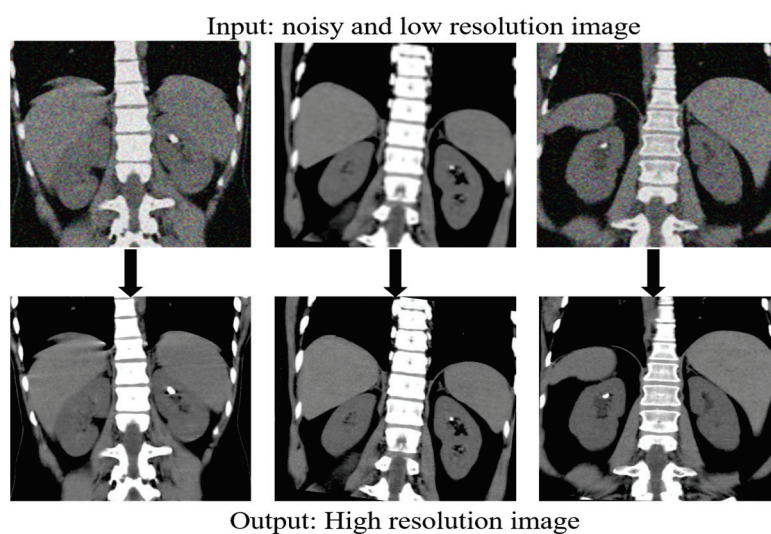


Figure 5. Presents a series of comparisons of our proposed model under noisy and low-contrast conditions.

The sequence starts with the original image, taken from the highlighted area, which appears at a lower resolution, with less distinct features. As we progress through the sequence, each model attempts to improve the clarity and detail of the image. SRCNN, as one of the pioneering models in this field, enhances the image to a PSNR of 26.98 dB and an SSIM of 0.6578, VDSR [12] advances this further to a PSNR of 27.56 dB and an SSIM of 0.7367, while SRGAN achieves a PSNR of 28.00 dB and an SSIM of 0.7812. In Figures 4, 6 and 7, each of these results is visible as almost the same. Within this cohort, SRGAN exhibits the most commendable performance, achieving a PSNR of 28.45 dB and an SSIM of 0.8423, indicating its superior capability in enhancing image quality and structural fidelity. Following SRGAN, VDSR [12] attains a PSNR of 27.96 dB and an SSIM of 0.7865, showcasing its effective resolution enhancement features. Conversely, SRCNN, despite being an early innovator in this domain, records a PSNR of 27.14 dB and an SSIM of 0.6701. The baseline model, noted for its lightweight architecture, surprisingly yields a higher PSNR of 28.24 dB and an SSIM of 0.8103, demonstrating efficient performance despite its simplicity. However, the proposed model surpasses all these metrics, delivering the highest refinement with a PSNR of 28.87 dB and an SSIM of 0.8803.

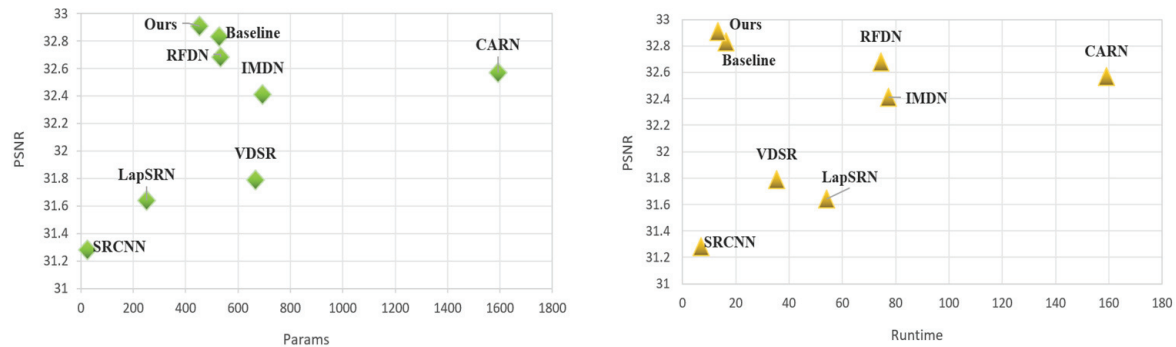


Figure 6. Illustration of the PSNR, Runtime, and Params for dataset1.

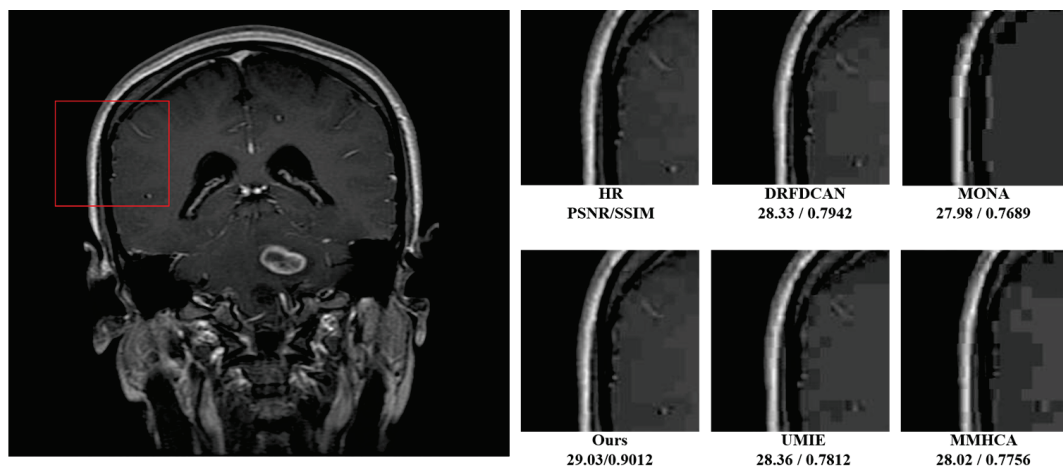


Figure 7. Visual comparison of the SOTA models.

Figure 5 presents a series of comparisons of our proposed model under noisy and low-contrast conditions. These examples highlight the model ability to mitigate noise and enhance subtle anatomical features, demonstrating its robustness and reliability for use in real-world clinical settings. The visual examples clearly show that our model effectively enhances image quality even in challenging situations, such as those with significant noise or low contrast, which are common in clinical practice. This further supports the suitability of our approach for medical applications where high fidelity is crucial.

4.5. Comparison of the Baseline Models

Table 2 offers a nuanced comparison of various super-resolution models, evaluating their performance across different scaling factors and showcasing their efficiency and effectiveness through parameters like runtime and the image quality metrics PSNR and SSIM. Dataset1 and Dataset2 refer to the Figshare and Kidney Stone datasets, respectively, which were used in our experiments. No additional datasets were introduced. We have explicitly stated this to ensure clarity and transparency in our manuscript. Furthermore, we performed several preprocessing modifications to the original datasets. These modifications included resizing the images to meet the input requirements of the proposed model, applying random augmentations such as rotations, flips, and brightness adjustments, and adding Gaussian noise to simulate realistic conditions. These preprocessing steps are detailed in the Data Preprocessing section to provide a comprehensive understanding of the dataset preparation.

Table 2. Results from the baseline models’ comparison.

Scale	Model	Params [K]	Runtime [ms]	Dataset1 PSNR/SSIM	Dataset2 PSNR/SSIM
2×	SRCNN [3]	24	6.92	31.28/0.9012	33.55/0.9312
	VDSR [12]	666	35.37	31.79/0.9056	33.78/0.9421
	IMDN [35]	694	77.34	32.41/0.9123	34.12/0.9512
	RFDN [36]	534	74.51	32.68/0.9154	34.45/0.9535
	CARN [37]	1592	159.10	32.57/0.9134	34.01/0.9486
	LapSRN [38]	251	53.98	31.64/0.9142	33.78/0.9356
	Baseline	527	16.41	32.83/0.9175	33.86/0.9369
	Ours	452	13.23	32.91/0.9188	34.07/0.9387
4×	SRCNN	57	1.90	27.78/0.7120	28.45/0.7276
	VDSR [12]	666	8.95	27.89/0.7165	28.61/0.7287
	IMDN [35]	715	20.56	27.95/0.7810	28.98/0.7301
	RFDN [36]	550	20.40	28.12/0.8023	29.23/0.7453
	CARN [37]	1592	39.96	27.86/0.7712	28.52/0.7282
	LapSRN [38]	502	66.81	27.15/0.6813	28.34/0.7145
	Baseline	543	16.41	28.34/0.8230	29.37/0.7478
	Ours	468	13.23	28.46/0.8256	29.48/0.7513

At a 2× scale, models such as IMDN [35] and RFDN [36] demonstrate high-quality image enhancement, with IMDN [35] achieving notably high PSNR and SSIM scores. SRCNN, while less complex, with the fewest parameters, offers a fast processing time, making it efficient though slightly less effective in terms of the quality metrics (Figure 6). VDSR [12] and CARN [37], with their higher complexity, show improved image quality at the cost of an increased computational load. The proposed model distinguishes itself by achieving the best balance between runtime and high-quality results, outperforming other models in both datasets at this scale. When scaling up to 4×, all models generally experience a decrease in performance, indicative of the increased challenge associated with higher scaling factors.

However, the proposed model maintains robust performance, surpassing other models in image quality, which highlights its superior design for handling more significant upscaling challenges effectively. RFDN [36] and the baseline model also exhibit commendable performances, suggesting their potential utility in applications where a balance between speed and image quality is crucial. The diversity in the model performances across the two datasets underscores the importance of selecting a model based on specific application needs, considering factors such as the desired balance between image quality and computational efficiency. The proposed model, with its exceptional performance metrics, exemplifies the advancements in super-resolution technology, promising significant improvements in applications requiring detailed image resolution enhancements.

4.6. Comparison with SOTA Models

This section presents a comparison of the proposed RFLN with various SOTA models in medical imaging. The evaluation was performed on two distinct datasets, Dataset1 and Dataset2, using PSNR and SSIM as the primary metrics for assessing image quality. Additionally, the computational efficiency of each model was compared by measuring the runtime. The performance of the models at both 2× and 4× scaling factors is summarized in Tables 2 and 3. The results highlight the superiority of the proposed RFLN model in preserving anatomical details, mitigating noise, and providing higher visual fidelity compared to previous methods.

Table 3. Performance comparison at a $2\times$ scaling factor.

Model	Params (K)	Runtime (ms)	Dataset1 PSNR/SSIM	Dataset2 PSNR/SSIM
MRI Super-Resolution [18]	1224	25.12	30.45/0.8921	32.78/0.9102
MONA [19]	1045	28.64	31.05/0.8998	33.22/0.9197
Fuzzy Neural Block [28]	880	22.51	30.12/0.8823	31.90/0.9051
MMHCA [29]	1342	55.78	31.56/0.9045	33.64/0.9228
UMIE [30]	1605	68.32	31.78/0.9107	33.85/0.9285
DRFDCAN [31]	953	26.91	32.14/0.9175	34.05/0.9322
Ours	452	13.23	32.91/0.9188	34.07/0.9387

In Table 3, the proposed RFLN model achieves the highest PSNR and SSIM for both datasets. The model's efficient residual learning and enhanced spatial attention ensure high-quality image reconstruction, while its runtime is the fastest among the compared models, demonstrating the balance between image quality and computational efficiency.

At the $4\times$ scaling factor, as shown in Table 4, the RFLN model continues to outperform the other methods, particularly in PSNR, which measures the sharpness and fidelity of the super-resolved images. Figure 6 presents visual comparisons between the different models for a patch from a brain MRI scan from Dataset1. The images reconstructed by our model exhibit the best clarity and structural accuracy, with fewer artifacts compared to the other methods.

Table 4. Performance comparison at a $4\times$ scaling factor.

Model	Params (K)	Runtime (ms)	Dataset1 PSNR/SSIM	Dataset2 PSNR/SSIM
MRI Super-Resolution [18]	1224	25.12	27.12/0.7521	28.95/0.7779
MONA [19]	1045	28.64	27.98/0.7689	27.78/0.7901
Fuzzy Neural Block [28]	880	22.51	26.85/0.7456	28.62/0.7654
MMHCA [29]	1342	55.78	28.02/0.7756	27.86/0.7922
UMIE [30]	1605	68.32	28.36/0.7812	28.01/0.8027
DRFDCAN [31]	953	26.91	28.33/0.7942	29.22/0.8115
Ours	468	13.23	28.46/0.8256	29.48/0.8513

While all models show a decline in performance at higher scaling factors, the proposed RFLN model maintains its advantage in preserving image quality with fewer parameters and lower runtime, making it highly efficient for clinical deployment.

5. Discussion

The experimental results presented in this paper demonstrate the effectiveness of the enhanced RFLN for single image super-resolution in medical imaging. The proposed model consistently outperformed existing state-of-the-art models in both quantitative metrics, such as PSNR and SSIM, and qualitative visual assessments. This section provides a discussion of the key findings, their implications, and the limitations of the current approach. One of the significant findings of this study is the importance of enhancing feature extraction through the use of deeper residual blocks. By replacing the original residual local feature blocks with standard residual blocks and increasing the network

depth, the model demonstrated a substantial improvement in its ability to extract and retain crucial features. This is particularly relevant for medical images, where capturing fine anatomical details is critical for accurate diagnosis. The ESA mechanism also played a vital role in improving the overall performance of the model by allowing it to focus more effectively on relevant regions of the input image, further boosting the quality of the super-resolved output. The integration of ESA proved to be highly beneficial in addressing one of the key challenges in medical imaging: distinguishing between essential anatomical features and noise. Medical images often contain regions with subtle differences that are critical for diagnosis. By incorporating ESA, the model was able to prioritize these regions, resulting in super-resolved images that not only had higher visual fidelity but also retained diagnostically important details. This is particularly useful in applications such as MRI and CT scans, where image clarity directly impacts clinical outcomes. The visual quality improvements provided by the proposed model have significant clinical implications, particularly in enhancing diagnostic accuracy. The enhanced resolution achieved by our method can lead to better detection of small pathologies, which are often challenging to identify in lower-resolution images. For instance, with CT scans, the model's ability to preserve fine anatomical details can aid in the early detection of small lesions, which is critical for early-stage diagnosis and timely intervention. Similarly, in ultrasound imaging, a higher resolution can improve the visibility of subtle abnormalities, such as small kidney stones or early-stage tumors, leading to more accurate and confident diagnoses. By refining the visual quality of medical images, our model has the potential to support radiologists and other healthcare professionals in identifying pathologies that may otherwise go unnoticed. This enhancement not only improves diagnostic accuracy but also contributes to more effective treatment planning and patient outcomes. These clinical implications underscore the value of our approach beyond mere image quality improvement, highlighting its practical utility in supporting healthcare professionals with precise and reliable image enhancement tools.

Despite the promising results, there are several limitations to the current approach that warrant further investigation. First, while the proposed model shows significant improvement over existing methods, the computational complexity remains relatively high. The increased model depth and the incorporation of attention mechanisms add to the computational load, which could limit the model's applicability in real-time clinical settings or in scenarios with limited computational resources. Future work could focus on optimizing the model to reduce inference time and computational requirements without compromising image quality. Another limitation is the reliance on specific datasets for training and evaluation. Although the proposed model performed well on the medical imaging datasets used in this study, the generalizability of the model to other types of medical images or imaging modalities needs to be further explored. Medical imaging data can vary significantly based on factors such as imaging equipment, acquisition protocols, and patient demographics. To ensure robustness and broad applicability, future research should include more diverse datasets and evaluate the model's performance across them.

6. Conclusions

In this paper, we presented an enhanced version of the RFLN specifically designed for single image super-resolution in medical imaging. By incorporating deeper residual blocks, ESA mechanisms, and increasing the model depth, the proposed architecture demonstrated significant improvements in reconstructing high-resolution medical images from their low-resolution counterparts. Our experimental results showed that the enhanced RFLN outperformed existing state-of-the-art models in terms of both quantitative metrics, such as PSNR and SSIM, and qualitative visual assessments. The integration of ESA played a crucial role in focusing on relevant anatomical features while suppressing noise, leading to better diagnostic utility of the super-resolved images. These enhancements make the proposed model particularly well-suited for medical applications, where image quality directly impacts diagnostic accuracy and patient outcomes. However, the increased computational

complexity and reliance on specific datasets present challenges that must be addressed in future work. Optimizing the model for real-time clinical deployment and expanding its applicability to diverse medical imaging modalities are essential next steps. By addressing these limitations, we believe that the proposed model can make a significant impact in the field of medical imaging, providing healthcare professionals with the tools needed for more precise and effective diagnoses.

Author Contributions: Methodology, A.A., S.U. and Y.-I.C.; software, A.A., S.U., Z.D., K.Z., S.M. and R.N.; validation, Z.D., K.Z., S.M. and R.N.; formal analysis, A.A., S.U., Z.D., S.M., K.Z. and R.N.; resources, Z.D., K.Z., Y.-I.C. and R.N.; data curation, Z.D., S.M., K.Z., Y.-I.C. and R.N.; writing—original draft, A.A., S.U., Z.D., K.Z., S.M., R.N. and Y.-I.C.; writing—review and editing, A.A., S.U., Z.D., K.Z., S.M., R.N. and Y.-I.C.; Supervision, Y.-I.C., A.A. and S.U.; project administration, Z.D., K.Z., S.M., R.N. and Y.-I.C. All authors have read and agreed to the published version of the manuscript.

Funding: This study was funded by the Korea Agency for Technology and Standards in 2022. The project numbers are 1415180835 (Development of International Standard Technologies based on AI Learning and Inference Technologies), 1415181629 (Development of International Standard Technologies based on AI Model Lightweighting Technologies), and 1415181638 (Establishment of standardization basis for BCI and AI Interoperability). Any correspondence related to this paper should be addressed to Young-Im Cho.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The original contributions presented in the study are included in the article, further inquiries can be directed to the corresponding author/s.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Lambert, B.; Forbes, F.; Doyle, S.; Dehaene, H.; Dojat, M. Trustworthy clinical AI solutions: A unified review of uncertainty quantification in deep learning models for medical image analysis. *Artif. Intell. Med.* **2024**, *150*, 102830. [CrossRef] [PubMed]
2. Bakhtiarnia, A.; Zhang, Q.; Iosifidis, A. Efficient high-resolution deep learning: A survey. *ACM Comput. Surv.* **2024**, *56*, 181. [CrossRef]
3. Dong, C.; Loy, C.C.; He, K.; Tang, X. Image super-resolution using deep convolutional networks. *IEEE Trans. Pattern Anal. Mach. Intell.* **2015**, *38*, 295–307. [CrossRef]
4. El-Shafai, W.; El-Nabi, S.A.; Ali, A.M.; El-Rabaie, E.S.M.; Abd El-Samie, F.E. Traditional and deep-learning-based denoising methods for medical images. *Multimed. Tools Appl.* **2024**, *83*, 52061–52088. [CrossRef]
5. Kong, F.; Li, M.; Liu, D.; He, J.; Bai, Y.; Chen, F.; Fu, L. Residual Local Feature Network for Efficient Super-Resolution. In Proceedings of the 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), New Orleans, LA, USA, 18–24 June 2022; pp. 766–776.
6. Ferdi, A.; Benierbah, S.; Nakib, A. Residual encoder-decoder based architecture for medical image denoising. *Multimed. Tools Appl.* **2024**, 1–18. [CrossRef]
7. Umirzakova, S.; Ahmad, S.; Khan, L.U.; Whangbo, T. Medical image super-resolution for smart healthcare applications: A comprehensive survey. *Inf. Fusion* **2023**, *103*, 102075. [CrossRef]
8. Liu, J.; Gan, Z.; Zhu, X. Directional Bicubic Interpolation—A New Method of Image Super-Resolution. In Proceedings of the 3rd International Conference on Multimedia Technology (ICMT-13), Guangzhou, China, 29 November–1 December 2013; pp. 463–470.
9. Wang, Y.H.; Qiao, J.; Li, J.B.; Fu, P.; Chu, S.C.; Roddick, J.F. Sparse representation-based MRI super-resolution reconstruction. *Measurement* **2014**, *47*, 946–953. [CrossRef]
10. Liu, D.; Huang, T.S. Single Image Super-Resolution: From Sparse Coding to Deep Learning. In *Deep Learning Through Sparse and Low-Rank Modeling*; Wang, Z., Fu, Y., Huang, T.S., Eds.; Elsevier: Cambridge, MA, USA, 2019; pp. 47–86.
11. Mishra, D.; Majhi, B.; Sa, P.K.; Dash, R. Development of robust neighbor embedding based super-resolution scheme. *Neurocomputing* **2016**, *202*, 49–66. [CrossRef]
12. Kim, J.; Lee, J.K.; Lee, K.M. Accurate Image Super-Resolution Using Very Deep Convolutional Networks. In Proceedings of the 2016 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 27–30 June 2016; pp. 1646–1654.
13. Kim, J.; Lee, J.K.; Lee, K.M. Deeply-Recursive Convolutional Network for Image Super-Resolution. In Proceedings of the 2016 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 27–30 June 2016; pp. 1637–1645.
14. Tian, C.; Zhang, X.; Lin, J.C.W.; Zuo, W.; Zhang, Y.; Lin, C.W. Generative adversarial networks for image super-resolution: A survey. *arXiv* **2022**, arXiv:2204.13620.

15. Ledig, C.; Theis, L.; Huszár, F.; Caballero, J.; Cunningham, A.; Acosta, A.; Aitken, A.; Tejani, A.; Totz, J.; Wang, Z.; et al. Photo-Realistic Single Image Super-Resolution Using a Generative Adversarial Network. In Proceedings of the 2017 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 21–26 July 2017; pp. 4681–4690.
16. Abdusalomov, A.B.; Nasimov, R.; Nasimova, N.; Muminov, B.; Whangbo, T.K. Evaluating synthetic medical images using artificial intelligence with the GAN algorithm. *Sensors* **2023**, *23*, 3440. [CrossRef]
17. Rasheed, Z.; Ma, Y.K.; Ullah, I.; Ghadi, Y.Y.; Khan, M.Z.; Khan, M.A.; Abdusalomov, A.; Alqahtani, F.; Shehata, A.M. Brain tumor classification from MRI using image enhancement and convolutional neural network techniques. *Brain Sci.* **2023**, *13*, 1320. [CrossRef] [PubMed]
18. Lim, B.; Son, S.; Kim, H.; Nah, S.; Mu Lee, K. Enhanced Deep Residual Networks for Single Image Super-Resolution. In Proceedings of the 2017 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 21–26 July 2017; pp. 136–144.
19. Zhang, Y.; Li, K.; Li, K.; Wang, L.; Zhong, B.; Fu, Y. Image Super-Resolution Using Very Deep Residual Channel Attention Networks. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018; pp. 286–301.
20. Liu, Y.; Yang, D.; Zhang, F.; Xie, Q.; Zhang, C. Deep recurrent residual channel attention network for single image super-resolution. *Vis. Comput.* **2024**, *40*, 3441–3456. [CrossRef]
21. Abdusalomov, A.; Umirzakova, S.; Safarov, F.; Mirzakhilov, S.; Egamberdiev, N.; Cho, Y.I. A Multi-Scale Approach to Early Fire Detection in Smart Homes. *Electronics* **2024**, *13*, 4354. [CrossRef]
22. Yu, M.; Shi, J.; Xue, C.; Hao, X.; Yan, G. A review of single image super-resolution reconstruction based on deep learning. *Multimed. Tools Appl.* **2024**, *83*, 55921–55962. [CrossRef]
23. El-Shafai, W.; Ali, A.M.; El-Nabi, S.A.; El-Rabaie, E.S.M.; Abd El-Samie, F.E. Single image super-resolution approaches in medical images based-deep learning: A survey. *Multimed. Tools Appl.* **2024**, *83*, 30467–30503. [CrossRef]
24. Liu, H.; Li, Z.; Shang, F.; Liu, Y.; Wan, L.; Feng, W.; Timofte, R. Arbitrary-scale super-resolution via deep learning: A comprehensive survey. *Inf. Fusion* **2024**, *102*, 102015. [CrossRef]
25. Feng, C.M.; Yan, Y.; Yu, K.; Xu, Y.; Fu, H.; Yang, J.; Shao, L. Exploring separable attention for multi-contrast MR image super-resolution. *IEEE Trans. Neural Netw. Learn. Syst.* **2024**, *35*, 12251–12262. [CrossRef]
26. Ayaz, A.; Boonstoppel, R.; Lorenz, C.; Weese, J.; Pluim, J.; Breeuwer, M. Effective deep-learning brain MRI super resolution using simulated training data. *Comput. Biol. Med.* **2024**, *183*, 109301. [CrossRef]
27. You, C.; Dai, W.; Liu, F.; Min, Y.; Dvornek, N.C.; Li, X.; Clifton, D.A.; Staib, L.; Duncan, J.S. Mine your own anatomy: Revisiting medical image segmentation with extremely limited labels. *IEEE Trans. Pattern Anal. Mach. Intell.* **2024**, *46*, 11136–11151. [CrossRef]
28. Wang, C.; Lv, X.; Shao, M.; Qian, Y.; Zhang, Y. A novel fuzzy hierarchical fusion attention convolution neural network for medical image super-resolution reconstruction. *Inf. Sci.* **2023**, *622*, 424–436. [CrossRef]
29. Georgescu, M.I.; Ionescu, R.T.; Miron, A.I.; Savencu, O.; Ristea, N.C.; Verga, N.; Khan, F.S. Multimodal Multi-Head Convolutional Attention with Various Kernel Sizes for Medical Image Super-Resolution. In Proceedings of the 2023 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), Waikoloa, HI, USA, 2–7 January 2023; pp. 2195–2205.
30. He, C.; Li, K.; Xu, G.; Yan, J.; Tang, L.; Zhang, Y.; Wang, Y.; Li, X. Hqg-net: Unpaired medical image enhancement with high-quality guidance. *IEEE Trans. Neural Netw. Learn. Syst.* **2023**; early access.
31. Umirzakova, S.; Mardieva, S.; Muksimova, S.; Ahmad, S.; Whangbo, T. Enhancing the Super-Resolution of Medical Images: Introducing the Deep Residual Feature Distillation Channel Attention Network for Optimized Performance and Efficiency. *Bioengineering* **2023**, *10*, 1332. [CrossRef] [PubMed]
32. Huang, W.; Liao, X.; Chen, H.; Hu, Y.; Jia, W.; Wang, Q. Deep local-to-global feature learning for medical image super-resolution. *Comput. Med. Imaging Graph.* **2024**, *115*, 102374. [CrossRef] [PubMed]
33. He, J.; Ma, H.; Guo, M.; Wang, J.; Wang, Z.; Fan, G. Research into super-resolution in medical imaging from 2000 to 2023: Bibliometric analysis and visualization. *Quant. Imaging Med. Surg.* **2024**, *14*, 5109. [CrossRef] [PubMed]
34. Baranwal, N.; Singh, K.N.; Singh, A.K. YOLO-based ROI selection for joint encryption and compression of medical images with reconstruction through super-resolution network. *Future Gener. Comput. Syst.* **2024**, *150*, 1–9.
35. Hui, Z.; Gao, X.; Yang, Y.; Wang, X. Lightweight Image Super-Resolution with Information Multi-Distillation Network. In Proceedings of the 27th ACM International Conference on Multimedia, Nice, France, 21–25 October 2019; pp. 2024–2032.
36. Liu, J.; Tang, J.; Wu, G. Residual Feature Distillation Network for Lightweight Image Super-Resolution. In Proceedings of the 2010 Edition of the European Conference on Computer Vision, Glasgow, UK, 23–28 August 2020; pp. 41–55.
37. Ahn, N.; Kang, B.; Sohn, K.A. Fast, Accurate, and Lightweight Super-Resolution with Cascading Residual Network. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018; pp. 252–268.
38. Lai, W.S.; Huang, J.B.; Ahuja, N.; Yang, M.H. Deep Laplacian Pyramid Networks for Fast and Accurate Super-Resolution. In Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 21–26 July 2017; pp. 624–632.

Disclaimer/Publisher’s Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.

Article

Improving Radiology Report Generation Quality and Diversity through Reinforcement Learning and Text Augmentation

Daniel Parres ^{1,*}, Alberto Albiol ¹ and Roberto Paredes ^{1,2}

¹ Campus de Vera, Universitat Politècnica València, Camí de Vera s/n, 46022 Valencia, Spain; alalbiol@iteam.upv.es (A.A.); rparedes@prhlt.upv.es (R.P.)

² Valencian Graduate School and Research Network of Artificial Intelligence, Camí de Vera s/n, 46022 Valencia, Spain

* Correspondence: dparres@prhlt.upv.es

Abstract: Deep learning is revolutionizing radiology report generation (RRG) with the adoption of vision encoder–decoder (VED) frameworks, which transform radiographs into detailed medical reports. Traditional methods, however, often generate reports of limited diversity and struggle with generalization. Our research introduces reinforcement learning and text augmentation to tackle these issues, significantly improving report quality and variability. By employing RadGraph as a reward metric and innovating in text augmentation, we surpass existing benchmarks like BLEU4, ROUGE-L, F1CheXbert, and RadGraph, setting new standards for report accuracy and diversity on MIMIC-CXR and Open-i datasets. Our VED model achieves F1-scores of 66.2 for CheXbert and 37.8 for RadGraph on the MIMIC-CXR dataset, and 54.7 and 45.6, respectively, on Open-i. These outcomes represent a significant breakthrough in the RRG field. The findings and implementation of the proposed approach, aimed at enhancing diagnostic precision and radiological interpretations in clinical settings, are publicly available on GitHub to encourage further advancements in the field.

Keywords: radiology report generation; reinforcement learning; text augmentation; machine learning; deep learning; vision transformer; chest X-rays; medical image; text generation

1. Introduction

Radiology report generation (RRG) is a challenging task where the goal is to interpret radiographic images and generate detailed reports on potential patient pathologies. In contrast to typical computer vision (CV) tasks, which aim to identify objects in images, RRG focuses on diagnosing potential pathologies and determining their presence, absence, or uncertainty. Moreover, limited data availability and the diverse nature of medical reports pose significant challenges. This task is crucial for streamlining and enhancing patient care, as it enables the analysis of individuals' health and the rapid detection of diseases. Furthermore, it serves as assistance and support for medical professionals. The RRG task parallels other works such as [1], where complex data are handled, and models trained from such data must ensure specific security criteria. This similarity arises from the need to develop models capable of diagnosing pathologies precisely, as in this type of problem, patients' health is put at risk. Current approaches mainly rely on deep learning (DL), specifically the vision encoder–decoder (VED) architecture [2–8], incorporating components like memories [9,10] or reinforcement learning (RL) [5,11,12] to improve performance.

This study presents a two-stage VED pure-transformer architecture for chest RRG. In the first stage, conventional negative log-likelihood (NLL) training is employed, while the second stage focuses on RL optimization for various metrics. These metrics encompass embedding comparison for semantic coherence [13], entity and relationship graph generation for pathology descriptions [14], and NLL. Additionally, we propose the integration of text augmentation and hard negative mining techniques. This comprehensive approach sur-

passes current chest RRG methodologies, enhancing report variability, factual correctness, completeness, and overall model generalization.

The key contributions of this study are as follows:

- Harnessing vision encoder–decoder transformer frameworks and reinforcement learning to enhance radiology report quality, factual correctness, and completeness.
- Proposing text augmentation to the training workflow. This technique allows for leveraging the scarcity of data by generating new reports. This leads to an improvement in diversity, averaging a threefold increase compared to the state of the art.
- Due to the challenge of measuring the quality of generated reports, we focus on employing specialized natural language processing-based metrics to evaluate radiology reports comprehensively.
- Open-sourcing our methodology on GitHub at <https://github.com/dparres/Diversifying-Radiology-Report-Generation> (accessed on 1 April 2024) to propel progress in diagnostic precision and radiological interpretation within clinical settings.

Related Work

In recent years, deep learning techniques have led various approaches to classify and detect pathologies in X-ray images. Notable efforts in this domain include Schlegl et al. [15], who proposed a convolutional network (CNN) for classifying tissue patterns in tomographies, utilizing semantic descriptions in reports as labels. Building on this success, subsequent neural network models were explored for X-rays, such as Shin et al. [16], who introduced a CNN for chest X-ray images and a recurrent network (RNN) for annotations, jointly trained to annotate diseases, anatomy, and severity. Other approaches, like that of Moradi et al. [17], focused on annotation through the concatenation of a CNN and an RNN block to identify regions of interest. Notably, Rubin et al. [18] employed parallelly trained CNNs for frontal and lateral chest X-ray views to estimate possible pathologies. The interest in RRG has grown, led by impactful models like TieNet [4], as well as innovative contributions from Li et al. [3], Jing et al. [19], and Jing et al. [19], each leveraging advanced techniques such as pretrained networks, coattention mechanisms, and retrieval policy modules for efficient disease classification and report generation.

Deep learning applications in radiology encounter obstacles due to limited and unstructured data availability, exacerbating the normal–abnormal case imbalance and complicated by ambiguous radiologist reports [20]. CNN models such as AlexNet [21], VGG-16 [22], GoogLeNet [23], and ResNet [24] dominate medical text–image mining, with the increasing prevalence of end-to-end training [20]. However, the model comparison is hindered by dataset diversity, although incorporating radiology reports is expected to grow [20]. Radiology reports, diverse in style and influenced by biases, pose challenges in training robust models, with few datasets meeting scalability and accessibility criteria [25]. Most RRG systems concentrate on X-ray tests, with emerging applications for CT scans and MRI datasets, each encountering unique challenges [25].

Alfarghaly et al. [26] introduced a novel technique for generating radiology reports from chest X-rays with DistilGPT2 [27]. This approach highlights faster training and improved metrics. Although limitations persist, larger datasets' release is suggested for enhanced model generalization and critical evaluation of quantitative methods [26]. Yang et al. [28] presented a method combining general medical and specific medical knowledge with a multihead attention mechanism for chest RRG. Furthermore, Pan et al. [29] proposed a method for generating chest radiology reports through cross-modal multiscale feature fusion. This architecture aims to enhance location sensitivity and disease characterization. Additionally, these advancements pave the way for aligning scales and integrating knowledge graphs, enhancing the accuracy of report generation [29]. Yang et al. [30] introduced a highly accurate and automated radiology generation framework coupled with a novel automatic medical knowledge updating mechanism, enhanced by a multimodal alignment approach. Zhao et al. [31] presented a knowledge enhancement technique leveraging medical knowledge in dictionary form alongside historical knowledge, comple-

mented by a multilevel alignment method to mitigate modal differences between text and image. Nicolson et al. [32] introduced CvT2DistilGPT2, leveraging a convolutional vision transformer for optimal encoder warm starting and highlighting GPT2's [33] superiority in decoder warm starting over BERT [34]. Despite the proposals' novelty, they all rely on metrics such as BLEU and ROUGE-L for comparison, which may not be entirely suitable for measuring the presence, absence, or uncertainty of pathologies. Furthermore, these metrics fail to consider the interrelation of pathologies within the report. Additionally, the monotony and repetition in generated reports are noteworthy factors attributed to the inherent difficulty of the RRG task.

Due to the increasing interest in automating RRG, specific architectures have been designed to address chest RRG tasks. Liu et al. [5] proposed a CNN for extracting visual features, followed by a sentence decoder and word decoder for generating report topics and composing the final report. The model is fine-tuned using RL with CheXpert [35] labels. Chen et al. [9] introduced the memory-driven transformer, employing relational and memory-driven layers to enhance information retention and incorporation into the decoder. Meanwhile, Chen et al. [10] presented a VED based on cross-model memory networks, leveraging shared memory to capture and utilize visual–textual alignments.

Liu et al. [6] and Liu et al. [7] proposed leveraging unsupervised construction of knowledge graphs to replicate radiologists' patterns to generate reports. Focusing on RL, Miura et al. [11] suggested employing two metrics: one ensuring the generation of domain entities (fact_{ENT}) and the other maintaining coherent entity descriptions (fact_{ENTNLI}). These metrics are optimized alongside BERTScore [13], a semantic equivalence metric. Delbrouck et al. [12] propose a competitive chest RRG approach utilizing a VED. The model employs a DenseNet-121 [36] as the optical encoder and a single-layer BERT [34] as the decoder. Trained with RL, the optimization involves three metrics: RadGraph [14], BERTScore, and NLL. RadGraph, a neural network, constructs semantic annotations by forming a graph of entities and relationships in reports. The reward is computed by comparing the hypotheses' entities and relationships against the reference, resulting in higher-quality reports surpassing values achieved by fact_{ENT} and fact_{ENTNLI} . These proposals employ chest-RRG-oriented metrics to enable more precise comparisons of report quality compared to metrics such as BLEU or ROUGE-L. Despite this and using RL for training, the reports exhibit repetitiveness and a lack of diversity.

2. Materials and Methods

This section of the paper presents the primary approach for the chest RRG task. Firstly, the neural models are introduced, and we propose four different architectures to analyze the most suitable for the RRG task. Subsequently, the training workflow of the models is outlined. Then, the text augmentation technique is introduced to leverage the scarcity of data. The final subsection presents the databases to be utilized and the metrics for comparing models and measuring the quality of the generated reports.

2.1. Proposed Architectures

Our model architectures for addressing the chest RRG problem are VED models, with Swin [37] as the vision model and BERT [34] as the language model or decoder. We opted for Swin for its outstanding performance in various computer vision tasks such as classification, object detection, and semantic segmentation [37]. Compared to ViT [38], Swin's hierarchical nature and more efficient attention mechanism based on shifted windows make it more suitable for X-ray analysis. Our models utilize both the base and small variants, denoted as SwinB and SwinS. These architectures are pretrained with ImageNet [39] and feature an input image size of three channels with a width and height of 224 pixels. They have architectural depths of 2, 2, 18, and 2 for each layer, with a patch size of 4. The primary difference lies in the number of heads, with SwinB utilizing sizes 4, 8, 16, and 32 and SwinS using 3, 6, 12, and 24.

For the decoder model, BERT was selected due to its proven effectiveness across various natural language processing (NLP) tasks [34,40,41]. However, we diverge from the standard configuration by employing only three layers with a hidden size of 1024, rather than the original 12 layers. Moreover, our investigation includes the analysis of two distinct decoders: one word-based, with a vocabulary of approximately 9.8k words, and another subword-based, featuring a vocabulary of 30k subwords. The integration of the vision model with the decoder model is facilitated through cross-attention layers, enabling the introduction of visual features extracted from the radiograph into the decoder. This integration allows for the autoregressive generation of reports.

Following this approach, the proposed VED models for chest RRG analysis are SwinB+BERT9k, SwinB+BERT30k, SwinS+BERT9k, and SwinS+BERT30k.

2.2. Training Workflows

In radiological contexts, multiple images (such as anterior–posterior and lateral views) are commonly generated per patient study. We employ a multi-image approach to address this, restricting it to three images per study. This involves concatenating features extracted by the Swin encoder for each image and processing them in the decoder. Furthermore, we apply random transformations—including translation, scaling, rotation, and adjustments to brightness and contrast—to augment training images.

Our training process comprises two stages, as illustrated in Figure 1. In the initial stage, the NLL loss function is employed, utilizing teacher forcing to train the model based on the generated hypothesis and the reference report. This stage typically encompasses approximately twelve training epochs. Subsequently, in the second stage, RL is introduced using the self-critical sequence training (SCST) algorithm [42]. SCST performs two forward passes of the VED: greedy decoding and beam-search multinomial sampling (sampling decoding). The first pass involves inference through greedy decoding to obtain the report Y_g without calculating gradients. Meanwhile, the second pass employs sampling decoding, calculating gradients to obtain the sampling report Y_s .

RL aims to optimize specific metrics by using them as rewards. In our approach, we propose utilizing two metrics: BERTScore [13] and RadGraph [14]. BERTScore effectively improves grammar and semantics within models, while F_1RG_{ER} aids the model in prioritizing pathology-related entities and relationships.

After obtaining the reports Y_g and Y_s , we must calculate a loss for each metric to optimize ($Loss_{metric}$). This $Loss_{metric}$ is subsequently used to compute a weighted sum to obtain the final $Loss_{RL}$ and update the model weights. Equation (1) outlines the process to derive the $Loss_{metric}$, where the metric rewards are computed for Y_s and Y_g using the reference report Y_{ref} . These rewards are then subtracted and multiplied by the logarithm of the probabilities calculated during the generation of the Y_s report.

$$Loss_{metric}(Y_s, Y_g, Y_{ref}) = -(r_{metric}(Y_s, Y_{ref}) - r_{metric}(Y_g, Y_{ref})) \log(\Pr(Y_s)) \quad (1)$$

Each $Loss_{metric}$ contributes as a distinct weighted term to the final RL loss, as presented in Equation (2). To obtain the $Loss_{RL}$ for updating the model weights, $Loss_{BERTScore}$, $Loss_{F_1RG_{ER}}$, and the same loss utilized in the initial training stage, the $Loss_{NLL}$, are summed. In our experiments, we set $\alpha = \beta = 0.495$ and $\gamma = 0.010$ as the weighting factors.

$$Loss_{RL} = \alpha Loss_{BERTScore} + \beta Loss_{F_1RG_{ER}} + \gamma Loss_{NLL} \quad (2)$$

2.3. Text Augmentation

Data augmentation (DA) is the most prevalent technique for addressing machine learning problems with limited data. This approach significantly contributes to enhanced generalization and improved data utilization. However, regarding RRG, the focus has primarily been on augmenting input images. To the best of our knowledge, no prior approach has investigated the impact of augmenting radiology reports, a technique known as text augmentation (TA). In chest RRG, reports are often drafted by different individuals

in diverse manners, lacking specific templates, which results in significant variability inherent to natural language. Presently, state-of-the-art RRG algorithms encounter this data scarcity issue, often generating highly repetitive and monotonous reports.

In this study, we introduce a TA technique designed to enhance the quality and diversity of generated reports. Our TA approach involves splitting the reference report into phrases, considering the typical structure of reports with multiple sentences. These phrases are then randomly reorganized to create new reports while maintaining the original diagnosis, as depicted in Figure 1. This method mitigates overfitting concerns and promotes a more targeted learning of the mentioned pathologies. Furthermore, our TA significantly improves the quality, factual accuracy, and comprehensiveness of the generated reports, as demonstrated subsequently.

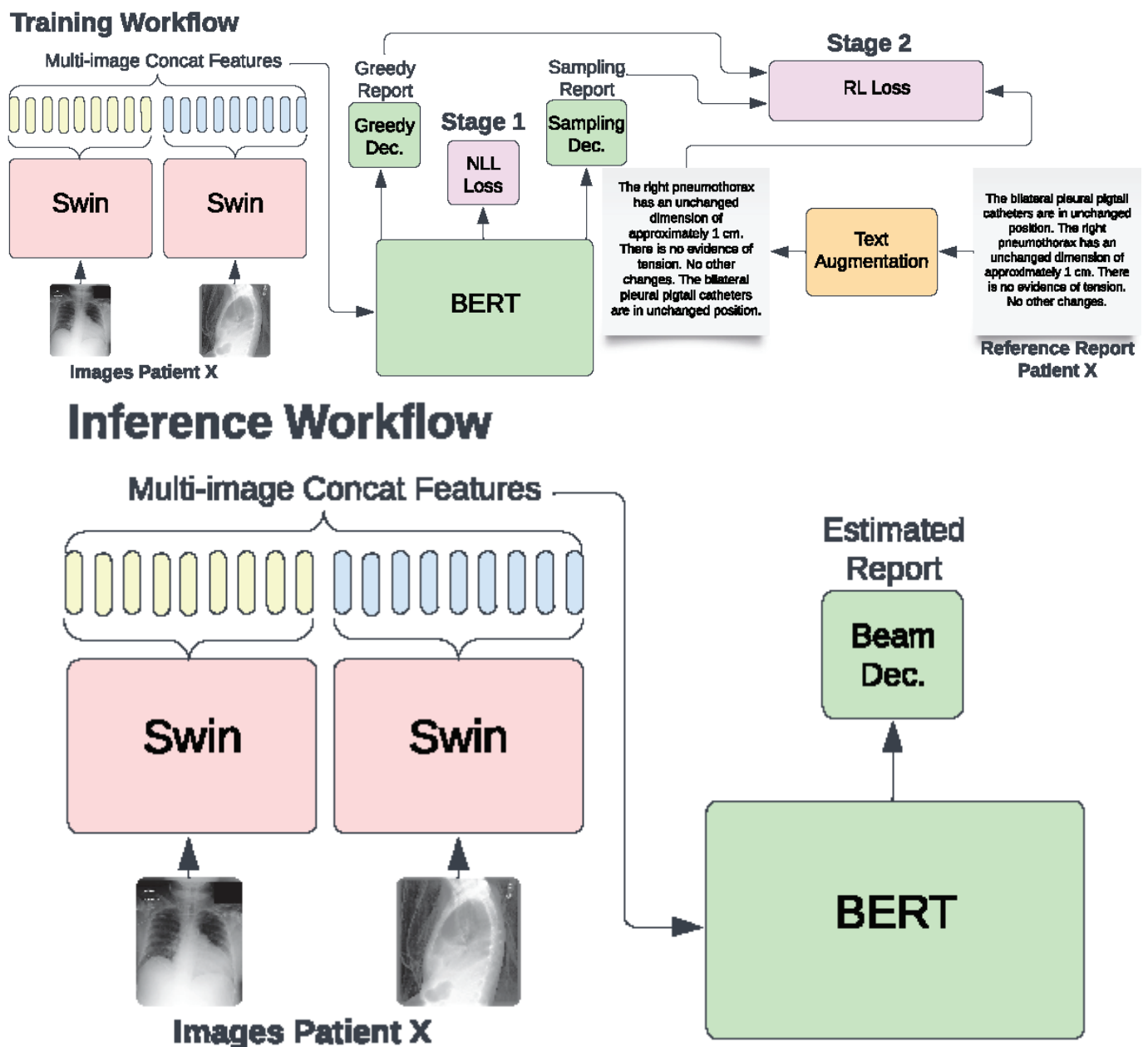


Figure 1. Our training workflow using RL and TA to enhance report quality. In the initial stage, training with NLL employs teacher forcing. Subsequently, during RL training, the SCST algorithm computes rewards utilizing two different reports from two distinct decoding strategies: greedy search and beam-search multinomial sampling. Inference on validation and test sets is conducted using beam search in both stages.

The TA technique described and utilized in this work is exclusively employed in the training workflow, specifically during the NLL and RL stages. This technique aims to leverage the reference reports to enhance the model training process and increase its robustness. The report sentences will be randomly ordered in each training epoch for each training sample.

2.4. Datasets and Metrics

To evaluate the competitiveness of our models, we employed two publicly available datasets: MIMIC-CXR [43] and Open-i [44]. Due to the relatively small size of the Open-i dataset, containing 3.3 k reports, it was exclusively utilized for testing. The MIMIC-CXR dataset comprises 152 k reports for training, 2.3 k for validation, and 2.3 k for testing. The workflow involves training with the 152 k reports, saving the model weights based on the best performance on the validation set, and finally using these weights to perform inference on both the MIMIC-CXR and Open-i test sets. It is important to note that in this study, the models were specifically designed to generate the findings section of the reports. Additionally, samples with empty findings sections were excluded from consideration.

Metrics play a crucial role in comparing models, with BLEU4 [45] and ROUGE-L [46] currently considered as the most widely used natural language generation (NLG)-oriented metrics for reports. Additionally, two chest-RRG-oriented metrics, specifically designed to evaluate the quality of radiology reports, have been employed: $F_1\text{CheXbert}$ ($F_1\text{cXb}$) [47] and $F_1\text{RG}_{ER}$ [12,14]. These metrics leverage neural networks to assess the quality of generated reports, offering a higher semantic evaluation. $F_1\text{cXb}$ utilizes CheXbert, a transformer model capable of identifying the presence of the 14 CheXpert [48] pathologies (atelectasis, cardiomegaly, consolidation, edema, enlarged cardiomeastinum, fracture, lung lesion, lung opacity, pleural effusion, pneumonia, pneumothorax, pleural other, support devices, and no finding) in a hypothesis report, and calculates the F_1 -score based on the pathologies present in the reference report. On the other hand, $F_1\text{RG}_{ER}$ employs a transformer model to analyze reports, generating graphs for entities and relationships. It evaluates comparisons between entity and relationship structures in hypotheses and reference graphs to compute the F_1 -score.

3. Results

This section details the experiments carried out to develop models for chest RRG. We analyze the proposed architectures and then conduct an ablation study on effective training techniques and strategies. Finally, we assess the competitiveness of our proposal by comparing it with state-of-the-art models. All experiments were conducted using a computer equipped with an NVIDIA RTX 4090 GPU.

3.1. Analysis of Our Architectures

This study proposes four VED models, distinguished primarily by the number of trainable parameters in each component. Table 1 illustrates how SwinB+BERT30K is the largest model, while SwinS+BERT9k is the smallest. The remaining two models fall within a similar parameter range. Comparing architecture sizes is interesting because the RRG task operates in a data-limited environment due to the high variability of medical reports and their scarcity. Therefore, training models with high parameter counts can become challenging to tune efficiently. The number of parameters, along with the nature and size of the dataset, are crucial factors for effective learning that avoids underfitting and overfitting. Furthermore, the results presented in Table 1 encourage analysis of which components of the architectures may be more critical for obtaining meaningful reports.

In addition to the number of parameters, Table 1 showcases BLEU4, ROUGE-L, $F_1\text{cXb}$, and $F_1\text{RG}_{ER}$ metrics. These metrics were obtained following the same workflow and hyperparameters across all four models. During the first stage, involving training with NLL, we conduct training for 12 epochs with a learning rate of 3×10^{-4} . Subsequently, in the second stage, corresponding to RL training, we train for 15 epochs with a learning

rate of 5×10^{-5} . This learning rate value is crucial as it allows a good starting point for improving results based on previous training. The value of learning rates has been empirically explored using grid search with learning rate values of 5×10^{-3} , 3×10^{-3} , 1×10^{-3} , 5×10^{-4} , 3×10^{-4} , 1×10^{-4} , 5×10^{-5} , 3×10^{-5} , 1×10^{-5} , 5×10^{-6} , 3×10^{-6} , and 1×10^{-6} for the first and second training stages. Values above 5×10^{-5} in RL training lead to highly repetitive reports, while values below it yield only marginal improvements, failing to achieve significantly better results than training with NLL. The training epochs are set to 12 and 15 for the first and second stages, respectively, proving sufficient to achieve satisfactory results. Increasing the epochs does not significantly improve report quality due to the task's data scarcity and variability. During both stages, the learning rate is linearly decreased until reaching a value of zero at the end of the epochs. The results presented in Table 1 do not utilize image data augmentation, hard negative mining (HNM), or text augmentation.

Table 1. Architecture comparison using four report quality metrics on the MIMIC-CXR and Open-i test sets, RL training hours per epoch, and the number of parameters. The metric values are provided in (%). Bold font indicates the best result obtained for each metric.

Model	F ₁ cXb	F ₁ RG _{ER}	BLEU4	ROUGE-L	h/epoch	# Params
MIMIC-CXR						
SwinS+BERT9k RL	61.1	34.9	11.5	25.9	12	109 M
SwinS+BERT30k RL	61.4	34.3	11.3	25.2	24	130 M
SwinB+BERT9k RL	63.1	35.4	11.7	26.5	14	147 M
SwinB+BERT30k RL	62.9	35.1	11.6	26.5	30	168 M
Open-i						
SwinS+BERT9k RL	52.8	44.5	14.8	33.7	-	109 M
SwinS+BERT30k RL	52.8	44.1	14.6	33.4	-	130 M
SwinB+BERT9k RL	54.7	45.6	15.1	34.3	-	147 M
SwinB+BERT30k RL	54.4	45.3	14.9	34.1	-	168 M

Regarding report metrics on MIMIC-CXR, the SwinB+BERT9k model stands out, exhibiting superior performance compared to its closest counterpart, SwinB+BERT30k, and achieving a training time reduction of 16 h per epoch. For the SwinS models, results demonstrate a similarity, albeit with the BERT9k variant showcasing marginally better metrics across all parameters except F₁cXb. Interestingly, F₁cXb remains nearly identical between the two SwinS and SwinB models, indicating a direct impact of the encoder's size on this metric's competitiveness. Conversely, the choice between word-based or subword-based decoders does influence metrics such as BLEU4, ROUGE-L, and F₁RG_{ER}. Moreover, adopting word-based models yields superior metrics and significantly reduces training time, enhancing overall efficiency. The results from Open-i also rank the SwinB+BERT9k model as the most competitive and SwinS+BERT30k as the least competitive. Once again, this reaffirms that word-based models yield better results in the chest RRG task. Furthermore, the SwinS model demonstrates superior capability in generating reports by obtaining the best representations of radiographs.

Based on the results of the analysis, SwinB+BERT9k emerges as the most competitive and efficient model for chest RRG. However, various strategies can prove crucial in data-limited environments like chest RRG. This study specifically proposes using TA to leverage reports and significantly increase the training data. However, employing image augmentation techniques such as slight rotations, scale changes, shifts, random crops, and adjustments of brightness and contrast helps exploit the number of images. Moreover, since most datasets typically include more challenging samples than others, HNM is an option to consider. During training, samples with errors greater than the mean error plus their standard deviation are reutilized before moving to the next epoch.

Table 2 presents the ablation study of our best model, considering these strategies. The first row showcases the metrics achieved at the end of the NLL stage. The second row demonstrates how training with RL succeeds in improving these metrics; while image data augmentation shows a minor impact attributed to the robustness of transformers, HNMM marginally enhances results by revisiting challenging samples at the end of each epoch. In contrast, TA emerges as a crucial technique, significantly improving chest-RRG-oriented metrics and BLEU4 by over one point. This underscores the significance and effectiveness of TA in efficiently leveraging the number of reports and furnishing the network with enhanced generalization capabilities.

Table 2. Ablation study of our best model on the MIMIC-CXR test set. The metric values are provided in (%). Bold font indicates the best result obtained for each metric.

Model	F ₁ cXb	F ₁ RG _{ER}	BLEU4	ROUGE-L
SwinB+BERT9k NLL	52.3	22.4	10.1	20.5
+ RL	63.1	35.4	11.7	26.5
+ Image Augment.	63.3	35.7	11.9	26.6
+ Hard Neg. Mining	64.2	35.8	12.0	26.8
+ Text Augment.	66.2	37.8	13.2	27.1

3.2. Benchmarking with the State of the Art

To assess the competitiveness of our top model against other state-of-the-art models, we selected the proposals with the most competitive results on both the MIMIC-CXR and Open-i datasets, as detailed in Table 3. The table is divided into two sections: models trained with NLL and models trained with RL, comparing them across the same metrics presented in Table 2. Metrics optimized for models trained with RL are indicated in parentheses. The benchmark model we employ is SwinB+BERT9k utilizing RL and TA, as depicted in Table 2.

The results underscore the importance of utilizing RL for training, as models trained with NLL exhibit inferior performance. In NLL models for MIMIC-CXR, it can be observed that Nicolson et al. [32]’s proposal achieves the highest BLEU4 score while securing a second position in terms of ROUGE-L. Pan et al. [29]’s proposal obtains the highest ROUGE-L value. Regarding chest RRG metrics, due to their novelty, not all proposals have registered values, as image-captioning metrics such as BLEU4 or ROUGE-L are commonly used. Our proposal achieved the highest F₁cXb values, followed by Delbrouck et al. [12]’s model; this trend was also observed for F₁RG_{ER} values. When compared to the NLL models of the Open-i dataset, the highest BLEU4 score is achieved by Miura et al. [11], while Chen et al. [10] obtain the highest ROUGE-L score. Similar to MIMIC-CXR, in chest-RRG-oriented metrics, our model obtains the highest values, followed by Delbrouck et al. [12].

Table 3 also presents the results for methods trained with RL. The metric used within the RL reward is indicated in parentheses. In the MIMIC-CXR dataset, it is apparent that the BLEU4 score reported by Miura et al. [11] deteriorates in both models when compared to the one trained with NLL. However, it still manages to enhance the F₁cXb score. Regarding Delbrouck et al. [12]’s models, there is an improvement of about one point in BLEU4 and ROUGE-L, with their model achieving the second-best scores for F₁cXb and F₁RG_{ER} at 62.2 and 34.7, respectively. Meanwhile, our SwinB+BERT9k model surpasses all RL models across all metrics, demonstrating superior quality, factual correctness, and completeness. In the case of the Open-i dataset, a similar trend is observed, with our model outperforming other proposals by approximately nine and six points in F₁cXb and F₁RG_{ER}, respectively, compared to Delbrouck et al. [12]. The good results obtained in this database demonstrate the generalization and adaptability capacity of our training workflow to the chest RRG task.

Upon analyzing the results in Table 3, it becomes evident that the conventional approach to compare RRG systems relies on image-captioning metrics such as BLEU4 and ROUGE-L. As a result, all proposals rely on NLG-oriented metrics in their original works; however, the majority lack values for chest-RRG-oriented metrics due to their novelty.

Nevertheless, NLG-oriented metrics may lack precision in clinical settings as they do not gauge syntax, semantics, or the comprehension of report meanings. Consequently, metrics tailored to the RRG task, like F_1cXb and F_1RG_{ER} , have emerged to evaluate pathologies' presence, absence, or uncertainty and their relationships within reports. These chest-RRG-oriented metrics offer a more accurate report quality assessment than NLG-oriented metrics. For instance, despite minimal differences in BLEU4 scores between Delbrouck et al. [12] and our approach in MIMIC-CXR, a significant gap is observed when measured with chest-RRG-oriented metrics. Another example of the imprecision of NLG-oriented metrics in assessing report quality can be seen in nearly identical BLEU4 scores between Miura et al. [11]'s models trained with NLL and RL in MIMIC-CXR, yet exhibiting a 12-point difference in F_1cXb . High values in NLG-oriented metrics do not reflect the quality of a report, as neither BLEU4 nor ROUGE-L measures whether the mentioned pathologies are present, absent, or uncertain.

Table 3. Comparison of our model, SwinB+BERT9k, against top state-of-the-art models for chest RRG on the MIMIC-CXR test set and Open-i dataset. The metric values are provided in (%). Bold font indicates the best result obtained for each metric.

State of the Art	Chest-RRG-Oriented		NLG-Oriented	
	F_1cXb	F_1RG_{ER}	BLEU4	ROUGE-L
MIMIC-CXR: NLL models				
Yang et al. [28]	-	-	11.5	28.4
Pan et al. [29]	-	-	11.2	28.8
Yang et al. [30]	-	-	11.1	27.4
Zhao et al. [31]	-	-	10.9	27.5
Nicolson et al. [32]	-	-	12.7	28.6
Liu et al. [5]	29.2	-	7.6	-
Chen et al. [9]	34.6	-	8.6	27.7
Miura et al. [11]	44.7	-	11.5	-
Chen et al. [10]	40.5	-	10.6	27.8
Delbrouck et al. [12]	44.8	20.2	10.5	25.3
Ours: NLL stage	57.8	27.1	10.3	22.8
MIMIC-CXR: RL models				
Miura et al. [11] (BERTScr+fact _{ENT})	56.7	-	11.1	-
Miura et al. [11] (BERTScr+fact _{ENTNLI})	56.7	-	11.4	-
Delbrouck et al. [12] (BERTScr+F ₁ RG _{ER})	62.2	34.7	11.4	26.5
Ours: RL stage (BERTScr+F₁RG_{ER})	66.2	37.8	13.2	27.1
Open-i: NLL models				
Miura et al. [11]	32.2	-	12.1	28.8
Chen et al. [10]	-	-	12.0	29.8
Alfarghaly et al. [26]	-	-	11.1	28.9
Donahue et al. [49]	-	-	9.9	27.8
Delbrouck et al. [12]	33.1	26.4	11.4	-
Ours: NLL stage	46.6	34.5	11.2	23.4
Open-i: RL models				
Miura et al. [11] (BERTScr+fact _{ENT})	48.3	-	12.0	-
Miura et al. [11] (BERTScr+fact _{ENTNLI})	47.8	-	13.1	-
Delbrouck et al. [12] (BERTScr+F ₁ RG _{ER})	49.1	41.2	13.9	32.7
Ours: RL stage (BERTScr+F₁RG_{ER})	54.7	45.6	15.1	34.3

4. Discussion

This study presents an efficient workflow for training VED models explicitly designed for chest RRG. This workflow demonstrates remarkable performance in chest-RRG-oriented metrics. One pivotal finding lies in our analysis of the proposed architectures, where we show that utilizing the SwinB encoder yields superior radiograph representations. Notably, the word-based approach proves more competitive in the decoder component

than its subword-based counterpart. This finding contrasts with strategies commonly employed in current large language models (LLMs). However, this discrepancy can be attributed to the inherent variability and scarcity of data in the chest RRG task. Furthermore, our ablation study underscores the critical role of TA in effectively leveraging data for this task, demonstrating substantial improvements in metrics over strategies like image augmentation or HNM.

A prevalent issue in RRG proposals is the high repetition and monotony frequently encountered in reports generated by other state-of-the-art approaches. Typically, models tend to converge to a local minimum, yielding nearly identical standardized reports across most patients. This phenomenon arises as deep learning algorithms settle at a midpoint where the generated reports broadly apply to most cases. These reports adhere to a template naming common pathologies in a fixed order and syntactic structure, irrespective of the patient. Consequently, this pattern results in reports of dubious quality, often overlooking less common pathologies, thereby undermining the reliability of diagnostic models.

Given the high repetition and monotony in the generated reports of state-of-the-art models, we propose an analysis of report diversity using N-grams in the MIMIC-CXR test set. This analysis allows for measuring diversity by examining the number of unique N-grams present in generated reports, as illustrated in Figure 2. Specifically, we compare reports generated by our model with those produced by the approach presented in Delbrouck et al. [12], acknowledged for being the most competitive proposal in the field, as outlined in Table 3. The graph illustrates a notably greater diversity in our generated reports compared to those of the Delbrouck et al. [12] model. Moreover, our model achieves a diversity difference averaging nearly three times their model.

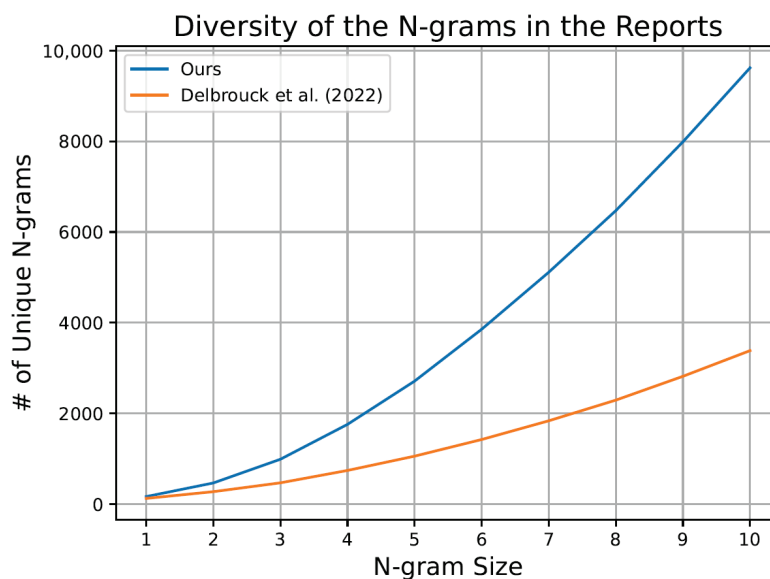
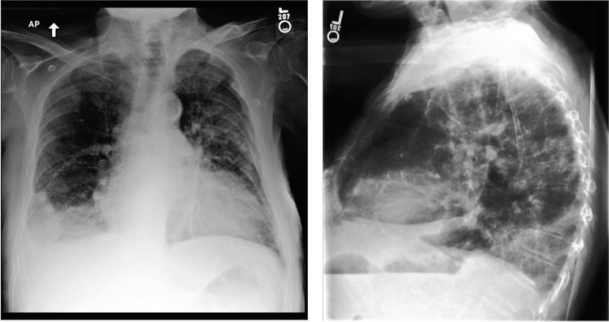
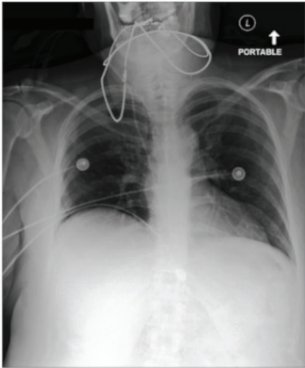
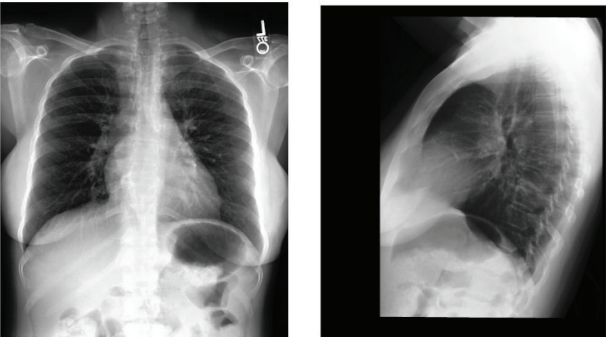


Figure 2. The graph showcases the diversity of N-grams within the generated reports using the MIMIC-CXR test set, encompassing the proposal by Delbrouck et al. [12] and our approach.

Consequently, the TA strategy substantially augments the volume of reports for training and significantly boosts the performance of VED models, as demonstrated by chest-RRG-oriented metrics. An additional qualitative advantage of our TA-based approach lies in its capacity to generate less repetitive and more diverse reports compared to prior methods. We illustrate this with examples of reports from SwinB+BERT9k and their corresponding RadGraph entities in Table 4. ANAT-DP denotes anatomical body parts mentioned in the report, while OBS signifies observations associated with radiology images, categorized as definitely present (OBS-DP), definitely absent (OBS-DA), or uncertain (OBS-U). Despite improving report quality with techniques like TA, specific repetitions and

monotony persist. This will need to be addressed using new metrics and strategies in the future.

Table 4. Comparison of reports generated by our SwinB+BERT9k model trained with RL and TA against reference reports on the MIMIC-CXR test set. The first and third cases involve a patient whose study contains two images, while the second involves a single image.

Images	Hypothesis Report	Reference Report
<p>Input images = 2</p> 	<p>$F_1R_{G_{ER}} 63.6 \%$</p> <p>the cardiac ANAT-DP silhouette ANAT-DP is moderate OBS-DP enlarged OBS-DP . moderate OBS-DP cardiomegaly OBS-DP is unchanged OBS-DP compared to the previous radiograph . small OBS-DP left ANAT-DP pleural ANAT-DP effusion OBS-DP is present . there is mild OBS-DP pulmonary ANAT-DP edema OBS-DP . there is increased OBS-DP opacity OBS-DP in the right ANAT-DP lower ANAT-DP lobe ANAT-DP is seen . there is mild OBS-DP atelectasis OBS-DP demonstrated at the right ANAT-DP lung ANAT-DP . there is no pneumothorax OBS-DA .</p>	<p>as compared to ____ , interval worsening OBS-DP moderate OBS-DP pulmonary ANAT-DP edema OBS-DP . right ANAT-DP moderate OBS-DP pleural ANAT-DP effusion OBS-DP has also slightly OBS-DP increased OBS-DP . small OBS-DP left ANAT-DP effusion OBS-DP persists . left ANAT-DP lower ANAT-DP lobe ANAT-DP parenchymal ANAT-DP opacity OBS-DP in the superior ANAT-DP segment ANAT-DP is now obscured OBS-DP by increasing OBS-DP pulmonary ANAT-DP edema OBS-DP . moderate OBS-DP cardiomegaly OBS-DP . no pneumothorax OBS-DA .</p>
<p>Input images = 1</p> 	<p>$F_1R_{G_{ER}} 43.2 \%$</p> <p>lung ANAT-DP volumes ANAT-DP are low OBS-DP . the lungs ANAT-DP are clear OBS-DP . the cardiomeastinal ANAT-DP silhouette ANAT-DP is normal OBS-DP . a right ANAT-DP central ANAT-DP venous ANAT-DP catheter OBS-DP is seen with tip OBS-DP terminating in the right ANAT-DP atrium ANAT-DP . there is no focal OBS-DA consolidation OBS-DA , pleural ANAT-DP effusion OBS-DA or pneumothorax OBS-DA . the heart ANAT-DP size ANAT-DP is normal OBS-DP .</p>	<p>since most recent chest radiograph , there has been interval placement of a right ANAT-DP ij ANAT-DP central ANAT-DP venous ANAT-DP catheter OBS-DP which terminates projecting over the right ANAT-DP atrium ANAT-DP . there is no pneumothorax OBS-DA . lungs ANAT-DP are clear OBS-DP . persistent elevation OBS-DP the right ANAT-DP hemidiaphragm ANAT-DP is noted . radiopaque OBS-DP lucencies OBS-DP overlie OBS-DP the right ANAT-DP upper ANAT-DP mediastinum ANAT-DP .</p>
<p>Input images = 2</p> 	<p>$F_1R_{G_{ER}} 27.6 \%$</p> <p>the lungs ANAT-DP are clear OBS-DP . the cardiomeastinal ANAT-DP silhouette ANAT-DP is normal OBS-DP . there is no focal OBS-DA consolidation OBS-DA , pleural ANAT-DP effusion OBS-DA or pneumothorax OBS-DA . the heart ANAT-DP size ANAT-DP is normal OBS-DP . mediastinal ANAT-DP and hilar ANAT-DP contours ANAT-DP are</p>	<p>heart ANAT-DP size ANAT-DP is normal OBS-DP . lung ANAT-DP fields OBS-DP are clear OBS-DP . the superior ANAT-DP mediastinum ANAT-DP appears slightly OBS-DP widened OBS-DP , but this may be projectional OBS-U . patient is mildly OBS-DP rotated OBS-DP . followup films in four to six weeks ' time are recommended to keep this area under observation . because of varying degrees of rotation OBS-DP , comparison to the previous examination of ____ is difficult .</p>

In addition to the inherent complexity of this task, focusing on medical text generation necessitates the analysis and proposal of additional chest-RRG-oriented metrics based on NLP models. Exploring novel methods to assess the reference report against an estimated one can facilitate a more comprehensive evaluation of state-of-the-art models. Our study demonstrates the utility of F_1cXb in verifying whether generated reports discuss patholo-

gies present in the reference reports. Furthermore, F_1RG_{ER} ensures that the present, absent, and unknown pathologies are correctly related in the reports. Considering these findings, these two metrics are essential for model comparison. Although these metrics effectively and efficiently enhance RL training, models frequently produce reports with incomplete sentences. This deficiency is evident in numerous state-of-the-art proposals, and despite the efficacy of TA, it still occurs occasionally. It is evidenced at the end of the third report in Table 4. This phenomenon arises from models learning that specific syntactic structures can benefit chest-RRG-oriented metrics like F_1RG_{ER} , even if they do not constitute syntactically correct sentences. Moving forward, we advocate for further research into novel metrics using alternative deep learning algorithms, such as LLMs like GPT-4 [50] or LLaMA [51].

Several aspects of the proposed approach for real-world RRG scenarios merit attention, including multilanguage support, standardization of reports, and improvements in picture archiving and communication systems (PACS). Since the decoder model primarily operates with English words, it encounters specific difficulties when used in other languages. Thus, it would be necessary to remove the last layer of the decoder model and adjust it based on the desired language; however, the rest of the model remains language-independent. Another challenge lies in the lack of standardization in report writing across different institutions, leading to variations in reporting guidelines. Consequently, RRG models can assist in integrating new reporting guidelines to standardize and unify databases for RRG. Furthermore, our approach can enhance PACS, which are commonly used in healthcare for securely storing and transmitting electronic images and clinical reports. Collaborating with expert radiologists, we can integrate RRG models to streamline processes within PACS, such as preparing automatic draft reports that radiologists can interactively correct. This approach can improve diagnostic efficiency and expert assistance times while enhancing the standardization and homogenization of reports.

The strengths and weaknesses of our model have been highlighted using different metrics. However, another essential aspect to analyze is the integration of the proposed workflow into the clinical setting. Given that the system relies on a VED deep learning model, its deployment only requires a computer with a GPU with at least 12 GB of memory. The model specializes in chest radiographs from different position views, such as antero-posterior, lateral, posteroanterior, and lateral decubitus. Therefore, it can integrate and analyze a wide variety of chest radiographs.

The scope of this work is focused on chest radiographs. However, our approach is not limited solely to its application in chest radiographs. The proposed approach is flexible and independent, allowing it to be applied to other radiological scenarios. Therefore, our approach can be seamlessly adapted for radiographs of different body parts. The only aspect requiring modification is the reward optimization during RL training. Since F_1RG_{ER} primarily specializes in chest-related pathologies, this metric should be replaced with another relevant to the radiological scenario where our approach is intended for application. The remaining components, such as the encoder, decoder, TA, and decoding strategies, are independent of the radiological scenario. Despite the flexibility and independence of the approach, the major limitation for application to other radiological scenarios depends on the availability of training data. Acquiring large datasets to train VED models is crucial for improving the quality of reports and enhancing their applicability to different scenarios.

5. Conclusions

Chest RRG poses a challenge due to the limited data availability and the diverse nature of medical reports and variations in pathology expressions. Four transformer-based models were analyzed, highlighting the encoder as the key component, with SwinB being the best choice. This transformer encoder has not been previously explored in RRG. Regarding the decoder, a word-based approach trains faster and achieves more competitive results than a subword-based one.

Techniques like image data augmentation, hard negative mining, and TA were introduced. TA was shown to be an effective method to improve generalization and the quality

of the generated reports regarding variability, quality, factual correctness, and completeness, which yields a model that outperforms the state of the art. Moreover, this approach paves the way for new TA approaches based on augmentation with more complex NLP techniques, such as LLM models.

Additionally, NLG-oriented metrics may not be optimal for measuring report quality, as they compare sentences and words without considering semantic meaning or alternative expressions of the same diagnosis. Thus, metrics based on NLP models specialized in chest radiology reports like F_1cXb and F_1RG_{ER} appear to be crucial in evaluating report quality.

Author Contributions: Investigation, D.P.; methodology, D.P.; software, D.P.; supervision, A.A. and R.P.; writing—original draft, D.P.; writing—review and editing, A.A. and R.P. All authors have read and agreed to the published version of the manuscript.

Funding: Work was partially supported by the Generalitat Valenciana under the predoctoral grant CIACIF/2022/289, with the support of valgrAI—Valencian Graduate School and Research Network of Artificial Intelligence and the Generalitat Valenciana, and cofunded by the European Union.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The data used in this study are publicly available.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Zhao, J.; Lv, Y.; Zeng, Q.; Wan, L. Online Policy Learning-Based Output-Feedback Optimal Control of Continuous-Time Systems. *IEEE Trans. Circuits Syst. II Express Briefs* **2024**, *71*, 652–656. [CrossRef]
2. Gale, W.; Oakden-Rayner, L.; Carneiro, G.; Bradley, A.P.; Palmer, L.J. Producing radiologist-quality reports for interpretable artificial intelligence. *arXiv* **2018**, arXiv:1806.00340.
3. Li, Y.; Liang, X.; Hu, Z.; Xing, E.P. Hybrid Retrieval-Generation Reinforced Agent for Medical Image Report Generation. In Proceedings of the Advances in Neural Information Processing Systems, Montreal, BC, Canada, 3–8 December 2018; pp. 1530–1540.
4. Wang, X.; Peng, Y.; Lu, L.; Lu, Z.; Summers, R.M. TieNet: Text-Image Embedding Network for Common Thorax Disease Classification and Reporting in Chest X-Rays. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 9049–9058.
5. Liu, G.; Hsu, T.M.H.; McDermott, M.; Boag, W.; Weng, W.H.; Szolovits, P.; Ghassemi, M. Clinically Accurate Chest X-ray Report Generation. In Proceedings of the 4th Machine Learning for Healthcare Conference, Ann Arbor, MI, USA, 9–10 August 2019; pp. 249–269.
6. Liu, F.; Wu, X.; Ge, S.; Fan, W.; Zou, Y. Exploring and Distilling Posterior and Prior Knowledge for Radiology Report Generation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Nashville, TN, USA, 20–25 June 2021; pp. 13753–13762.
7. Liu, F.; You, C.; Wu, X.; Ge, S.; Wang, S.; Sun, X. Auto-Encoding Knowledge Graph for Unsupervised Medical Report Generation. In Proceedings of the Advances in Neural Information Processing Systems, Online, 6–14 December 2021; pp. 16266–16279.
8. Windsor, R.; Jamaludin, A.; Kadir, T.; Zisserman, A. Vision-Language Modelling For Radiological Imaging and Reports In The Low Data Regime. In Proceedings of the Medical Imaging with Deep Learning, Nashville, TN, USA, 10 July 2023; pp. 53–73.
9. Chen, Z.; Song, Y.; Chang, T.H.; Wan, X. Generating Radiology Reports via Memory-driven Transformer. *arXiv* **2020**, arXiv:2010.16056.
10. Chen, Z.; Shen, Y.; Song, Y.; Wan, X. Cross-modal Memory Networks for Radiology Report Generation. *arXiv* **2022**, arXiv:2204.13258.
11. Miura, Y.; Zhang, Y.; Tsai, E.B.; Langlotz, C.P.; Jurafsky, D. Improving Factual Completeness and Consistency of Image-to-Text Radiology Report Generation. *arXiv* **2021**, arXiv:2010.10042.
12. Delbrouck, J.B.; Chambon, P.; Bluethgen, C.; Tsai, E.; Almusa, O.; Langlotz, C.P. Improving the Factual Correctness of Radiology Report Generation with Semantic Rewards. *arXiv* **2022**, arXiv:2210.12186.
13. Zhang, T.; Kishore, V.; Wu, F.; Weinberger, K.Q.; Artzi, Y. BERTScore: Evaluating Text Generation with BERT. *arXiv* **2020**, arXiv:1904.09675.
14. Jain, S.; Agrawal, A.; Saporta, A.; Truong, S.Q.; Duong, D.N.; Bui, T.; Chambon, P.; Zhang, Y.; Lungren, M.P.; Ng, A.Y.; et al. RadGraph: Extracting Clinical Entities and Relations from Radiology Reports. *arXiv* **2021**, arXiv:2106.14463.
15. Schlegl, T.; Waldstein, S.M.; Vogl, W.D.; Schmidt-Erfurth, U.; Langs, G. Predicting Semantic Descriptions from Medical Images with Convolutional Neural Networks. In Proceedings of the Information Processing in Medical Imaging, Isle of Skye, UK, 28 June–3 July 2015; pp. 437–448.

16. Shin, H.C.; Roberts, K.; Lu, L.; Demner-Fushman, D.; Yao, J.; Summers, R.M. Learning to Read Chest X-Rays: Recurrent Neural Cascade Model for Automated Image Annotation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 2497–2506.
17. Moradi, M.; Madani, A.; Gur, Y.; Guo, Y.; Syeda-Mahmood, T. Bimodal Network Architectures for Automatic Generation of Image Annotation from Text. In Proceedings of the Medical Image Computing and Computer Assisted Intervention, Granada, Spain, 16 September 2018; pp. 449–456.
18. Rubin, J.; Sanghavi, D.; Zhao, C.; Lee, K.; Qadir, A.; Xu-Wilson, M. Large Scale Automated Reading of Frontal and Lateral Chest X-Rays using Dual Convolutional Neural Networks. *arXiv* **2018**, arXiv:1804.07839.
19. Jing, B.; Xie, P.; Xing, E. On the Automatic Generation of Medical Imaging Reports. In Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics, Melbourne, Australia, 15–20 July 2018; pp. 2577–2586.
20. Monshi, M.M.A.; Poon, J.; Chung, V. Deep learning in generating radiology reports: A survey. *Artif. Intell. Med.* **2020**, *106*, 101878. [CrossRef]
21. Krizhevsky, A.; Sutskever, I.; Hinton, G.E. ImageNet Classification with Deep Convolutional Neural Networks. In Proceedings of the Advances in Neural Information Processing Systems, Lake Tahoe, NV, USA, 3–6 December 2012; Volume 25, pp. 1097–1105.
22. Simonyan, K.; Zisserman, A. Very Deep Convolutional Networks for Large-Scale Image Recognition. *arXiv* **2014**, arXiv:1409.1556.
23. Szegedy, C.; Liu, W.; Jia, Y.; Sermanet, P.; Reed, S.; Anguelov, D.; Erhan, D.; Vanhoucke, V.; Rabinovich, A. Going Deeper With Convolutions. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Boston, MA, USA, 7–12 June 2015; pp. 1–9.
24. 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.
25. Liao, Y.; Liu, H.; Spasić, I. Deep learning approaches to automatic radiology report generation: A systematic review. *Inform. Med. Unlocked* **2023**, *39*, 101273. [CrossRef]
26. Alfarghaly, O.; Khaled, R.; Elkorany, A.; Helal, M.; Fahmy, A. Automated radiology report generation using conditioned transformers. *Inform. Med. Unlocked* **2021**, *24*, 100557. [CrossRef]
27. Sanh, V.; Debut, L.; Chaumond, J.; Wolf, T. DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter. *arXiv* **2019**, arXiv:1910.01108.
28. Yang, S.; Wu, X.; Ge, S.; Zhou, S.K.; Xiao, L. Knowledge matters: Chest radiology report generation with general and specific knowledge. *Med. Image Anal.* **2022**, *80*, 102510. [CrossRef] [PubMed]
29. Pan, Y.; Liu, L.J.; Yang, X.B.; Peng, W.; Huang, Q.S. Chest radiology report generation based on cross-modal multi-scale feature fusion. *J. Radiat. Res. Appl. Sci.* **2024**, *17*, 100823. [CrossRef]
30. Yang, S.; Wu, X.; Ge, S.; Zheng, Z.; Zhou, S.K.; Xiao, L. Radiology report generation with a learned knowledge base and multi-modal alignment. *Med. Image Anal.* **2023**, *86*, 102798. [CrossRef] [PubMed]
31. Zhao, G.; Zhao, Z.; Gong, W.; Li, F. Radiology report generation with medical knowledge and multilevel image-report alignment: A new method and its verification. *Artif. Intell. Med.* **2023**, *146*, 102714. [CrossRef] [PubMed]
32. Nicolson, A.; Dowling, J.; Koopman, B. Improving chest X-ray report generation by leveraging warm starting. *Artif. Intell. Med.* **2023**, *144*, 102633. [CrossRef]
33. Radford, A.; Wu, J.; Child, R.; Luan, D.; Amodei, D.; Sutskever, I. Language Models are Unsupervised Multitask Learners. *OpenAI Blog* **2019**, *1*, 9.
34. Devlin, J.; Chang, M.W.; Lee, K.; Toutanova, K. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *arXiv* **2019**, arXiv:1810.04805.
35. Irvin, J.; Rajpurkar, P.; Ko, M.; Yu, Y.; Ciurea-Illcus, S.; Chute, C.; Marklund, H.; Haghighi, B.; Ball, R.; Shpanskaya, K.; et al. CheXpert: A Large Chest Radiograph Dataset with Uncertainty Labels and Expert Comparison. In Proceedings of the AAAI Conference on Artificial Intelligence, Honolulu, HI, USA, 27 January–1 February 2019; Volume 33, pp. 590–597.
36. Huang, G.; Liu, Z.; van der Maaten, L.; Weinberger, K.Q. Densely Connected Convolutional Networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 4700–4708.
37. Liu, Z.; Lin, Y.; Cao, Y.; Hu, H.; Wei, Y.; Zhang, Z.; Lin, S.; Guo, B. Swin Transformer: Hierarchical Vision Transformer Using Shifted Windows. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Montreal, BC, Canada, 11–17 October 2021; pp. 10012–10022.
38. Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; et al. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. *arXiv* **2020**, arXiv:2010.11929.
39. Deng, J.; Dong, W.; Socher, R.; Li, L.J.; Li, K.; Fei-Fei, L. ImageNet: A large-scale hierarchical image database. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Miami, FL, USA, 20–25 June 2009; pp. 248–255.
40. Liu, Y.; Ott, M.; Goyal, N.; Du, J.; Joshi, M.; Chen, D.; Levy, O.; Lewis, M.; Zettlemoyer, L.; Stoyanov, V. RoBERTa: A Robustly Optimized BERT Pretraining Approach. *arXiv* **2019**, arXiv:1907.11692.
41. Lewis, M.; Liu, Y.; Goyal, N.; Ghazvininejad, M.; Mohamed, A.; Levy, O.; Stoyanov, V.; Zettlemoyer, L. BART: Denoising Sequence-to-Sequence Pre-training for Natural Language Generation, Translation, and Comprehension. *arXiv* **2019**, arXiv:1910.13461.
42. Rennie, S.J.; Marcheret, E.; Mroueh, Y.; Ross, J.; Goel, V. Self-Critical Sequence Training for Image Captioning. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 7008–7024.

43. Johnson, A.E.W.; Pollard, T.J.; Greenbaum, N.R.; Lungren, M.P.; Deng, C.Y.; Peng, Y.; Lu, Z.; Mark, R.G.; Berkowitz, S.J.; Horng, S. MIMIC-CXR-JPG, a large publicly available database of labeled chest radiographs. *arXiv* **2019**, arXiv:1901.07042.
44. Demner-Fushman, D.; Antani, S.; Simpson, M.; Thoma, G.R. Design and development of a multimodal biomedical information retrieval system. *J. Comput. Sci. Eng.* **2012**, *6*, 168–177. [CrossRef]
45. 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 on Association for Computational Linguistics, Philadelphia, PA, USA, 7–12 July 2002; pp. 311–318.
46. Lin, C.Y. ROUGE: A Package for Automatic Evaluation of Summaries. In Proceedings of the Text Summarization Branches Out, Barcelona, Spain, 4–10 July 2004; pp. 74–81.
47. Zhang, Y.; Merck, D.; Tsai, E.B.; Manning, C.D.; Langlotz, C.P. Optimizing the Factual Correctness of a Summary: A Study of Summarizing Radiology Reports. *arXiv* **2020**, arXiv:1911.02541.
48. Smit, A.; Jain, S.; Rajpurkar, P.; Pareek, A.; Ng, A.Y.; Lungren, M.P. CheXbert: Combining Automatic Labelers and Expert Annotations for Accurate Radiology Report Labeling Using BERT. *arXiv* **2020**, arXiv:2004.09167.
49. Donahue, J.; Hendricks, L.A.; Rohrbach, M.; Venugopalan, S.; Guadarrama, S.; Saenko, K.; Darrell, T. Long-Term Recurrent Convolutional Networks for Visual Recognition and Description. *IEEE Trans. Pattern Anal. Mach. Intell.* **2017**, *39*, 677–691. [CrossRef]
50. OpenAI. GPT-4 Technical Report. *arXiv* **2023**, arXiv:2303.08774.
51. Touvron, H.; Lavril, T.; Izacard, G.; Martinet, X.; Lachaux, M.A.; Lacroix, T.; Rozière, B.; Goyal, N.; Hambro, E.; Azhar, F.; et al. LLaMA: Open and Efficient Foundation Language Models. *arXiv* **2023**, arXiv:2302.13971.

Disclaimer/Publisher’s Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.

Article

Optimizing Skin Cancer Survival Prediction with Ensemble Techniques

Erum Yousef Abbasi ^{1,*}, Zhongliang Deng ¹, Arif Hussain Magsi ², Qasim Ali ³, Kamlesh Kumar ⁴ and Asma Zubedi ⁵

¹ State Key Laboratory of Wireless Network Positioning and Communication Engineering Integration Research, School of Electronics Engineering, Beijing University of Posts and Telecommunications, Beijing 100876, China; dengzhl@bupt.edu.cn

² State Key Laboratory of Networking and Switching Technology, Beijing University of Posts and Telecommunications, Beijing 100876, China; ahmagsi@bupt.edu.cn

³ Department of Software Engineering, Mehran University of Engineering and Technology, Jamshoro 76062, Pakistan; qasim.arain@faculty.muett.edu.pk

⁴ School of Electronics Engineering, Beijing University of Posts and Telecommunications, Beijing 100876, China; kamleshsoothar@gmail.com

⁵ School of Economics and Management, Beijing University of Posts and Telecommunications, Beijing 100876, China; asmazubedi@bupt.edu.cn

* Correspondence: erumzubedi@bupt.edu.cn

Abstract: The advancement in cancer research using high throughput technology and artificial intelligence (AI) is gaining momentum to improve disease diagnosis and targeted therapy. However, the complex and imbalanced data with high dimensionality pose significant challenges for computational approaches and multi-omics data analysis. This study focuses on predicting skin cancer and analyzing overall survival probability. We employ the Kaplan–Meier estimator and Cox proportional hazards regression model, utilizing high-throughput machine learning (ML)-based ensemble methods. Our proposed ML-based ensemble techniques are applied to a publicly available dataset from the ICGC Data Portal, specifically targeting skin cutaneous melanoma cancers (SKCM). We used eight baseline classifiers, namely, random forest (RF), decision tree (DT), gradient boosting (GB), AdaBoost, Gaussian naïve Bayes (GNB), extra tree (ET), logistic regression (LR), and light gradient boosting machine (Light GBM or LGBM). The study evaluated the performance of the proposed ensemble methods and survival analysis on SKCM. The proposed methods demonstrated promising results, outperforming other algorithms and models in terms of accuracy compared to traditional methods. Specifically, the RF classifier exhibited outstanding precision results. Additionally, four different ensemble methods (stacking, bagging, boosting, and voting) were created and trained to achieve optimal results. The performance was evaluated and interpreted using accuracy, precision, recall, F1 score, confusion matrix, and ROC curves, where the voting method achieved a promising accuracy of 99%. On the other hand, the RF classifier achieved an outstanding accuracy of 99%, which exhibits the best performance. We compared our proposed study with the existing state-of-the-art techniques and found significant improvements in several key aspects. Our approach not only demonstrated superior performance in terms of accuracy but also showcased remarkable efficiency. Thus, this research work contributes to diagnosing SKCM with high accuracy.

Keywords: skin cancer; melanoma; machine learning; ensemble technique; feature selection

1. Introduction

In recent years, the alarming surge in malignant diseases has become a critical global health concern. Among these malignancies, skin cutaneous melanoma cancer (SKCM) is one of the most aggressive variants, demanding thorough investigation and understanding [1]. According to an International Agency for Research on Cancer (IARC) report, cancer is

the leading cause of mortality. The report exhibits that nearly 10 million deaths have resulted from various types of cancer [2]. The World Health Organization (WHO) 2023 report illustrates that cancer is the second leading cause of death (16%), followed by cardiovascular disease (27%) [3]. In such a situation, the early diagnosis of a disease can cure and prevent the patients from further jeopardy. In general, there are two main forms of skin cancer: melanoma (cancers resulting from melanocyte malfunction) and non-melanoma skin cancers (from cells generated from the epidermis) [4]. Among various types of cancers, SKCM has become one of the most prevalent cancers in the last ten years [5] with tumors made of melanocyte cells. It is currently a major public health issue worldwide, and the increasing prevalence of the disease might significantly impact the world's population and economy [6]. However, early diagnosis and effective tumor therapy lead to a cure rate of over 90% in individuals with incipient melanoma [7]. There are several factors for an increased number of skin cancers. One of the most common occurrences of skin cancer is due to ultraviolet (UV) rays [8]. Other reasons include sun exposure, depletion of the ozone layer, genetic predisposition, and so on.

Several studies have shown that SKCM results from abnormalities in transcriptional and epigenetic factors, including the expression of messenger ribonucleic acid (mRNAs) and micro ribonucleic acid (miRNAs), the aberration in methylation patterns of CpG islands of genes, and histone modifications, which opens the door for the development of potential molecular biomarkers in melanoma [9,10]. As predictive indicators for cutaneous melanoma, miRNA expression has been implicated in several past studies.

Various healthcare sectors, including dermatology, have leveraged artificial intelligence (AI), revolutionizing diagnostic and therapeutic processes. Diverse biomedical data from health records, medical images, IoT sensor data, and text can be used to predict SKCM. Specifically, machine learning (ML) and deep learning (DL) significantly contribute to predicting the disease on publicly available datasets. The most recent skin cancer detection technique includes dermoscopy with AI, which leverages the handheld device for magnifying the skin and allows dermatologists to examine moles and lesions in detail. The ML and DL algorithms require structured data for classification, with lower prediction accuracy, and require more computational time. Due to its superiority over traditional analytical methods, AI has significantly uplifted the healthcare industry. Applications of AI in healthcare are being used with increasing optimism, and they range from speeding up the research of new drugs to helping with medical diagnosis, treatment, and administrative support. Additionally, using it as an adjuvant in clinical decision making can be advantageous [11,12]. Various ML algorithms are leveraged to predict different diseases in the early stage after diagnosing different attributes of the disease. Those diseases include different cancer types, diabetes, kidney disease, and other diseases [13]. Specifically, ensemble learning is a learning method that combines multiple baseline models to create a powerful single model. It reduces overfitting risk and has been successfully applied in various fields. Common ensemble techniques include averaging, bagging, boosting, stacking, and voting [14]. Traditional ensemble learning integrates ML models across various fields, but efforts have shifted to DL, focusing on complex models and integrating them across various fields [15]. The other recent skin cancer detection technique involves reflectance confocal microscopy (RCM), a non-invasive imaging approach, that enables high-resolution skin imaging at the cellular level. It aids dermatologists in visualizing skin structures and identifying abnormal cells without a biopsy procedure. Another recent technique leverages smartphone applications where mobile applications utilize smartphone cameras for skin self-examination. These apps often use AI algorithms to analyze photos and provide risk assessments. The main objective of this study is to propose four ensemble methods for predicting skin cancer by utilizing ML algorithms. We experimented with five transcriptomic technologies from the ICGC portal [16]. In this research, three features were leveraged—recursive feature elimination (RFE), forward feature selection (FFE), and backward feature elimination (BFE)—for the ensemble method. This paper discusses the five phases for ensemble methods based on ML algorithms that use transcriptomic technology

data to predict SKCM. Figure 1 illustrates the workflow of our integrative study. Figure 1 (1) depicts the data collection source and 5 different transcriptomic technology datasets. Figure 1 (2) illustrates preprocessing and analysis steps. We handled missing data with the MICE imputation technique and applied three methods for best feature selections. Figure 1 (3) represents the experimental achievements of our study using different ML algorithms as baseline classifiers to create an ensemble method. Overall survival was analyzed with the Kaplan–Meier and Cox hazard regression model. Different from traditional methods in existing literature, this study contributes to predict skin cancer using various ML techniques. The novelty of this work lies in its comprehensive approach, combining high-throughput ML-based ensemble methods with the analysis of multi-omics data, particularly addressing the challenges posed by complex and imbalanced datasets with high dimensionality. Figure 1 (4) shows the biological interpretation and comparative study.

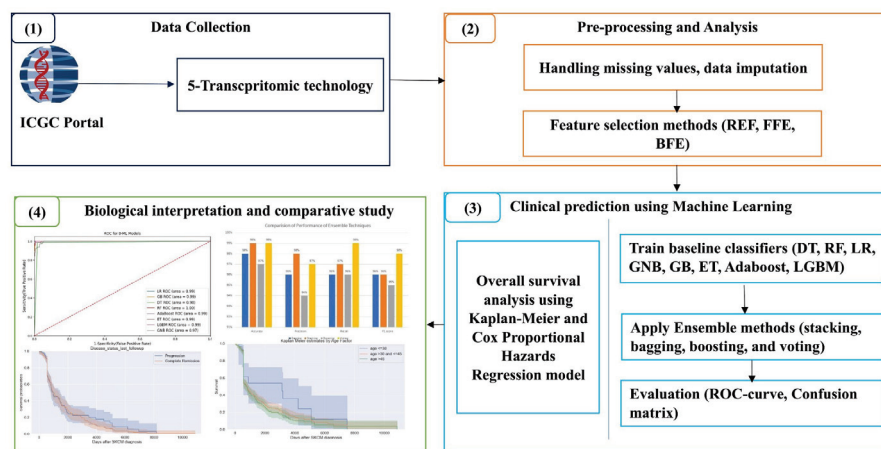


Figure 1. Workflow of the proposed research.

The key contributions of this study are as follows:

- We evaluated various techniques for SKCM prediction considering their suitability and effectiveness in this context.
- We used RFE, FFE, and BFE features for ensemble methods.
- We analyzed the overall survival (OS) analysis and progression through the Kaplan–Meier estimator and the Cox hazard proportional regression model.
- We used eight baseline classifiers, namely, random forest (RF), decision tree (DT), gradient boosting (GB), AdaBoost, Gaussian naïve Bayes (GNB), extra tree (ET), logistic regression (LR), and light GBM in this research work.
- We applied ML algorithms for predicting the disease with various selected features.
- We trained four ensemble learning methods, including stacking, bagging, boosting, and voting, to achieve the best results.

The rest of the article is organized as follows: Section 2 provides a detailed description of related work. In Section 3, we provide ML-based materials and methods that include data collection, preprocessing techniques, and classification methods. Section 4 exhibits the achieved results and discussion. Finally, Section 5 concludes the article.

2. Related Work

In recent years, cancer has been a very undeniable global health challenge. There are various cancer types, such as lymphoma, leukemia, breast cancer [17], lung cancer, skin cancer, and so on. Early skin cancer detection significantly impacts prognosis, and various techniques have been exploited. From histopathological examination to advanced imaging modalities [18], the quest for optimizing predictive models has given rise to the integration of ensemble techniques. This literature review delves into the multifaceted landscape of

skin cancer detection methodologies, focusing on the evolving role of ensemble techniques in enhancing survival prediction accuracy. On the other hand, ML is greatly contributing to anomaly detection in various fields, including health care, vehicular networks [19,20], the Internet of Things (IoT), E-commerce, and so on. ML and DL algorithms significantly aid in identifying skin cancer, with early detection potentially leading to successful treatment, making melanoma a significant health concern [21–24]. Various ML and DL techniques have been applied in existing literature, such as in [25], where the authors presented a convolutional neural network (CNN) based DL stacked ensemble framework for melanoma skin cancer detection using transfer learning. The model uses multiple CNN sub-models and a meta-learner to predict malignant melanoma moles. The model achieves a high accuracy of 95.76%, precision of 95.60%, recall of 96.67%, specificity of 94.67%, F1 score of 94.67%, and area under the curve (AUC) of 0.957% identifying both benign and malignant melanoma. Although this research is important, it could not achieve better accuracy. Similarly, another work in [26] proposed a DL-based skin cancer detection system on an imbalanced dataset. The authors employed the MNIST: HAM10000 dataset that contains seven classes of skin lesions. In order to classify the skin cancer, the authors utilized AlexNet, InceptionV3, and RegNetY-320 techniques. However, the achieved accuracy (91%), F1-score (88.1%), and ROC curve (95%) reflect a poor accuracy as compared to our proposed study.

Moreover, the authors in [27] proposed a CNN-based skin cancer detection system using a publicly available dataset, HAM10000, that includes seven skin cancer types. The authors achieved the following: accuracy (86%), precision (84%), recall (86%), and F-1 score (86%). Thus, all the achieved results fall in the 80s, which reflects the poor performance of the proposed study. Authors in [28] employed a CNN-based approach using a HAM10000 dataset that comprises 6705 benign and 1113 malignant samples and 2197 unknown lesion samples. The proposed model achieved an accuracy of 93.16% on training and 91.93% on testing. Moreover, the authors balanced the dataset of both classes, resulting in an enhanced accuracy of categorization. Despite training several transfer learning models on the same dataset, the outcomes did not surpass those of their proposed model. Another similar work in [29] proposed a CNN-based skin cancer diagnosis that is evaluated using the ISIC 2019 dataset. This work is based on multiclassification system that classifies the cancer types including benign keratosis, melanoma, melanocytic nevi, and basal cell carcinoma. The achieved results depicted an accuracy of 96.91%, which is inefficient as compared to our proposed study. Similar to our work, authors in [9] studied three immune-related mRNAs (SUCO, BTN3A1, and TBC1D2) linked to melanoma prognosis. This study used univariate Cox regression and Kaplan–Meier analysis to compare the overall survival probability between high-risk and low-risk groups, analyzing the time-dependent ROC curve. However, the accuracy of various classifiers is lower as compared to our achieved results.

Furthermore, reference [10] developed a combination of ML and DL-based tools to predict the short-term survival of cutaneous malignant melanoma (CMM), a common malignancy. The study found that additional clinical variables such as sex, tumor site, histotype, growth phase, and age were significantly linked to overall survival, with DNN and RF models showing the best prognostic performance with an accuracy of 91% and 88%, respectively. Reference [30] analyzed mRNA expressions of m5C regulators in colorectal cancer tissues and identified high mutation frequency. NOP2 and YBX1 were highly expressed in prostate, gallbladder, lung, and renal cancers. NSUN6 functions as a tumor suppressor in pancreatic cancer. UV radiation was identified as the primary environmental driver. The authors in [31] trained a HAM10000 ISIC dataset using DL for multiclass skin cancer diagnosis. The proposed model detects the skin lesion with an accuracy of 96.26%.

Limitations of Existing Studies

Skin cancer has become an interesting topic in current research. Most past studies preferred survival analysis using KM and Cox proportional hazards regression model. Unlike those traditional models, our study proposes ensemble methods for predicting

SKCM and analyzing the survival probability using the KM and Coz hazard regression model. Table 1 demonstrates the limitations of previous studies in comparison with our proposed research.

Table 1. Summarized related work and its limitations.

Ref.	Method	Study Area	Dataset	Results	Limitations
[25]	A CNN-based melanoma skin cancer detection	Skin cancer	Open access dataset	95%	No multi-omics exploited
[26]	DL-based skin cancer detection system	Skin cancer	HAM 10000	91%	The accuracy of proposed study is poor
[27]	DL-based melanoma detection	Skin cancer	HAM 10000	86%	Various models and datasets call for different hyperparameter settings
[28]	CNN-based skin cancer detection	Skin cancer	HAM 10000	91.93%	The limited size of the datasets employed in this study may have led to local optimizations
[29]	A DL-based framework for the multi-classification of skin cancer using dermoscopy images	Skin cancer	ISIC 2019	92%	Lower accuracy
[9]	Immune cell infiltration pattern of CM	SKCM	-	-	The study does not utilize any ML/DL algorithms to show a better performance
[10]	DL-based short-term survival of cutaneous malignant melanoma (CMM)	SKCM	RNA-seq	91%	Lower accuracy
[30]	The mRNA expressions of m5C regulators in colorectal cancer tissues	Various cancer types	RNA-seq	-	Performance metrics were not evaluated
[31]	DL-based multiclass skin cancer diagnosis	Skin cancer	HAM 10000 ISIC	96.26%	The accuracy of this study is lower

Table 2 depicts the notations and their descriptions used in this paper.

Table 2. Notations and descriptions.

Notation	Description
A	Input set of features for backward selection
B_0	Initialize the function with a full set of features
w	Subset of output features
r	A finite set of input features
D_w	Output set of features
T^+	Selection criteria function
g_i	Subset of output features
L	The desired set of features
m	A finite set of output features
X	Input set features for forward selection
X_0	Initialize the function with an empty set
p^+	Selection criteria function

Table 2. *Cont.*

Notation	Description
h	The desired set of features
Z_t	Output set of features
z_b	Subset of output features
t	Size of a subset of output features
Pro	The likelihood of the event occurred
ti	Time at which event occurred or did not occur
Y	Random duration of survival function
za	Number of patients
ca	Number of incidents
tia	Life risk at a time
d	Survival time
$E(d e)$	Hazard function
e_n	Set of factors
$h_0(d)$	Baseline hazard function
y_n	Measure of the impact of covariates on a subject's hazard

3. Materials and Methods

This study focuses on overall survival analysis and ensemble methods to predict SKCM and is described as an integrative omics study in this paper. The suggested integrative model generates trained ML classifiers that can be utilized as SKCM prediction and feature selection strategies. Our proposed research methodology involves various steps, as mentioned below:

3.1. Dataset Collection

We initially collected a dataset from a publicly available source [16] to propose ensemble methods. As illustrated in Table 3, There were five categories of multi-omics data in the datasets: donor, simple somatic mutation, miRNA seq, copy number somatic mutation, and specimen. Our cohort of 471 patients includes information on the patient's history, such as age, gender, length of survival, and donor relapse type. There are 377,735 samples in the copy number somatic mutation file and 369,409 samples in the miRNA seq file, all of which were examined and approved by the Illumina HiSeq verification platform. There are 1,048,575 samples in the simple somatic mutation file, which were examined and verified by Illumina GA sequencing and the Illumina HiSeq platform. There are 947 samples in the specimen file. Figure 2 shows the benign and malignant samples.

Table 3 illustrates the detailed description of the dataset.

Table 3. Dataset description.

Class	Records per Class	Features
Donor	471	9
Simple_somatic_mutation	1,048,576	12
Copy-number_soamtic_mutation	377,735	3
Mirna_seq	369,409	5
Specimen	947	2
Total Records	1,797,138	31

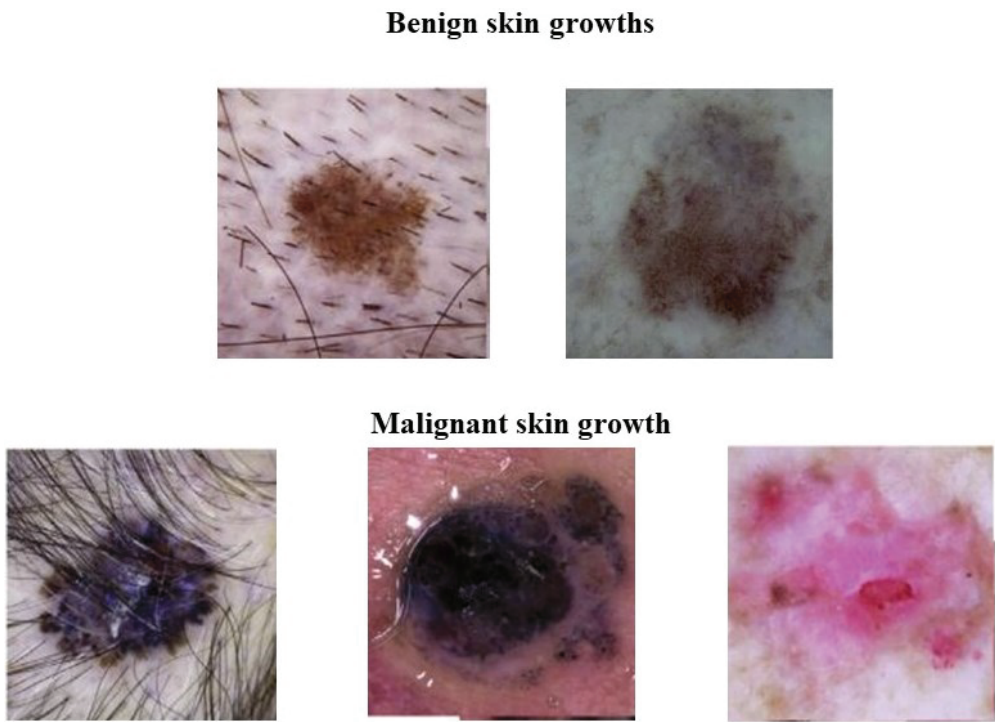


Figure 2. Various skin samples.

Figure 3 depicts the detailed dataset description in graphical form as below.

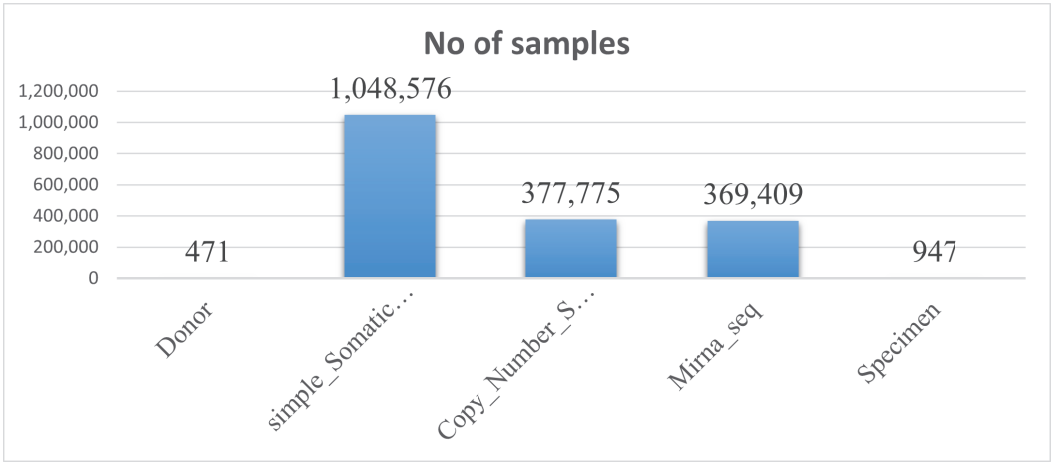


Figure 3. Dataset description.

3.2. Preprocessing

The data preprocessing plays a significant role in achieving better accuracy results in ML. Considering the importance of preprocessing, we applied various preprocessing techniques, including removing noisy data, dividing the dataset into training and testing, and feature selection. The detail of each technique is elaborated below. Initially, we removed unreliable noisy data. The features with missing value scores of more than 70% and less than 10% were excluded. The features with more than 10% and less than 70% of the data missing score were included [32]. Missing data were imputed using the multiple imputation chained equation (MICE) technique that applies the k-neighbor algorithms criteria [33]. A widely recognized Python programming language at an advanced level was employed in this research paper. These preprocessing techniques collectively contribute to the enhancement of model accuracy and robustness. Removing noisy data and selecting relevant features ensure that the subsequent machine learning models are trained on a cleaner and more

informative dataset, ultimately leading to improved predictive performance. Furthermore, the impact of these preprocessing techniques on the results is noteworthy. By systematically cleaning the data and selecting features judiciously, we mitigate the risk of model overfitting and improve generalization to new, unseen data.

Feature Selection

Precision Health uses statistical modeling based on clinical and biological data to predict patient outcomes more accurately. Traditional approaches struggle with large datasets, leading to feature selection research in various fields [34]. The following approaches improve model performance, deliver features quickly and cost-effectively, facilitate data visualization, and offer a better understanding of the data-generating process. For solving and reducing the difficulty of learning tasks, feature selection aims at removing irrelevant or redundant features. For selecting the best features, we have applied three different feature selection methods, which are discussed below:

a. Forward Feature Elimination Method (FFE): The FFE method is the reverse of the backward elimination method, starting with empty features and adding them one by one until any excluded features can significantly contribute to the model's outcome. The most significant feature is added first, and the model is refitted with the new feature. Test statistics or p values are recomputed for all remaining features. The features with the largest test statistic are chosen from the remaining features and added to the model [35]. Suppose X is an input set of features with n size of features that can be defined using Equation (1). Initially, we have an empty set of features $X_0 = \emptyset$ with the t size of the subset, and it is initialized with a null set, and $t = 0$ where t denotes the size of subset features, and it can be defined using Equation (2). After initializing the input variable, we have a subset of features $Z_t = z_b | b=1,2,3,\dots,t; Z_b \in X$ based on which the method refits the features. Let us define the subset of features using Equation (3). In Equation (4), assume p^+ to be the features that will find the $\arg \max (Z_t + z)$ here $z \in X - Z_t$ and maximize our selection criteria, which are associated with the classifier having the best score; score can be accuracy, mean absolute error (MAE), residual square R^2 on the output set of features that is Z_t . This process continues until we get the desired set h of features with a good score. The iterative process is described by Equations (5)–(7). Equation (5) updates the feature subset Z_{t+1} by adding the most significant feature p^+ to the existing subset Z_t . Equation (6) increments the variable t to continue the iterative process, and Equation (7) marks the termination of the process when t reaches the desired set of features h .

$$X = \{X_1, X_2, X_3, \dots, X_n\} \quad (1)$$

$$X_0 = \emptyset, \quad t = 0 \quad (2)$$

$$Z_t = \{z_b \mid b = 1, 2, 3, \dots, t; z_b \in X\}, \quad \text{where } t = \{1, 2, 3, \dots, n\} \quad (3)$$

$$p^+ = \arg \max (Z_t + Z) \text{ where } z \in X - Z_t \quad (4)$$

$$Z_{t+1} = Z_t + p^+ \quad (5)$$

$$t = t + 1 \quad (6)$$

$$t = h \quad (7)$$

Algorithm 1 shows the process of the forward selection method. We have input features X , and we want the best features. These features will be selected based on the value of score Z_t . This method starts with an empty set and then fits the model with a good score; simultaneously, features will be added and updated. This process will terminate when we get the desired set of features.

Algorithm 1 Forward feature elimination.

```

1: Input:  $X = \{X_1, X_2, X_3, \dots, X_n\}$ 
2: Output:  $Z_t = \{z_b \mid b = 1, 2, 3, \dots, t; Z_b \in X\}$ 
3: Start
4:   Prepare an empty array using (1)
5:   Evaluate the fitness of the best feature using (4)
6:   IF  $Z_t + p^+ > X_0$ 
7:     Update the features using (5)
8:   End If
9:    $t = t + 1$ 
10:  Repeat Step 2.
11:  Terminate using (7)
12: End

```

b. Backward Feature Elimination Method (BFE): BFE is a simple feature selection method that starts with a full model and deletes features until all remaining features have significant contributions. The least significant feature is deleted first, followed by refitting the model without the deleted feature and recompiling test statistics [36]. To understand the workings of this method, we have a set of features A with r size of dimensions that can be interpreted using Equation (8). We initialize the method using Equation (9) with a given set of features. Once the input variable is initialized, we have a subset of features D_w based on which method refits the features using Equation (10). Assume T^- to be the features that will find the $\arg \max (D_w - g)$ where $g \in A - D_w$ maximize our selection criteria that is associated with the classifier having the best score; score can be accuracy, MAE, r^2 on the set of features that is D_w . This process continues until we have the desired set L of features with a good score. Equations (11)–(14) detail the steps involved in the iterative feature elimination process.

$$A = \{A_1, A_2, A_3, \dots, A_r\} \quad (8)$$

$$B_0 = A, \quad w = r \quad (9)$$

$$D_w = \{g_i \mid i = 1, 2, 3, \dots, w; g_i \in A\}, \quad \text{where } w = \{1, 2, 3, \dots, m\} \quad (10)$$

$$T^- = \arg \max (D_w - g), \quad \text{where } g \in A - D_w \quad (11)$$

$$D_{w+1} = D_w - T^- \quad (12)$$

$$w = w + 1 \quad (13)$$

$$W = L \quad (14)$$

Algorithm 2 shows the process of the backward selection method. This method takes the full set of input features, calculates the score of classifiers, takes the features with good results, iteratively repeats step 3 until it achieves the desired number of features, and then terminates.

Algorithm 2 Backward feature elimination.

```

1: Input:  $A = \{A_{(1)}, A_{(2)}, A_{(3)}, \dots, A_r\}$ 
2: Output:  $D_w = \{g_i \mid i = 1, 2, 3, \dots, w; g_i \in A\}$ 
3: Start
4:   Begin with the full set of input features using (8)
5:   Evaluate the fitness of the best feature using (10)
6:   IF  $D_w - T^- > B_0$ 
7:     Update the features using (11)
8:      $w = w - 1$ 
9:   End If
10:  Repeat step 2
11:  Terminate using (14)
12: End

```

c. Recursive Feature Elimination (RFE): RFE is a method that selects the optimal feature subset based on the learned model and classification accuracy. We have calculated the feature importance using the training RF model. Algorithm 3 describes the process of RFE. This method works as a ranking procedure.

Algorithm 3 Recursive feature elimination.

```

1: Input:
2:   a. Training set  $W$ 
3:   b. Set of  $C$  features  $M = \{M_{(1)}, M_{(2)}, \dots, M_C\}$ 
4:   c. Ranking Method  $N(W, M)$ 
5: Output:
6:   Ranking  $J$ 
7: Start
8:   Initialize training set  $W$ 
9:   Repeat for  $i$  in  $\{1 : C\}$ 
10:    Set the Rank  $C$  using  $N(W, M)$ 
11:     $M^* \leftarrow$  last ranked feature in  $M$ 
12:     $J(C - i + 1) \leftarrow M^*$ 
13:     $M \leftarrow M - M^*$ 
14: End

```

3.3. Proposed Methodology

Unlike traditional research for detecting and predicting SKCM disease, our proposed research exploits ensemble methods (stacking, bagging, boosting, and voting). The proposed research includes the latest ensemble methods to predict SKCM disease using various ML classifiers and analyze the overall survival using the Kaplan–Meier and Cox proportional hazards regression models. To train ensemble methods, we initially create and train baseline classifiers (RF, GB, NB, LR, ET, AdaBoost, DT, LGBM). The performance was evaluated for accuracy, precision, recall, and F1 score. ROC curve and confusion matrix were generated to illustrate the performance. Following is the detail of baseline classifiers.

RF: The RF is a highly powerful ML classifier, which amalgamates diverse DT outputs using a majority voting mechanism. This technique increases the resilience of the solution, specifically in challenging problem domains. The overall prediction is derived by computing the average of the results generated by individual DTs.

GB: It is a powerful ensemble learning ML classifier. Unlike RF, which combines various DTs, the GB creates a sequential DT, with each subsequent tree correcting the errors. The classifiers optimize a loss function by iteratively adding weak learners, typically shallow DT, to the ensemble. Each tree is trained to emphasize the instances where the model performs poorly, gradually refining the overall predictive capability. The GB is

known for its high predictive accuracy and adaptability to various data types, making it a popular choice for classification and regression tasks.

NB: It is a simple classifier that leverages Bayes' theorem for predicting the unlabeled data points. It involves the computation of previous probabilities related to various classes and their application to the latest data. The simplicity and computational efficiency of GNB arise from the assumption of feature independence, making it a streamlined approach for classification tasks.

LR: It is used to predict the probability of the categorical data. LR utilizes a logistic function for calculating the probability in binary classification, where the output is dichotomous, representing two classes. It is also named the sigmoid function, which transforms the linear combination of input features into a value between 0 and 1, signifying the likelihood of belonging to a particular class. This makes LR particularly well-suited for problems with binary outcomes, such as in spam detection or medical diagnosis.

ET: It is an ensemble learning method related to the DT algorithm. Similar to RF, ET develops a forest of DTs for prediction. In ET, for each split of the DT, the ETs randomly select the feature to split on, leading to a higher level of diversity among individual trees in the ensemble. This increased randomness often results in a more robust model and can be particularly useful in mitigating overfitting. ET is famous for its efficiency and accuracy in handling high-dimensional data, making it a valuable classifier.

AdaBoost: This is a popular ensemble learning classifier for regression and classifications. It combines multiple weak learners' predictions to create an efficient and accurate prediction model. The algorithm assigns weights to each data point, and, in each iteration, it focuses on the misclassified instances, adjusting their weights to prioritize correct classification in the subsequent iteration.

DT: It is a highly recognized ML algorithm that evaluates the samples to categories as per their feature values. The DT creation process entails evaluating training samples and considering the most reliable features to partition the data into subsets, guided by principles like information gain or the Gini index. The motive is to create a tree capable of precisely predicting outcomes for new data based on the available features.

LGBM: This is an effective gradient-boosting framework exploited for advanced ML tasks. Unlike traditional GB methods, it uses a "leaf-wise" tree growth approach. This technique follows to expand the structure of the tree, integrating leaves that result in the maximum reduction of the loss function, ultimately leading to faster training times.

Algorithm 4 interprets the baseline classifiers and ensemble methods training in general. It depicts the working of ensemble models. Initially, we train base algorithms such as RF, DT, NB, GB, LGBM, LR, and AdaBoost. Then we train four ensemble methods that include stacking, bagging, boosting, and voting. Here, meta-algorithm H denotes the ensemble methods.

Figure 4 demonstrates the flow of our study. After data collection, the features with missing scores greater than 70% and less than 10% will be eliminated. Any features with missing scores less than 70% and greater than 10% will be included, and a new set of features will be defined. For selecting the best features among newly defined features, we applied three different feature selection methods: REF, FFE, and BFE. Our study evaluates four ensemble methods and an overall survival analysis using the Kaplan–Meier estimator and the Cox hazard regression model. To create ensemble methods, we must first create and train baseline classifiers.

Algorithm 4 ML-based ensemble methods.

```

1: Input:
2:   Training set  $R = \{(y_1, u_1), (y_2, u_2), \dots, (y_v, u_v)\}$ 
3:   Base algorithms  $G = \{G_1, G_2, \dots, G_s\}$ 
4:   Meta algorithm  $H$ 
5: Output: Ensemble Model
6: Start
7: Step-1:
8:   Train the base algorithms by applying algorithms  $G_i$  to  $R$ 
9:   For  $i = 1, 2, \dots, k$  do
10:     $E_i = G_i(R)$ 
11:   End For
12: Step-2:
13:   Generate a new dataset for making predictions  $R$ 
14:   For  $j = 1, 2, \dots, n$  do
15:    Classify the training samples  $x_j$ 
16:     $z_{ij} = E_i(x_j)$ 
17:   End For
18:    $R = \{y_j, u_j\}$ , where  $y_j = \{z_{1j}, z_{2j}, \dots, z_{sj}\}$ 
19: End For
20: Step-3:
21:   Train the meta-algorithm  $H$ 
22:    $H = G(R)$ 
23:   Return  $H$ 
24: End

```

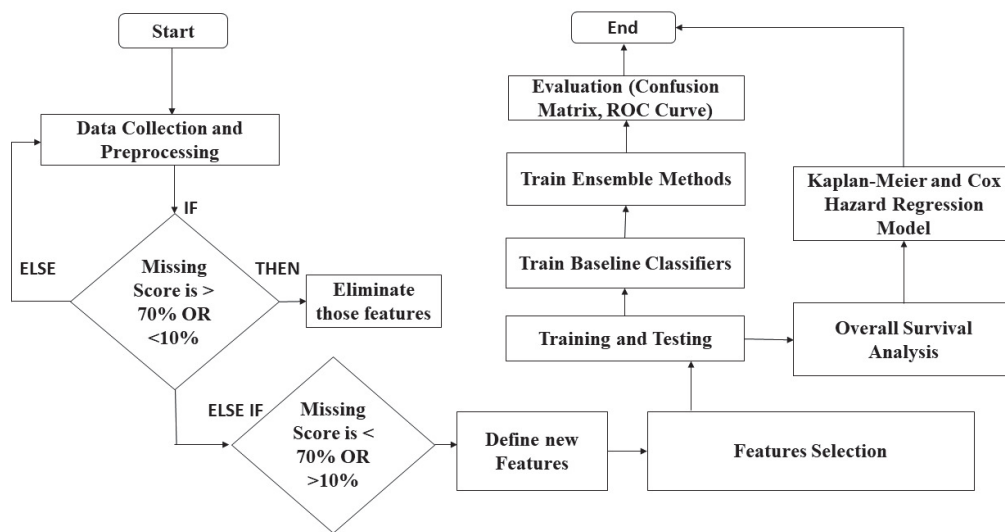


Figure 4. Flow chart of proposed study.

4. Experimental Results

The experiments in this study are conducted using Python programming language on Windows 10 @ 1.80 GHZ. The motivation of this study is to propose an ensemble model for the prediction of SKCM disease by utilizing different ML classifiers and to analyze overall survival using Kaplan–Meier and Cox hazard regression models. Initially, we applied three different feature selection methods, i.e., RFE, FFE, and BEF, to select the best features, as discussed in Section 3.

4.1. Survival Analysis Clinical Endpoint

The survival analysis with the log-rank test was examined in this study.

4.1.1. Kaplan–Meier Estimator

One of the most popular statistical methods used to estimate the likelihood of an event, such as death, a recurrence of a disease, the emergence of a new disease entity, or an adverse response, is survival analysis [37]. First, we need to understand the survival function to understand survival analysis. For example, consider Y as the random duration taken from the dataset under study as a duration that can be infinite but not a negative value, and the survival function can be denoted as $Pro(ti)$, where $Pro(ti)$ denotes the probability that an event has occurred or not yet at a time ti . It denotes the survival function calculated as Equation (15).

$$Pro(ti) = x(Y > ti) \quad (15)$$

The survival analysis can be achieved using the Kaplan–Meier estimator. Re-estimating the survival probability upon each event occurrence can be achieved using the Kaplan–Meier (KM) approach. This non-parametric method does not assume a specific distribution for the outcome variable, such as time. This approach is very simple, and complexity arises as the number of observations increases. We can say that the main idea of the KM approach, depending on the observed event time, is to split the estimation of the survival function into small chunks. The probability for each interval can be formulated using the following Equation (16):

$$Pro(ti) = \prod_{tia < t} \frac{Z_{(a-c_a)}}{Z_a} \quad (16)$$

where z_a denotes the number of patients whose lives are at risk at time t_{ia} , and c_a denotes the number of incidents that occurred in the event at a time t_{ia} (See Figure 5). Figure 5 shows the overall survival analysis of patients. We find the overall survival probability with significance (p -value is 0.05) of patients after diagnosing SKCM. The graph shows a higher probability of survival beyond the age of 20 and less than 20 years.



Figure 5. Overall survival (OS) analysis.

Figure 6 shows the survival probabilities of patients. We find that male patients have a higher probability than female patients. The male patients aged 20 years and below have higher (about 0.8 or 80%) survival probability. Above 80 years and somehow below 80 years, patients have less about (0.2 or 20%) of survival probability.

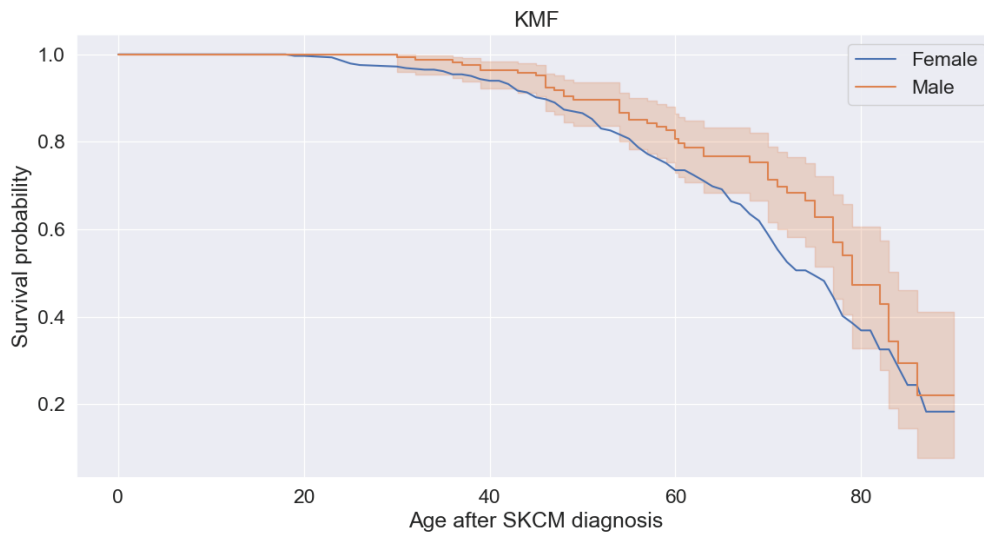


Figure 6. Survival probabilities.

Figure 7 depicts the survival analysis of different age groups of patients. We analyzed that patients in the age group greater than 30 years and less than or equal to 45 years and those in the age group equal to 45 years are nearly overlapping and have higher (about 0.6 or 60% and above) survival probability. For the patients in the age group less than or equal to 30, the curve shows step-wise increments in the probability starting near the survival probability (0.1 or above) and increasing steadily. All age groups overlap when survival probability reaches between 0.5 or 50% and up to 0.8 or 80%.

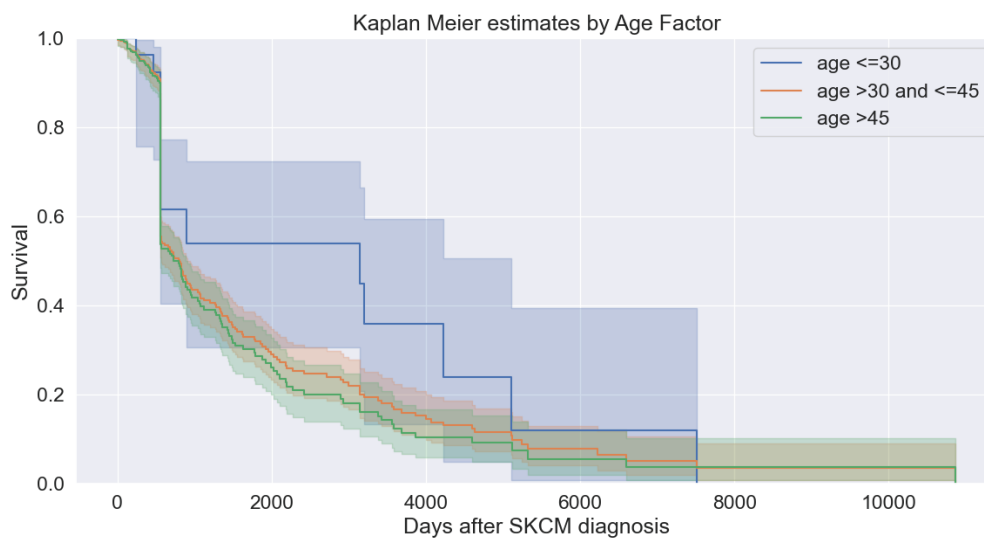


Figure 7. Survival analysis with different age groups.

Figure 8 describes the progression and complete remission of survival after diagnosis. For each cohort, there are two survival curves. We can observe that the progression curve increases in a step-wise curve. As the days passed, the probability of progression increased gradually. When survival probability reaches between 70% and 85%, both curves overlap. However, after diagnosis, patients start recovering.

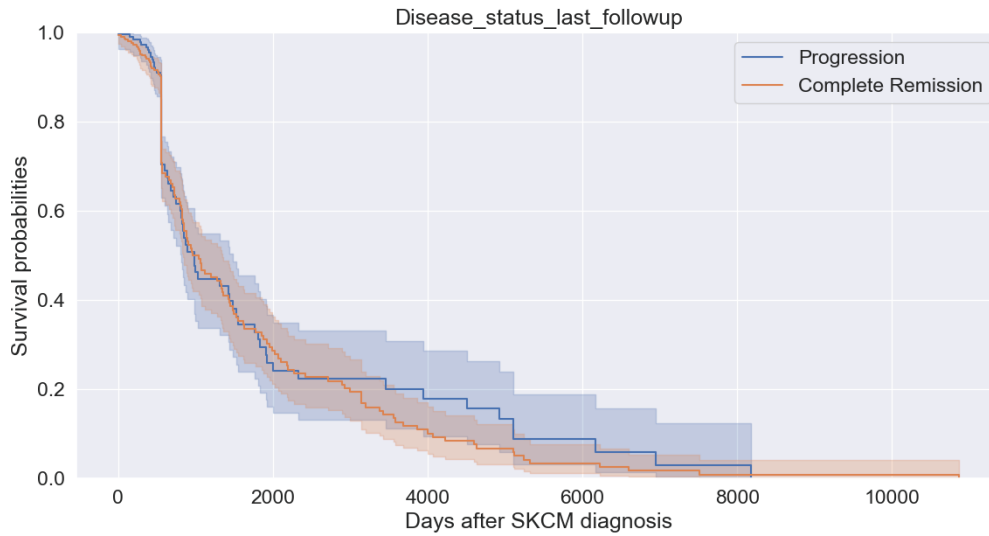


Figure 8. Survival probability with last follow up.

4.1.2. Cox Proportional Hazards Regression Model

The proportional hazards model, developed by David Cox in 1972 [38], uses the proportional risks assumption to produce reliable estimates of covariate effects. The Cox proportional hazards regression model is a semi-parametric approach for estimating weights in a proportional hazard model. It uses gradient descent to fit the data and minimizes errors. The model works by estimating the log hazard of patients as a linear function of their static covariates and a population-level baseline hazard function that changes over time [24]. It can be defined mathematically as Equation (17).

$$E(d|e) = h_0(d) \exp\left(\sum_{n=1}^e y_n(e_n)\right) \quad (17)$$

where

- d represents survival time;
- $E(d|e)$ represents the hazard function determined by a set of factors, i.e., $e_1, e_2, e_3, \dots, e_n$;
- $h_0(d)$ defines the baseline hazard function representing event probability when all covariates are zero. Hazard value equals 1 when all e_n are zero. The model assumes a parametric form for covariates' effect on hazard without baseline assumptions;
- $\exp\left(\sum_{n=1}^b y_n(e_n)\right)$ represents partial hazard as a time-invariant scalar factor that increases or decreases baseline hazard like the intercept in ordinary regression;
- The coefficients ($y_1, y_2, y_3, \dots, y_n$) measure the impact of covariates on a subject's hazard. The sign of the coefficient b affects the baseline hazard. A positive sign indicates higher risk, whereas a negative sign indicates lower risk. The magnitude of the coefficient b is estimated by maximizing partial likelihood. It assumes a proportional rate ratio throughout the study period, offering increased flexibility. This model can handle right-censored data but not left-censored or interval-censored data directly. The Cox model accepts the following three assumptions:
 1. A constant hazard ratio;
 2. The multiplicativity of explanatory variables;
 3. The independent failure times for individual subjects.

Table 4 describes the Cox proportional hazard regression. We evaluate the Cox hazard model and log-rank test to find the hazard ratio (HR) and significant association among the groups. We find that the value of the hazard ratio $HR < 1$, which means there is a reduction in the risk. The significance (p -value) < 0.05 is considered to find out the association among groups. We observed that the covariate Age at the last followup and interval has p -values

of 0.02 and 0.03, respectively, less than the significant p -value (0.05). We can say that there is an association between the groups.

Table 4. Cox hazard proportional method.

Covariate	Coef	Exp (Coef)	Se (Coef)	Coef Lower (95%)	Coef Upper (95%)	Exp Coef Lower (95%)	Exp Coef Upper (95%)	z	p	$\log_2(p)$
Sex	0.02	1.02	0.10	−0.17	0.21	0.84	1.23	1.02	0.10	−0.17
Status	−0.12	0.98	0.19	−0.39	0.35	0.68	1.41	0.98	0.19	−0.39
Disease status last followup	−0.00	0.98	0.10	−0.21	0.17	0.81	1.18	0.98	0.10	−0.21
Age at last followup	−0.50	0.61	0.02	−0.55	−0.46	0.58	0.63	0.61	0.02	−0.55
Diagnosis	−0.00	1.00	0.00	−0.01	0.00	0.99	1.00	1.00	0.00	−0.01
Interval	0.49	1.64	0.03	0.44	0.54	1.56	1.72	1.64	0.03	0.44

Figure 9 shows the hazard ratio (HR) for different covariates. We find that most of the covariates have $HR > 1$, meaning there is a risk reduction. Only for one covariate is there no effect, since $HR = 1$.

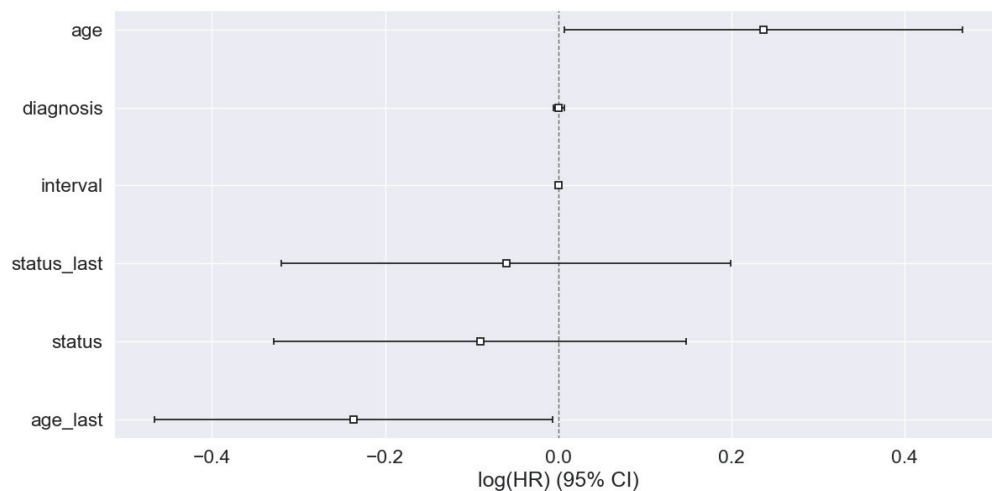


Figure 9. Hazard ratio.

4.1.3. ML-Based Ensemble Methods

We trained eight different ML classifiers to create ensemble models. To evaluate the performance of the proposed ML classifiers, we used two performance measures: ROC and confusion matrix. The results are presented in accuracy, precision, recall, F1 score, and ROC curve. The assessment includes four performance metrics: true positive (TP), denoting the accurate classification of ‘Positive Reputation’ in positive samples; false positive (FP), representing the misclassification of samples not belonging to the class; true negative (TN), indicating the accurate classification of negative samples; and false negative (FN), signifying the misclassification of samples as positive when they actually belong to the negative class.

Table 5 compares the performance of feature selection methods. We have trained an RF classifier with scoring matrix r^2 to select the best features. The performance of the RFE method is better than the other two methods.

Table 5. Feature selection methods and scores.

Feature Selection Methods	Score
Backward Feature Elimination	0.99993
Forward Feature Elimination	0.99988
Recursive Feature Elimination	0.99400

Table 6 describes the performance of eight different ML algorithms in terms of accuracy, precision, recall, and F1 score on the test dataset. It can be concluded that most of the algorithms achieved the highest accuracy rate of 98%. Only AdaBoost and GNB achieve 97% and 96% accuracy rates. It can be observed that the highest precision achieved by RF is 98%, while the highest recall rate obtained by LR, GB, RF, and light gradient boosting machine (LGBM) is 99%. LR, GB, and LGBM attained the highest F1 score rate. It is noteworthy to mention that the above performance metrics are evaluated on the test dataset.

Table 6. Performance metrics of ML algorithms.

ML Algorithms	Accuracy	Precision	Recall	F1 Score
LR	98%	97%	99%	98%
DT	98%	96%	98%	97%
GB	98%	97%	99%	98%
RF	99%	98%	99%	99%
ET	98%	97%	98%	97%
AdaBoost	97%	96%	95%	90%
LGBM	98%	97%	99%	98%
GNB	96%	88%	94%	94%

In Figure 10a, we present the confusion matrix for the GNB algorithm, showing its robust accuracy in correctly predicting 96.05% out of 482 samples, with only 3.95% samples being predicted inaccurately. On the other hand, in Figure 10b, we demonstrate the confusion matrix for the RF algorithm. There are a total of 482 samples, out of which 99.38% were accurately predicted while 0.62% were incorrectly forecasted. As compared to the other related studies, such as [9,10], our results are highly accurate.

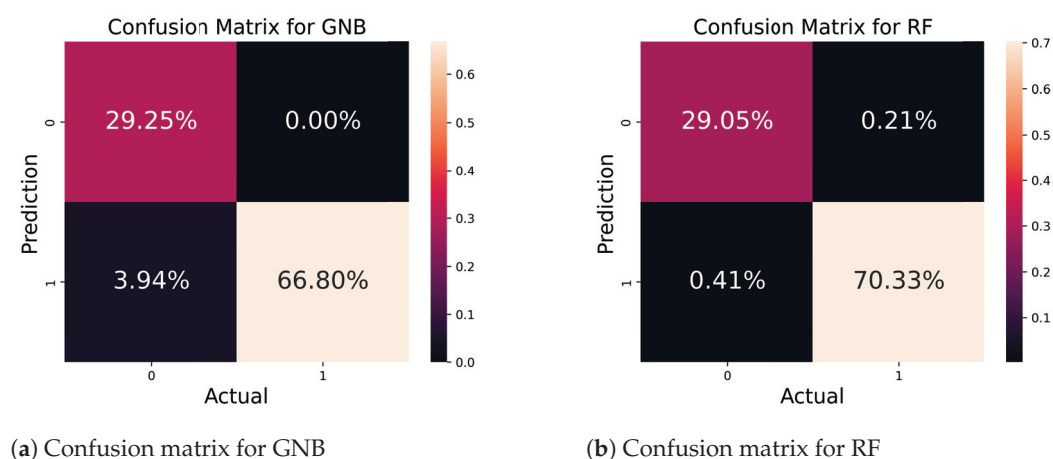
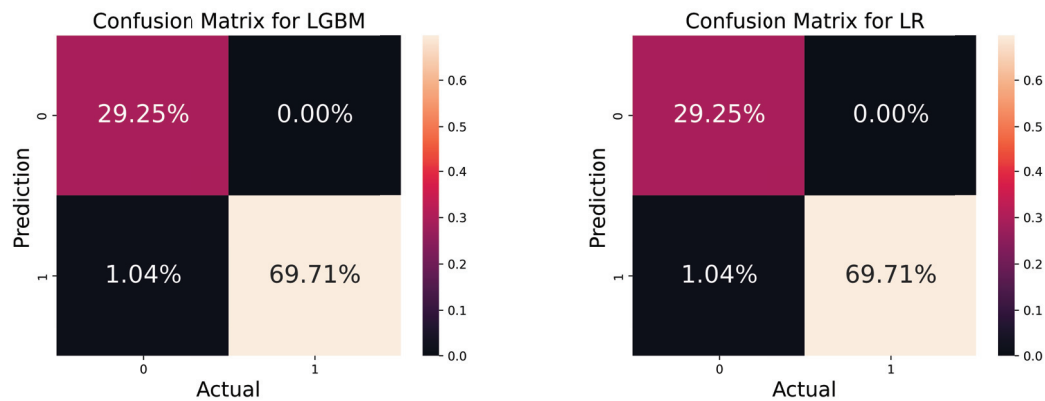


Figure 10. Confusion matrix for GNB and RF algorithms.

In Figure 11a, we present the confusion matrix for the LGBM; the performance of the classifier was evaluated on a total of 482 samples. The classifier successfully predicted 98.96% samples correctly; only 1.04% samples were incorrectly predicted. Figure 11b depicts the matrix for the LR algorithm. There were 482 samples, out of which 98.96% were precisely predicted while 1.04% of samples were wrongly predicted.

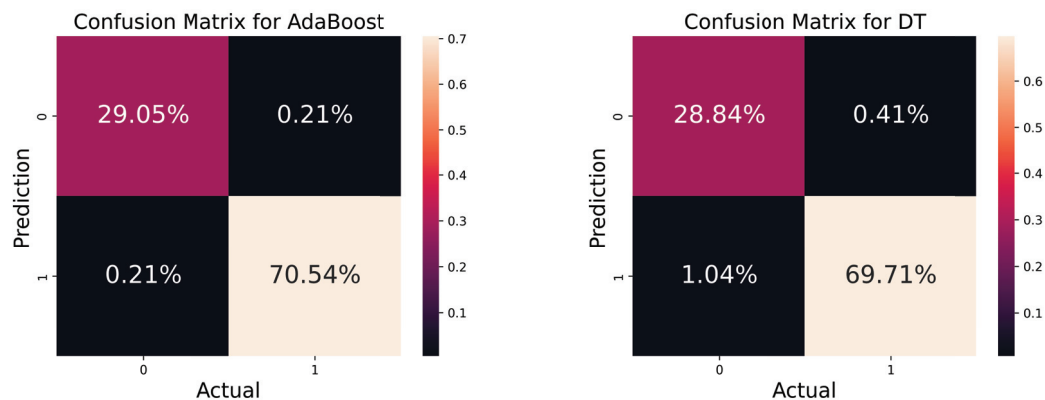


(a) Confusion matrix for LGBM

(b) Confusion matrix for LR

Figure 11. Confusion matrix for LGBM and LR classifiers.

In Figure 12a, we show the confusion matrix for the AdaBoost, where the classifiers impressively provide good accuracy by accurately predicting 99.59% out of a total of 100% on 482 samples. In contrast, the classifier wrongly predicted 0.41% samples. Figure 12b illustrates the matrix for the LR algorithm; the algorithm accurately predicted 98.13% out of 100% of samples, and only 1.87% of samples were incorrectly predicted. The classifiers' performance can be visualized from these insightful representations.



(a) Confusion matrix for AdaBoost

(b) Confusion matrix for DT

Figure 12. Confusion matrix for AdaBoost and DT classifiers.

Figure 13 depicts the confusion matrix for ET and GB algorithms. Figure 13a depicts the matrix for the ET algorithm. The classifier accurately predicted 98.76% out of 100% of samples, and only 1.24% of samples were wrongly predicted. Figure 13b determines the confusion matrix for the GB algorithm. The classifier correctly predicted 98.96% of 100% samples, and the classifier incorrectly predicted 1.04% of total samples.

Figure 14 illustrates the ROC curves for specificity (false positive rate) and sensitivity (true positive rate). The model can be classified or perform well if the ROC curve is turned to the upper left corner. Most of the classifiers turn toward the left upper corner, which means the classifiers perform well. From the below graph, we can say that LR, RF, AdaBoost, and LGBM achieve the highest accuracy, which is 0.99, whereas GB, DT, and extra tree achieve an accuracy of 0.98. Only the NB classifiers attained 0.97 accuracy. It can be concluded that there is a slightly small difference in the accuracy of different classifiers.

These ensemble methods were generated by training the above eight base ML classifiers. It can be observed that the stacking and voting method achieved the highest accuracy rate, which is 99%, as illustrated in Table 7. The highest precision rate recorded is 98% and

is obtained by the stacking method. The voting method attained the highest recall rate, as well as F1 score, which was 99%.

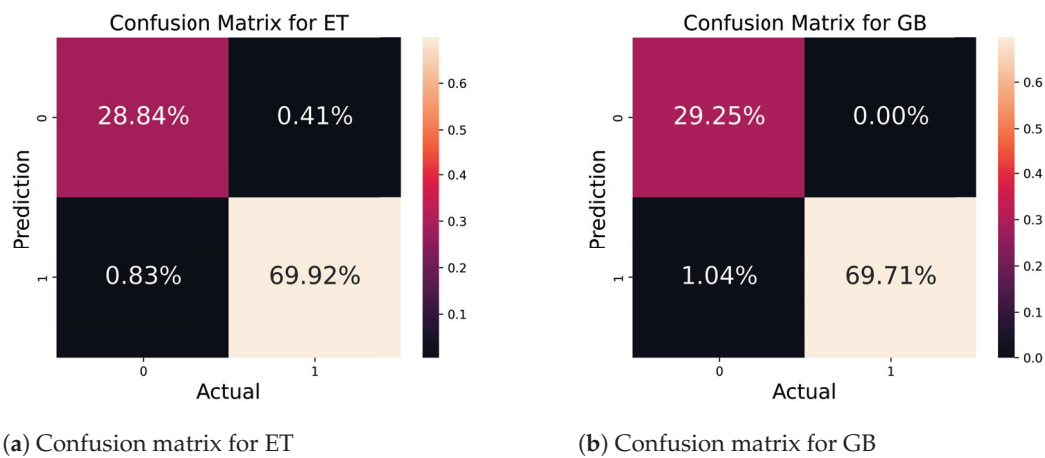


Figure 13. Confusion matrix for ET and GB classifiers.

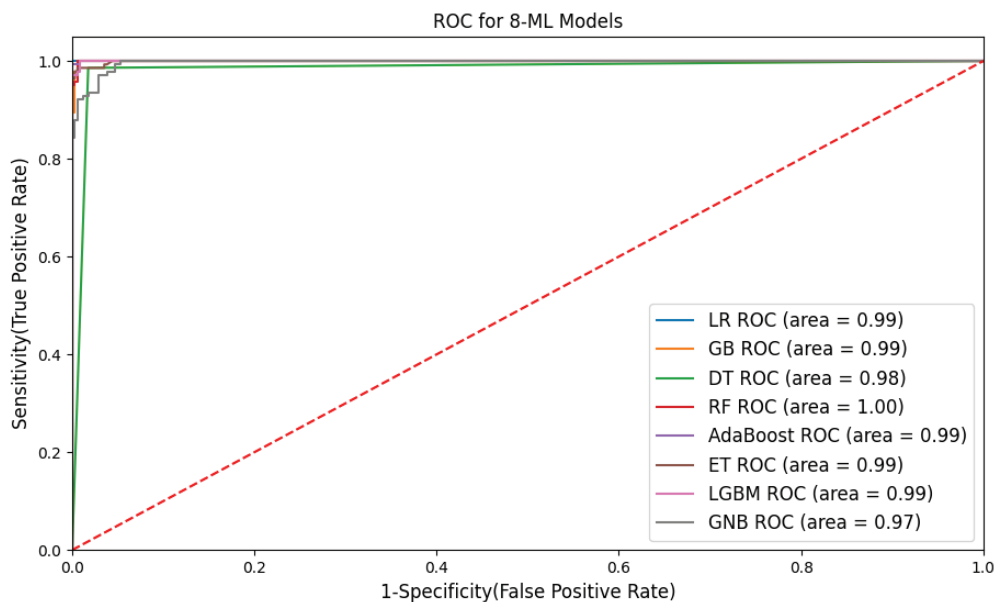


Figure 14. Consolidated ROC curve for ML classifiers.

Table 7. Performance metrics of ensemble methods.

Ensemble Methods	Accuracy	Precision	Recall	F1 Score
Bagging	98%	96%	96%	96%
Stacking	99%	98%	97%	96%
Boosting	97%	94%	96%	95%
Voting	99%	97%	99%	98%

Figure 15 portrays the matrix for the voting and stacking ensemble methods. From Figure 15a, out of 100% samples, 98.76% of samples were accurately predicted by this method, while 1.24% of samples were wrongly predicted. Figure 15b shows the matrix for the stacking method. This method correctly predicted 99.17% of samples, and only 0.83% of samples were predicted incorrectly.

Figure 16 portrays the matrix for the boosting and bagging ensemble methods. From Figure 16a, out of a total of 100% samples, 97.51% were correctly predicted while 2.49% were wrongly predicted. Figure 16b shows the matrix for the bagging method. Here, 97.09% of samples were predicted correctly, and only 2.91% of samples were wrongly predicted.

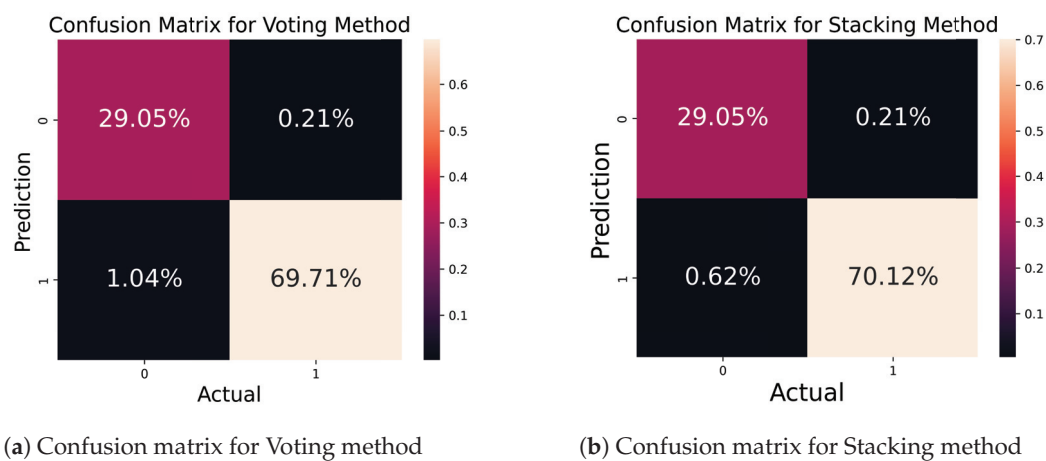


Figure 15. Confusion matrix for ensemble method.

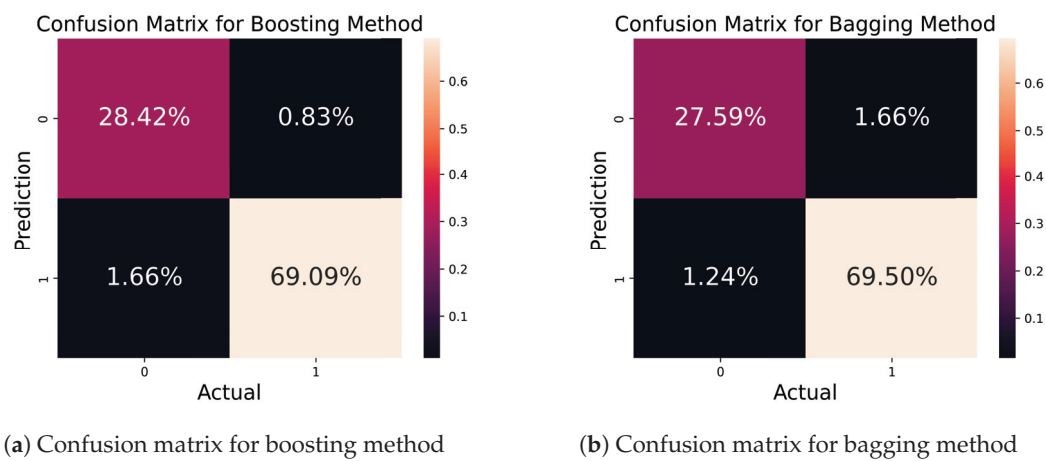


Figure 16. Confusion matrix for ensemble model.

4.2. Comparative Study

This study aims to propose an ensemble model for the prediction of SKCM cancer by utilizing ML classifiers. To create an ensemble model, we trained four different ensemble methods with eight different ML classifiers. The performance of ML classifiers and ensemble methods is compared and discussed below (see Figures 17 and 18).

Table 8 compares the performance of baseline classifiers and ensemble methods. We find that the stacking and voting ensemble methods achieved the highest performance as compared to baseline classifiers. Figure 17 illustrates the performance of different ensemble methods. The methods perform well. Furthermore, Figure 17 shows that stacking and voting outperform as compared to other methods.

Figure 18 represents the comparison among ML classifiers. The performance was evaluated in accuracy, precision, recall, and F1 score. However, all classifiers perform well, but the RF classifier outperforms all others. The main purpose of this study is to propose an ML-based ensemble method for the prediction of SKCM and to analyze the survival probability using the Kaplan–Meier and Cox proportional hazards regression models.

Table 8. Performance metrics of ML classifiers.

ML Classifiers	Accuracy	Precision	Recall	F1 Score
RF	99%	98%	99%	99%
GNB	96%	88%	94%	94%
LR	98%	97%	99%	98%
DT	98%	96%	99%	98%
GB	98%	97%	99%	98%
Adaboost	97%	96%	95%	90%
Extratree	98%	97%	98%	97%
LGBM	98%	97%	99%	98%
Bagging	98%	96%	96%	96%
Stacking	99%	98%	97%	96%
Boosting	97%	94%	96%	95%
Voting	99%	97%	99%	98%

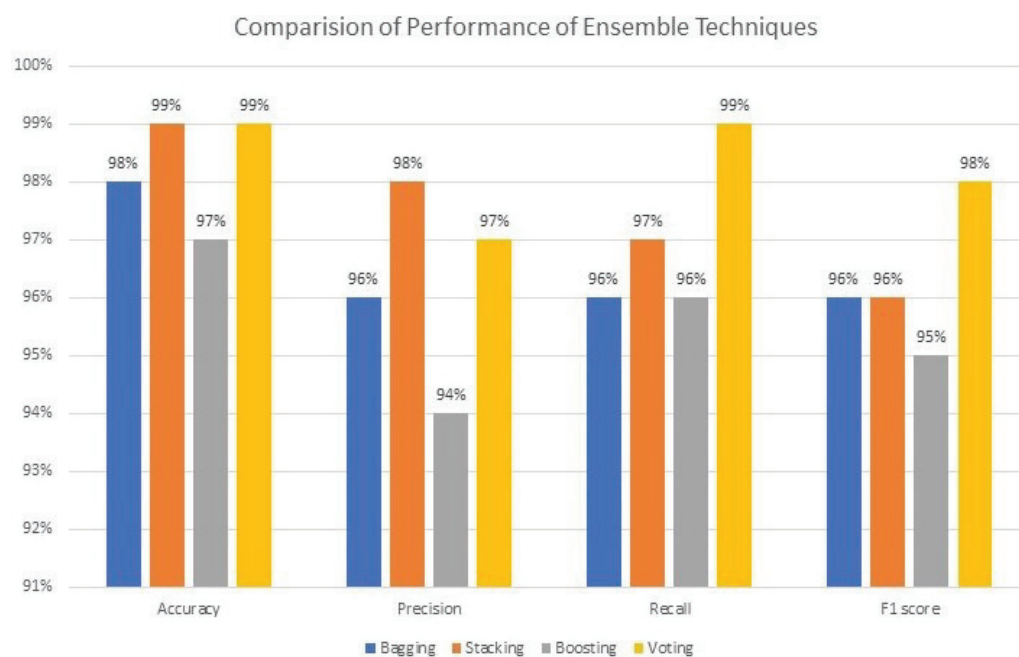


Figure 17. Performance of ensemble methods.

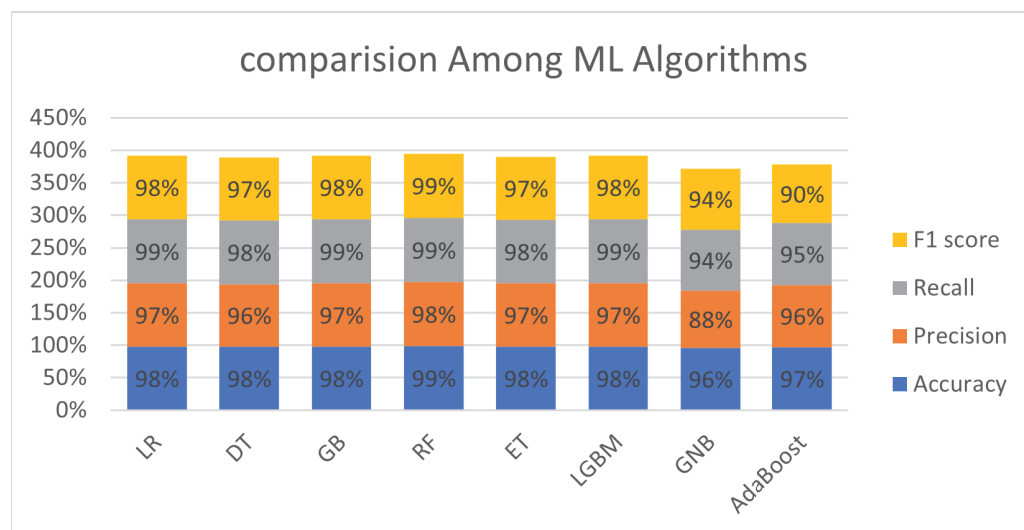


Figure 18. Comparison of different ML classifiers.

4.3. Baseline Classifiers Standard Error

Table 9 illustrates the standard error computed by eight baseline classifiers. This measure gauges the reliability of performance metrics like accuracy, precision, recall, and F1-score. Notably, an inverse relationship between accuracy and standard error was observed. Analysis of the table indicates that classifiers exhibiting the lowest standard error demonstrate greater stability and consistency in their performance across various metrics.

Table 9. Baseline classifiers standard error.

Baseline Classifiers	Standard Error
LR	0.45951
DT	0.45679
GB	0.459551
RF	0.45679
ET	0.45586
AdaBoost	0.45492
LGBM	0.45492
GNB	0.46127

Figure 19 displays the standard error associated with each baseline classifier. This metric offers crucial insights into the stability of model performance. Among the classifiers, GNB predicts a slightly higher standard error than others. Nonetheless, there are nuanced differences among the standard errors across the algorithms. AdaBoost and LGBM classifiers notably showcase the least standard error. A smaller standard error signifies a higher likelihood of consistent performance across various cross-validations, whereas higher standard errors suggest more variability in performance.

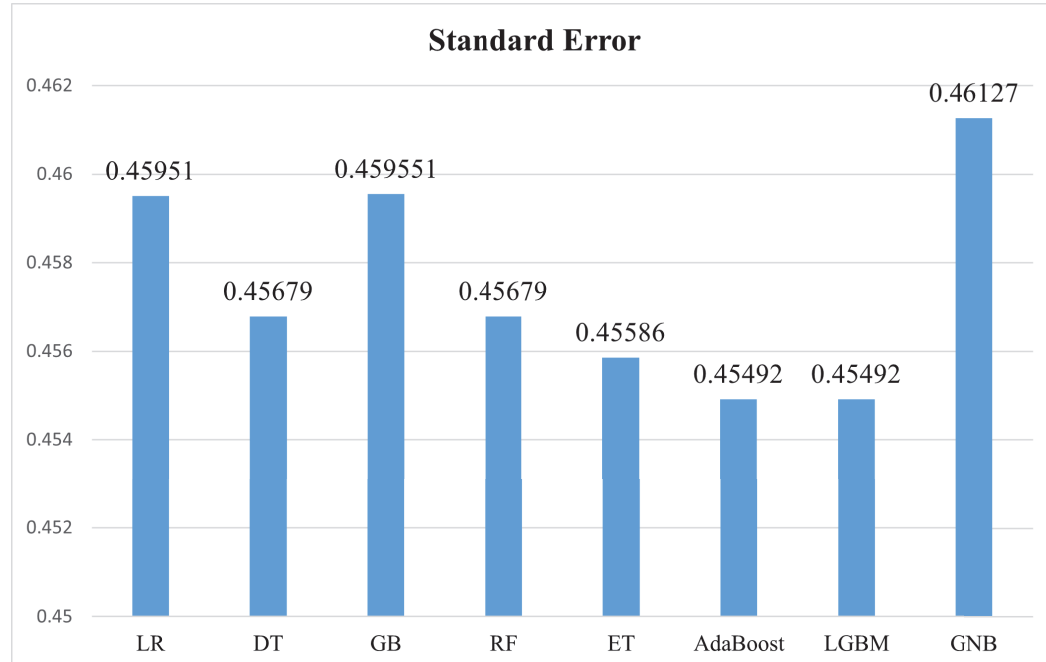


Figure 19. Baseline classifiers standard error.

Table 10 showcases the standard error forecasted by individual ensemble methods. Among these methods, the lowest standard error, at 0.45397, is observed in the bagging method, indicating superior performance compared to the other ensemble techniques.

Figure 20 illustrates the standard error produced by the four ensemble methods. A higher standard error signifies increased variability in the model's performance. Among these methods, the voting method attains the highest standard error of 0.45861, indicating

lower consistency in model performance. Conversely, the bagging method achieves the lowest standard error at 0.45397, signaling superior model performance.

Table 10. Ensemble methods standard error.

Ensemble Methods	Standard Error
Bagging	0.45397
Boosting	0.45771
Voting	0.45861
Stacking	0.45679

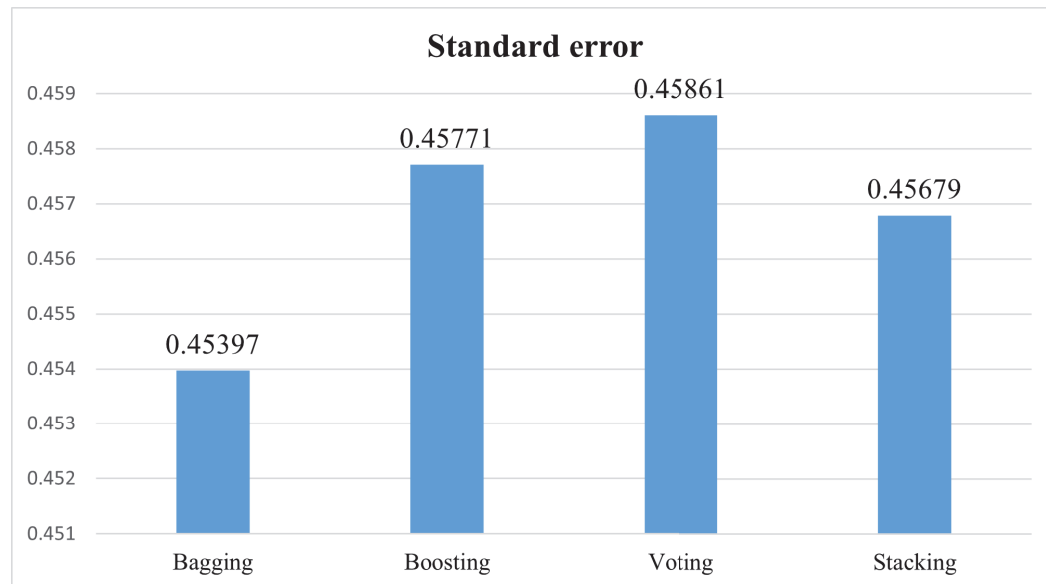


Figure 20. Ensemble methods standard error.

4.4. Discussion

Skin cancer is considered one of the most dangerous types of cancer. Many studies focus on early detection, treatment approaches, and suggesting prevention techniques. Numerous studies in the past have delved into these aspects, with a notable focus on employing ML and DL techniques that have yielded promising results for early detection and prognosis of the disease. Previous studies mainly focused on survival analysis and early detection using DL techniques. Our study proposes four ensemble methods (stacking, bagging, boosting, and voting) to predict SKCM and to analyze survival probability using KM and Cox proportional hazards regression models. In constructing our ensemble methods, which encompass stacking, bagging, boosting, and voting, we meticulously trained and tested eight baseline classifiers: RF, LR, DT, GB, ET, Adaboost, LightGBM (LGBM), and GNB. The performance of these methods was rigorously evaluated using a suite of metrics, including accuracy, precision, recall, F1 score, confusion matrix, and ROC curve. Remarkably, our results demonstrate a pinnacle of accuracy, reaching an impressive 99%, achieved by the stacking and voting ensemble methods. This exhibits the robustness and efficacy of the ensemble techniques employed in our study. Notably, among the individual algorithms, RF emerged as the top-performing classifier, depicting superior predictive capabilities. This exceptional performance across multiple metrics shows the potential applicability of our proposed ensemble methods in the realm of SKCM prediction. The high accuracy rates, especially with stacking and voting methods, suggest a synergistic enhancement of predictive power by combining diverse classifiers. Such findings hold significant implications for the development of more reliable and accurate predictive models in the context of skin cancer.

5. Conclusions

In this paper, Kaplan–Meier and Cox proportional hazards regression models are used to analyze overall survival, and ML-based ensemble methods are proposed to predict SKCM. Five distinct datasets using transcriptomic technologies were collected. To choose the best features, three distinct feature selection methods, i.e., REF, FFE, and BFE, were used. We trained and compared four ensemble approaches (stacking, bagging, boosting, and voting) using eight baseline classifiers (RF, DT, GNB, AdaBoost, GB, LR, ET, and LGBM). The performance of ensemble methods was evaluated with the help of the ROC curve, confusion matrix, accuracy, precision, recall, and F1 score. The overall performance of RF was good as compared to other classifiers. The recorded performance of the algorithms shows a slight variation. Voting and stacking strategies scored the best among ensemble techniques. The highest ROC was achieved using RF, LR, AdaBoost, and LGBM, which was 0.99. The RF classifier achieved the best accuracy, which was 99%, and the stacking and voting method achieved the highest accuracy rate, which was 99%. Finally, this study is limited to a specific dataset, which can be evaluated on various datasets to achieve better results.

Future Work

We will investigate deep learning methods for skin cancer early detection and prognosis in the future. We will investigate other multi-omic technologies in this area and investigate various skin cancers to further find out ways for early detection of diseases.

Author Contributions: Conceptualization, E.Y.A. and A.Z.; methodology, E.Y.A., Z.D., A.H.M., Q.A., K.K. and A.Z.; software, E.Y.A. and A.H.M.; validation, E.Y.A., Z.D. and A.Z.; formal analysis, K.K. and A.Z.; investigation, Z.D., A.H.M. and A.Z.; resources, E.Y.A., Z.D., A.H.M. and K.K.; data curation, E.Y.A., Z.D., A.H.M., Q.A. and A.Z.; writing—original draft, E.Y.A., A.H.M., Q.A. and K.K.; writing—review and editing, Q.A. and K.K.; visualization, Q.A.; supervision, Z.D. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Informed Consent Statement: Not applicable.

Data Availability Statement: The data in this study can be provided upon request.

Conflicts of Interest: The authors declare no conflicts of interest.

References

- Wang, X.; Xiong, H.; Liang, D.; Chen, Z.; Li, X.; Zhang, K. The role of SRGN in the survival and immune infiltrates of skin cutaneous melanoma (SKCM) and SKCM-metastasis patients. *BMC Cancer* **2020**, *20*, 378. [CrossRef] [PubMed]
- Ervik, F.; Ferlay, J.; Mery, L.; Soerjomataram, I.; Bray, F. *Cancer Today*; International Agency for Research on Cancer: Lyon, France, 2017.
- World Health Organization. *World Health Statistics*; Visual Summary; World Health Organization: Geneva, Switzerland, 2023.
- Naik, P.P. Cutaneous malignant melanoma: A review of early diagnosis and management. *World J. Oncol.* **2021**, *12*, 7. [CrossRef] [PubMed]
- Carr, S.; Smith, C.; Wernberg, J. Epidemiology and risk factors of melanoma. *Surg. Clin.* **2020**, *100*, 1–12. [CrossRef] [PubMed]
- Switzer, B.; Puzanov, I.; Skitzki, J.J.; Hamad, L.; Ernstoff, M.S. Managing metastatic melanoma in 2022: A clinical review. *JCO Oncol. Pract.* **2022**, *18*, 335–351. [CrossRef] [PubMed]
- Wu, Y.; Chen, B.; Zeng, A.; Pan, D.; Wang, R.; Zhao, S. Skin cancer classification with deep learning: A systematic review. *Front. Oncol.* **2022**, *12*, 893972. [CrossRef] [PubMed]
- Leiter, U.; Keim, U.; Garbe, C. Epidemiology of skin cancer: Update 2019. In *Sunlight, Vitamin D and Skin Cancer*; Springer: New York, NY, USA, 2020; pp. 123–139.
- Tang, Y.; Feng, H.; Zhang, L.; Qu, C.; Li, J.; Deng, X.; Zhong, S.; Yang, J.; Deng, X.; Zeng, X.; et al. A novel prognostic model for cutaneous melanoma based on an immune-related gene signature and clinical variables. *Sci. Rep.* **2022**, *12*, 20374. [CrossRef] [PubMed]
- Cozzolino, C.; Buja, A.; Rugge, M.; Miatton, A.; Zorzi, M.; Vecchiato, A.; Del Fiore, P.; Tropea, S.; Brazzale, A.; Damiani, G.; et al. Machine learning to predict overall short-term mortality in cutaneous melanoma. *Discov. Oncol.* **2023**, *14*, 13. [CrossRef]

11. Dildar, M.; Akram, S.; Irfan, M.; Khan, H.U.; Ramzan, M.; Mahmood, A.R.; Alsaiani, S.A.; Saeed, A.H.M.; Alraddadi, M.O.; Mahnashi, M.H. Skin cancer detection: A review using deep learning techniques. *Int. J. Environ. Res. Public Health* **2021**, *18*, 5479. [CrossRef]
12. Son, H.M.; Jeon, W.; Kim, J.; Heo, C.Y.; Yoon, H.J.; Park, J.U.; Chung, T.M. AI-based localization and classification of skin disease with erythema. *Sci. Rep.* **2021**, *11*, 5350. [CrossRef]
13. Verma, A.K.; Pal, S.; Kumar, S. Comparison of skin disease prediction by feature selection using ensemble data mining techniques. *Inform. Med. Unlocked* **2019**, *16*, 100202. [CrossRef]
14. Guo, P.; Xue, Z.; Mtema, Z.; Yeates, K.; Ginsburg, O.; Demarco, M.; Long, L.R.; Schiffman, M.; Antani, S. Ensemble deep learning for cervix image selection toward improving reliability in automated cervical precancer screening. *Diagnostics* **2020**, *10*, 451. [CrossRef] [PubMed]
15. Mamun, M.; Farjana, A.; Al Mamun, M.; Ahammed, M.S. Lung cancer prediction model using ensemble learning techniques and a systematic review analysis. In Proceedings of the 2022 IEEE World AI IoT Congress (AIoT), Seattle, WA, USA, 6–9 June 2022; pp. 187–193.
16. ICGC Data Portal—Skin Cutaneous Melanoma (SKCM)—US Project. Available online: <https://dcc.icgc.org/releases/current/Projects/SKCM-US> (accessed on 27 November 2023).
17. Aamir, S.; Rahim, A.; Aamir, Z.; Abbasi, S.F.; Khan, M.S.; Alhaisoni, M.; Khan, M.A.; Khan, K.; Ahmad, J. Predicting breast cancer leveraging supervised machine learning techniques. *Comput. Math. Methods Med.* **2022**, *2022*, 5869529. [CrossRef] [PubMed]
18. Shah, S.A.; Tahir, A.; Ahmad, J.; Zahid, A.; Pervaiz, H.; Shah, S.Y.; Ashleibta, A.M.A.; Hasanali, A.; Khattak, S.; Abbasi, Q.H. Sensor fusion for identification of freezing of gait episodes using Wi-Fi and radar imaging. *IEEE Sens. J.* **2020**, *20*, 14410–14422. [CrossRef]
19. Magsi, A.H.; Mohsan, S.A.H.; Muhammad, G.; Abbasi, S. A Machine Learning-Based Interest Flooding Attack Detection System in Vehicular Named Data Networking. *Electronics* **2023**, *12*, 3870. [CrossRef]
20. Magsi, A.H.; Ghulam, A.; Memon, S.; Javeed, K.; Alhussein, M.; Rida, I. A Machine Learning-Based Attack Detection and Prevention System in Vehicular Named Data Networking. *Comput. Mater. Contin.* **2023**, *77*, 1445–1465. [CrossRef]
21. Trang, K.; Nguyen, H.A.; TonThat, L.; Do, H.N.; Vuong, B.Q. An Ensemble Voting Method of Pre-Trained Deep Learning Models for Skin Disease Identification. In Proceedings of the 2022 IEEE International Conference on Cybernetics and Computational Intelligence (CyberneticsCom), Malang, Indonesia, 16–18 June 2022; pp. 445–450.
22. Verma, A.K.; Pal, S.; Tiwari, B. Skin disease prediction using ensemble methods and a new hybrid feature selection technique. *Iran J. Comput. Sci.* **2020**, *3*, 207–216. [CrossRef]
23. Thanka, M.R.; Edwin, E.B.; Ebenezer, V.; Sagayam, K.M.; Reddy, B.J.; Günerhan, H.; Emadifar, H. A hybrid approach for melanoma classification using ensemble machine learning techniques with deep transfer learning. *Comput. Methods Programs Biomed. Update* **2023**, *3*, 100103.
24. Bradburn, M.J.; Clark, T.G.; Love, S.B.; Altman, D.G. Survival analysis part II: Multivariate data analysis—An introduction to concepts and methods. *Br. J. Cancer* **2003**, *89*, 431–436. [CrossRef]
25. Shorfuzzaman, M. An explainable stacked ensemble of deep learning models for improved melanoma skin cancer detection. *Multimed. Syst.* **2022**, *28*, 1309–1323. [CrossRef]
26. Alam, T.M.; Shaikat, K.; Khan, W.A.; Hameed, I.A.; Almuqren, L.A.; Raza, M.A.; Aslam, M.; Luo, S. An efficient deep learning-based skin cancer classifier for an imbalanced dataset. *Diagnostics* **2022**, *12*, 2115. [CrossRef]
27. Alwakid, G.; Gouda, W.; Humayun, M.; Sama, N.U. Melanoma detection using deep learning-based classifications. *Healthcare* **2022**, *10*, 2481. [CrossRef] [PubMed]
28. Ali, M.S.; Miah, M.S.; Haque, J.; Rahman, M.M.; Islam, M.K. An enhanced technique of skin cancer classification using deep convolutional neural network with transfer learning models. *Mach. Learn. Appl.* **2021**, *5*, 100036. [CrossRef]
29. Naeem, A.; Anees, T.; Fiza, M.; Naqvi, R.A.; Lee, S.W. SCDNet: A Deep Learning-Based Framework for the Multiclassification of Skin Cancer Using Dermoscopy Images. *Sensors* **2022**, *22*, 5652. [CrossRef] [PubMed]
30. Huang, M.; Zhang, Y.; Ou, X.; Wang, C.; Wang, X.; Qin, B.; Zhang, Q.; Yu, J.; Zhang, J.; Yu, J.; et al. m5C-related signatures for predicting prognosis in cutaneous melanoma with machine learning. *J. Oncol.* **2021**, *2021*, 6173206. [CrossRef] [PubMed]
31. Agrahari, P.; Agrawal, A.; Subhashini, N. Skin cancer detection using deep learning. In *Futuristic Communication and Network Technologies: Select Proceedings of VICFCNT 2020*; Springer: Berlin/Heidelberg, Germany, 2022; pp. 179–190.
32. Wang, Y.; Singh, L. Analyzing the impact of missing values and selection bias on fairness. *Int. J. Data Sci. Anal.* **2021**, *12*, 101–119. [CrossRef]
33. Mera-Gaona, M.; Neumann, U.; Vargas-Canas, R.; López, D.M. Evaluating the impact of multivariate imputation by MICE in feature selection. *PLoS ONE* **2021**, *16*, e0254720. [CrossRef] [PubMed]
34. Hambali, M.A.; Oladele, T.O.; Adewole, K.S. Microarray cancer feature selection: Review, challenges and research directions. *Int. J. Cogn. Comput. Eng.* **2020**, *1*, 78–97. [CrossRef]
35. He, Z.; Li, L.; Huang, Z.; Situ, H. Quantum-enhanced feature selection with forward selection and backward elimination. *Quantum Inf. Process.* **2018**, *17*, 154. [CrossRef]
36. Chowdhury, M.Z.I.; Turin, T.C. Variable selection strategies and its importance in clinical prediction modelling. *Fam. Med. Community Health* **2020**, *8*, e000262. [CrossRef]

37. Abd ElHafeez, S.; D'Arrigo, G.; Leonardis, D.; Fusaro, M.; Tripepi, G.; Roumeliotis, S. Methods to analyze time-to-event data: The Cox regression analysis. *Oxidative Med. Cell. Longev.* **2021**, *2021*, 1302811. [CrossRef]
38. Nikulin, M.; Wu, H.D. *The Cox Model and Its Applications*; Springer: Berlin/Heidelberg, Germany, 2016.

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.

Article

Machine Learning-Based Ensemble Feature Selection and Nested Cross-Validation for miRNA Biomarker Discovery in Usher Syndrome

Rama Krishna Thelagathoti, Dinesh S. Chandel, Wesley A. Tom, Chao Jiang, Gary Krzyzanowski, Appolinaire Olou and M. Rohan Fernando *

Molecular Diagnostic Research Laboratory, Center for Sensory Neuroscience, Boys Town National Research Hospital, Omaha, NE 68010, USA; ramakrishna.thelagathoti@boystown.org (R.K.T.); wesley.tom@boystown.org (W.A.T.); chao.jiang@boystown.org (C.J.); gary.krzyzanowski@boystown.org (G.K.); appolinaire.olou@boystown.org (A.O.)

* Correspondence: m.rohan.fernando@boystown.org

Abstract: Usher syndrome (USH) is a rare genetic disorder affecting vision, hearing, and balance. Identifying reliable biomarkers is crucial for early diagnosis and understanding disease mechanisms. MicroRNAs (miRNAs), key regulators of gene expression, hold promise as biomarkers for USH. This study aimed to identify a minimal subset of miRNAs that could serve as biomarkers to effectively differentiate USH from controls. We employed ensemble feature selection techniques to select the top miRNAs appearing in at least three algorithms. Machine learning models were trained and tested using this subset, followed by validation on an independent 10% sample. Our approach identified 10 key miRNAs as potential biomarkers for USH. To further validate their biological relevance, we conducted pathway analysis, which revealed significant pathways associated with USH. Furthermore, our approach achieved high classification performance, with an accuracy of 97.7%, sensitivity of 98%, specificity of 92.5%, F1 score of 95.8%, and an AUC of 97.5%. These findings demonstrate that combining ensemble feature selection with machine learning provides a robust strategy for miRNA biomarker discovery, advancing USH diagnosis and molecular understanding.

Keywords: ensemble feature selection; biomarker discovery; usher syndrome; miRNA; machine learning; nested cross-validation

1. Introduction

Usher syndrome is a rare genetic disorder characterized by a combination of hearing loss, progressive vision loss due to retinitis pigmentosa and, in some cases, balance issues [1,2]. It is the most common cause of inherited deaf-blindness, accounting for approximately 50% of cases where individuals experience both hearing and vision impairment [3]. Usher syndrome presents in various clinical subtypes, each with varying severity and onset of symptoms, making its diagnosis particularly complex [4]. The heterogeneity in symptoms and the various genetic mutations associated with the syndrome make diagnosis challenging, often requiring a combination of clinical assessments, audiological tests, and genetic screening [5,6].

In recent years, microRNAs (miRNAs) have emerged as promising biomarkers for diagnosing various genetic and complex diseases, including Usher syndrome [7,8]. miRNAs are small, non-coding RNA molecules that regulate gene expression and play crucial roles

in various cellular processes [9,10]. Abnormal miRNA expression has been linked to a wide range of diseases, including cancers, neurological disorders, and genetic syndromes [10]. Given their stability in biological fluids and their specificity to certain pathological states, miRNAs have gained attention as potential biomarkers for early and accurate diagnosis. In the context of Usher syndrome, profiling miRNA expression can offer insights into the molecular mechanisms underlying the disorder and potentially aid in its detection [7,11].

While miRNAs hold significant promise for diagnosis, not all miRNAs contribute equally to disease progression. Identifying a minimal subset of miRNAs that are most relevant to disease is critical for both diagnostic accuracy and biological understanding [12–14]. A smaller miRNA set improves the interpretability of diagnostic models, aiding in better understanding the disease's molecular underpinnings [15–17]. Clinically, a minimal miRNA set reduces the complexity and cost of diagnostic tests, making them more feasible for large-scale or routine screening [18]. Moreover, focusing on a smaller, highly relevant miRNA subset facilitates the development of targeted therapies and personalized treatments [15].

Feature selection techniques are instrumental in reducing the dimensionality of miRNA datasets by identifying the most relevant miRNAs associated with any disease, including Usher syndrome [12,13,15]. Given the high dimensionality of miRNA expression profiles, selecting a minimal feature set is essential. Rare genetic disorder studies often have relatively small sample sizes. Feature selection methods, such as filter, wrapper, and embedded techniques, can help identify the most impactful miRNAs [19]. Therefore, choosing a compact set of features is crucial for developing robust and interpretable models. In the case of Usher syndrome, which is also a rare genetic disorder, applying feature selection to miRNA data can help isolate the key miRNAs associated with the disorder, facilitating a more efficient and accurate diagnostic process.

Given the complexity of Usher syndrome and the vast miRNA datasets generated from expression profiling, automating the diagnostic process with machine learning classifiers is essential. In this study, we propose a machine learning-based approach that integrates ensemble feature selection and nested cross-validation to identify the minimal miRNA feature set needed for the automated detection of Usher syndrome. This study has two main objectives: (1) to design and develop an ensemble feature selection method combined with nested cross-validation to identify the minimal miRNA set for classification; and (2) to utilize a range of supervised machine learning classifiers to train, test, and validate the models to identify the best-performing one. This approach not only reduces the time and labor involved in manual diagnosis, but also improves the accuracy and reliability of predictions. Furthermore, nested cross-validation is particularly beneficial when working with small datasets, where data acquisition is often a challenge.

2. Relevant Work

Ensemble feature selection has proven to be a promising approach in miRNA-based disease classification, particularly for high-dimensional datasets. Studies highlight its ability to identify minimal miRNA subsets that improve diagnostic accuracy and reduce dimensionality. For example, Cai et al. (2015) employed an ensemble method with multiple classifiers to differentiate lung cancer samples from controls, achieving enhanced robustness in classification [20]. Similarly, Sarkar et al. (2021) leveraged machine learning integrated with survival analysis to identify breast cancer subtype-specific miRNA biomarkers, demonstrating improved diagnostic and prognostic precision [15]. These methods underscore the value of ensemble strategies in addressing the challenges of high-dimensional biological data.

Lopez-Rincon et al. (2020) [21] applied an ensemble recursive feature selection approach to identify circulating miRNAs as biomarkers for cancer classification across various tumor types. This method improved the interpretability and reliability of cancer classification by pinpointing a minimal subset of miRNAs relevant to different tumor types [21]. In a related study, Lopez-Rincon et al. (2019) [12] developed an ensemble feature selection framework to identify a 100-miRNA signature for cancer classification. Their automated feature selection approach demonstrated the scalability of ensemble feature selection in cancer classification. However, the large number of selected features (100 miRNAs) may not be suitable for diseases associated with fewer miRNAs, such as rare genetic disorders [12]. Colombelli et al. (2022) [22] proposed a hybrid ensemble feature selection method for identifying miRNA biomarkers from transcriptomic data. This hybrid design enabled a more comprehensive selection of candidate biomarkers, enhancing model accuracy and interpretability [22].

The current study overcomes the limitations of previous research by incorporating an adaptive ensemble feature selection method combined with nested cross-validation for miRNA-based classification, specifically targeting Usher syndrome. This novel approach ensures robust feature selection and validation, addressing overfitting by employing nested cross-validation throughout both the feature selection and classification processes. Furthermore, our method dynamically updates the minimal feature set as new data becomes available, making it particularly well-suited for rare genetic disorders like Usher syndrome, which often suffer from small sample sizes. This integration of dynamic adaptability and scalability represents a significant advancement in miRNA-based disease classification, offering a more reliable and flexible framework than prior methods.

3. Materials and Methods

This section outlines the methodology employed in this study. As shown in Figure 1, the process begins with preprocessing the miRNA samples, followed by feature selection and model training using an ensemble feature selection approach combined with nested cross-validation. This methodology helps identify the best-performing model and the minimal subset of miRNA features. The selected model is then evaluated and validated using these features to assess its performance. Additionally, the minimal set of miRNAs is used for pathway analysis to extract relevant biological pathways. Furthermore, this section provides an overview of the feature selection techniques and machine learning classification algorithms used in this study.

3.1. miRNA Samples Extraction and Quantification

For each sample in the study, total miRNA was extracted using methods described previously [7]. Briefly, QIAzol reagent was used to isolate total RNA, and miRNA was samples in this study were collected from patient derived B-lymphocyte cell lines. MiRNA was purified from cell lines using miRNeasy Tissue/Cells Advanced Micro Kit (QIAGEN Sciences Inc., Germantown, MD, USA). MicroRNA expression was quantified for all samples using NanoString nCounter Human v3 miRNA Expression assays (cat. # CSO-MIR3-12, Bruker Corp., Billerica, MA, USA). Quality control and batch normalization of miRNA count data was performed using Nanostring quality Control dasHbOard (NACHO) package in R [23]. This resulted in a miRNA count matrix for 798 miRNAs for 60 samples, which was utilized for subsequent machine learning analysis. There were 31 Usher samples, and 29 control samples in the cohort.

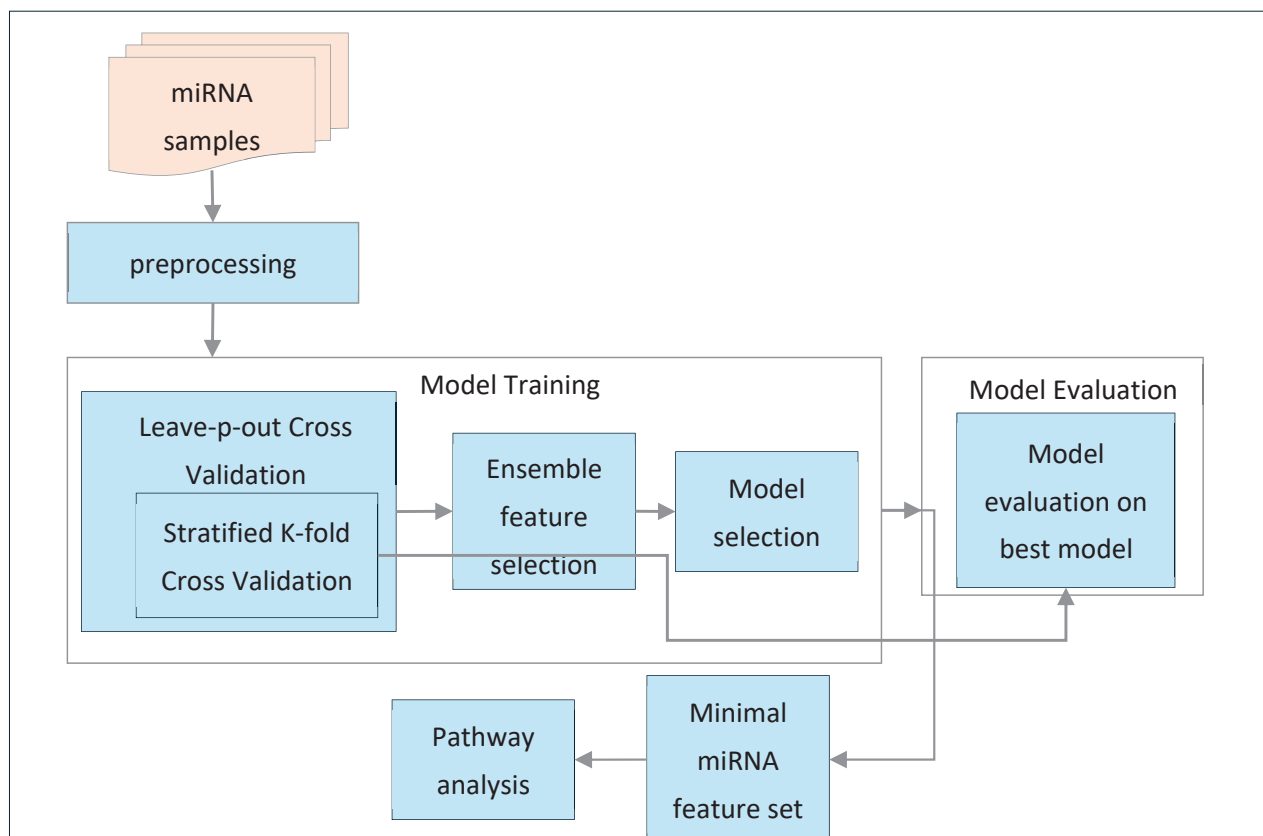


Figure 1. Overview of the methodology.

3.2. Feature Selection

Feature selection, also known as feature reduction or variable subset selection, is the process of identifying a minimal subset of the most relevant features from a larger set of features in a dataset [12,24]. When working with high-dimensional data, such as miRNA data, it becomes essential to select a small, optimal subset that captures the critical information of the entire dataset without significant loss of detail. This process is fundamental in building robust learning models. The importance of feature selection has been widely studied in fields like bioinformatics and pattern recognition [25,26]. In general, feature selection techniques are designed to minimize overfitting and enhance model performance, leading to improved predictive accuracy in supervised classification and better cluster detection in clustering tasks. Furthermore, they contribute to the creation of faster and more cost-efficient models while also providing valuable insights into the underlying processes that produced the data [16,27].

In the realm of classification, feature selection methods can be classified into three categories based on their selection strategy of the features: filter methods, wrapper methods, and embedded methods. Filter methods select the relevance of features based on certain properties such as statistical significance. Features are selected before choosing any machine learning model. These techniques are computationally efficient and help eliminate irrelevant features, allowing for a simplified model without risking overfitting [28]. Wrapper methods involve using a specific machine learning algorithm to evaluate the performance of different feature subsets by training the model repeatedly. This approach often yields better feature selection tailored to the model, but it can be computationally expensive [29]. Embedded methods integrate feature selection within the model training process, allowing for simultaneous feature selection and model optimization. This approach typically results

in models that are both efficient and accurate, as they consider feature interactions during the selection process [30].

When working with miRNA expression profile datasets, selecting the most relevant features (miRNAs) is essential for constructing accurate and interpretable models [12]. Due to the high dimensionality of miRNA datasets and the typically small sample sizes, effective feature selection methods can greatly enhance model performance and mitigate the risk of overfitting [24]. Research has shown that ensemble feature selection methods combined with cross-validation techniques can effectively identify an optimal minimal subset of features. This approach enhances model performance and ensures robust validation on unseen data while reducing the risk of overfitting, outperforming single-feature selection techniques [15]. In this study, we propose an ensemble feature selection approach combined with nested cross-validation to identify the minimal miRNA feature set for classification.

3.3. Ensemble Feature Selection Algorithms

Ensemble feature selection identifies an optimal feature set by combining results from multiple individual feature selection algorithms [31]. This approach differs from single-feature selection strategies, which identify key features using a singular method. For example, single methods might filter out low-variance features or recursively eliminate those that do not contribute to model performance. In contrast, ensemble selection integrates optimal feature sets from various techniques to find the best overall feature set. Previous studies show that ensemble learning results in more robust classification outcomes and produces a superior optimal feature set compared to single-feature selection approaches [12,15–17,20–22]. In this study, we selected four distinct feature selection methods, each representing a different category of techniques, to ensure a comprehensive approach to feature selection. Recursive Feature Elimination (RFE), taken from wrapper methods, Random Forest feature importance (RF), an embedded method, the k-best method (k-best), a filter technique, and Least Absolute Shrinkage and Selection Operator (LASSO) to introduce regularization by penalizing certain coefficients.

3.3.1. Recursive Feature Elimination

Recursive Feature Elimination (RFE) is a wrapper-based feature selection method that iteratively selects the most important features by recursively removing the least important ones [32]. The principle of RFE involves training a model using all available features, evaluating their importance scores, and then recursively eliminating the least important features until the desired number of features is reached [33]. The goal of RFE is to identify an optimal subset of features from the complete list of features in the dataset. Initially, a supervised learning model is trained using the full feature set to predict the target values. Afterward, the importance of each feature is evaluated based on the model's learned parameters, with less important features identified by their scores. The feature with the lowest importance score is then eliminated from the dataset. This process continues iteratively, with the model being retrained after each removal, until the desired number of features is selected.

3.3.2. Random Forest Feature Importance

Random Forest (RF) feature importance belongs to the embedded feature selection family that evaluates the significance of each feature based on its contribution to the prediction in a Random Forest model [34]. The primary goal of RF feature importance is to identify and rank features based on their predictive power, allowing for the selection of the most influential features. A Random Forest model is trained by constructing multiple decision trees using bootstrap samples from the dataset. During the training process, the importance of each feature is evaluated based on its contribution to the model's predictive accuracy,

typically using metrics such as Mean Decrease Gini (MDG) or Mean Decrease Accuracy (MDA) [35,36]. This method effectively selects important features while retaining the benefits of the Random Forest's robustness against overfitting.

3.3.3. Least Absolute Shrinkage and Selection Operator (LASSO)

LASSO is a regression analysis technique that combines both variable selection and regularization to improve the prediction accuracy and interpretability of a statistical model. It works by minimizing a specific objective function, which includes a regularization term. In LASSO regression, L1 regularization is applied, which adds a penalty to the model based on the absolute values of the coefficients. This regularization encourages sparsity in the model by pushing the coefficients of less important features towards zero. As the regularization parameter increases, more coefficients are driven towards zero, effectively removing less relevant predictors from the model. The result is a simpler and more interpretable model that tends to generalize better when applied to unseen data. This makes LASSO a powerful tool for feature selection and model simplification [37].

3.3.4. K-Best Feature Selection

K-Best is a filter-based feature selection method that assesses the relevance of each feature with respect to the target variable. The process starts by evaluating all features in the dataset using statistical tests like the Chi-squared test, ANOVA F-value, or mutual information. The goal is to identify the top k features that provide the most useful information for predicting the target variable. The selection process can be outlined as follows:

1. For each feature, compute a score using a chosen statistical measure.
2. Sort the features based on their scores in descending order.
3. Select the top k features with the highest scores.

K-Best feature selection helps reduce the dimensionality of the dataset by removing less informative features. This results in improved model performance while retaining the most relevant features for predictive modeling [28].

3.4. Overview of Machine Learning Techniques

This section presents the list of ML techniques we have utilized in this study including Logistic Regression (LR), Random Forest (RF), Support Vector Machine (SVM), Extreme Gradient Boosting (XGB), and Ada Boost (ADB).

3.4.1. Logistic Regression

Logistic regression is a statistical method used for binary classification, where it models the relationship between a dependent binary variable and one or more independent variables. The method estimates probabilities using the logistic function [38]. In this model, the probability that the target variable is equal to 1, given the features, is calculated. The coefficients for each feature are estimated by maximizing the likelihood function, which measures how well the model fits the observed outcomes [39]. Logistic Regression can be used to classify miRNA expression data by estimating the probability that a given miRNA profile corresponds to a specific disease class. Its simplicity and ability to handle binary outcomes make it a valuable tool for predicting the presence of disease based on miRNA expression levels.

3.4.2. Random Forest

Random Forest is an ensemble learning method used for classification. It builds multiple decision trees from bootstrapped samples of the dataset and averages their predictions [40]. For a given dataset, a Random Forest creates multiple decision trees, each using

a random subset of features. During training, the algorithm minimizes the Gini impurity or entropy at each node to effectively split the data [35,36]. Random Forest is particularly effective for miRNA data classification due to its ability to handle high-dimensional datasets and its robustness against overfitting. It identifies important miRNAs by averaging predictions from multiple decision trees, making it a useful tool for classifying diseases based on miRNA profiles.

3.4.3. Support Vector Machine (SVM)

A Support Vector Machine (SVM) is a supervised learning algorithm that constructs a hyperplane in a high-dimensional space to separate different classes [41]. It aims to find the optimal hyperplane that maximizes the margin between the classes. The SVM is particularly effective for classification problems with complex and high-dimensional data, such as miRNA expression profiles. The SVM is ideal for distinguishing between disease and healthy profiles due to its ability to handle non-linear relationships. The kernel trick allows SVMs to capture these non-linearities by transforming the input space into a higher-dimensional feature space, making it a powerful method for miRNA data classification [42].

3.4.4. Extreme Gradient Boosting (XGBoost)

XGBoost is an optimized gradient boosting algorithm that builds models in a sequential manner, where each new model corrects the errors made by the previous ones. It aims to improve the performance of weak learners by iteratively adjusting the model to enhance prediction accuracy [43]. The algorithm's objective function combines the loss function and a regularization term, ensuring that the model is both accurate and simple, avoiding overfitting. XGBoost is highly effective for miRNA classification tasks, as it boosts the performance of weak learners by iteratively improving prediction accuracy. Additionally, its ability to handle missing values and irregular data makes it a strong choice for miRNA-based disease classification, allowing it to effectively work with complex datasets [44].

3.4.5. Adaptive Boosting (AdaBoost)

AdaBoost is an ensemble learning method that combines multiple weak classifiers to form a strong classifier [45]. It works by training classifiers sequentially, with each new classifier focusing on the errors made by the previous ones. The algorithm assigns weights to each instance, increasing the weights of misclassified instances at each iteration, helping the model improve over time. The final model is a weighted sum of the predictions from each weak classifier, where each classifier's weight is determined by its accuracy. AdaBoost is particularly useful in miRNA classification because it enhances the model's ability to predict disease. It improves the accuracy of classifiers, even when miRNA data are complex or imbalanced, making it a strong choice for disease classification tasks where data may be noisy or challenging [46].

3.5. Ensemble Feature Selection with Nested Cross-Validation

In this section, we introduce the ensemble feature selection approach combined with nested cross-validation, a robust methodology designed to identify the most relevant features while minimizing overfitting. Cross-validation is a technique that involves partitioning data into multiple subsets. Some of these subsets are utilized for training the model, while the others are reserved for testing or validation. This process continues until every subset has served as both a training and validation set [41]. A commonly used method is k-fold cross-validation, where the dataset is divided into k-equally sized subsets (folds). In this method, the model is trained on k-1 folds and tested on the remaining fold.

This procedure is repeated k times, ensuring that each fold is tested at least once [47]. By employing different data subsets for training and testing, cross-validation offers a more accurate estimate of a model's performance on unseen data [48]. Nested cross-validation involves multiple levels of cross-validation, often structured as an outer loop and an inner loop [49]. This technique helps reduce overfitting and enhances overall model performance.

Following algorithm describes our proposed ensemble feature selection with nested cross-validation as shown in Algorithm 1.

Algorithm 1: Ensemble feature selection with nested cross-validation

Input: miRNA expression dataset $D \in \mathbb{R}^{N \times F}$, where N is the number of samples and F is the number of miRNA features.

Output:

- Minimal miRNA feature set F_{minimal}
- Best-performing model M^*
- Mean performance metrics across all validation sets

Step 1. Initialization

$F_{\text{minimal}} \leftarrow \emptyset$

$M^* \leftarrow \text{None}$

Let $p = \frac{N}{10}$ (i.e., Leave-6-Out Cross-Validation when $N = 60$)

Generate $p = 10$ non-overlapping folds:

$\{(T_i, V_i)\}_{i=1}^p$, where $T_i \in \mathbb{R}^{(N-p) \times F}$, $V_i \in \mathbb{R}^{p \times F}$

Step 2. Outer Cross-Validation (Leave-p-Out)

for each $i \in \{1, 2, \dots, p\}$ **do**

Let T_i be the outer training set and V_i be the outer validation set

Step 3. Inner Cross-Validation (Stratified k-Fold) and Feature Selection

Split T_i into k stratified folds : $\{(t_j, v_j)\}_{j=1}^k$

for each $j \in \{1, 2, \dots, k\}$ **do**

Feature Selection on t_j :

Apply RFE, Random Forest importance, LASSO, and SelectKBest

Model Training:

Train classifiers (LR, RF, SVM, XGBoost, AdaBoost) on t_j

Model Evaluation:

Evaluate on v_j using Accuracy, Sensitivity, Specificity, F1 Score, and

AUC

Model Selection:

Choose best-performing model M_j

Update Feature Set:

Add features to F_{minimal} if selected in ≥ 3 inner folds

Select most frequent model across inner folds as:

$M^* = \text{argmax}_{M_j} (\text{frequency of selection in inner folds})$

Step 4. Model Validation on Outer Fold

Use M^* and F_{minimal} to classify V_i

Evaluate performance using Accuracy, Sensitivity, Specificity, F1 Score,

and AUC

Return:

- F_{minimal} —final feature set
 - M^* —best-performing model
 - Average performance metrics over all V_i , $i = 1, 2, \dots, p$
-

The methodology begins with the input of a miRNA expression dataset, which contains both the samples and the associated features necessary for classification. The first step involves initializing key variables: N , the number of samples; F , the number of miRNA features; and an empty set called F_{minimal} to store the minimal set of miRNA features identified during the analysis. Additionally, a variable M^* is initialized to hold the best-performing model after evaluation. Two sets are initialized to facilitate the nested cross-validation process: outer training sets, which consists of the outer training sets (T_1, T_2, \dots, T_p), and validation sets, which includes the corresponding validation sets (V_1, V_2, \dots, V_p).

The second step involves executing Leave-P-Out Cross-Validation (LPOCV) as shown in Figure 2, where the input data are divided into outer training and validation sets. This method generates multiple splits, with each outer training set (T_i) containing 90% of the total data, while the corresponding validation sets (V_i) consist of the remaining 10%. This process ensures that each instance in the dataset is eventually used for validation, providing a robust evaluation framework for model performance.

In the third step, the focus shifts to the inner cross-validation and feature selection process. For each outer training set T_i , stratified k-fold cross-validation is employed to further divide the data. As shown in Figure 2, each outer training set (T_i) is split into multiple inner training sets (t_1, t_2, \dots, t_i) and corresponding inner test sets (v_1, v_2, \dots, v_i). Feature selection is then performed on each inner training set using various methods, such as Recursive Feature Elimination (RFE), Random Forest (RF) feature importance, LASSO, and the k-best method using the ANOVA F-score as the statistical indicator to rank features based on their discriminative power. Simultaneously, different classifiers, including Logistic Regression (LR), Random Forest (RF), Support Vector Machines (SVM), Extreme Gradient Boosting (XGB), and AdaBoost (ADB), are trained on these inner training sets. The performance of each model is evaluated against its corresponding inner test set using key metrics like accuracy, sensitivity, specificity, F1 score, and AUC. The model that achieves the best mean performance across all metrics for the current outer training set is designated as M^* . In addition to that, miRNA features that consistently appear at least 3 times across all iterations are appended to the F_{minimal} .

The final step involves validating the selected model and feature set. For each validation set V_i , the M^* identified from the inner cross-validation is used to perform classification, utilizing the minimal miRNA feature set derived from the previous step. The performance of this model is then evaluated on the validation sets using the same metrics: accuracy, sensitivity, specificity, F1 score, and AUC.

Ultimately, the output of this comprehensive methodology includes the minimal miRNA feature set consistently selected across all inner k-folds, the best-performing model based on mean performance across all inner train and test sets, and the mean metrics of this best model evaluated against the validation sets. This structured approach ensures a thorough and reliable framework for identifying significant miRNA features and achieving accurate classification outcomes.

3.6. Finding Enriched Pathways

3.6.1. miRNA Gene Target Prediction

TargetScanHuman version 8.0 was used to predict genes with binding sites matching the 10 miRNAs from the model, resulting in 6115 genes with matching miRNA target sites [50]. Of these, 572 gene targets had cumulative weighted context scores (CWCS) less than -0.5 , indicating strong gene suppression. All 6115 unique gene targets were included in gene ontology enrichment analysis and metabolic pathway analysis [50].

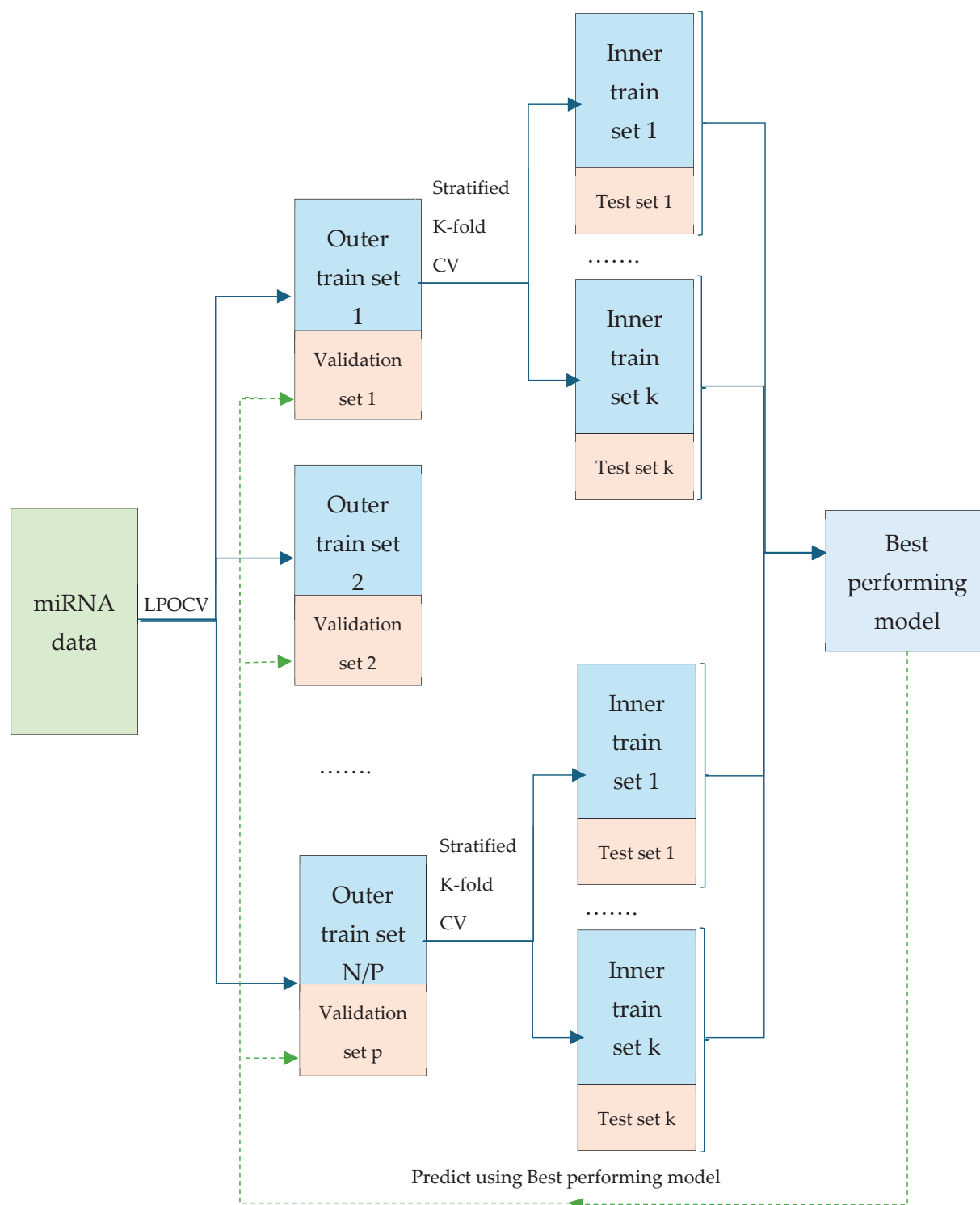


Figure 2. Overview of ensemble feature selection with nested cross-validation.

3.6.2. Gene Ontology Enrichment and Pathway Analysis

The R package ‘clusterProfiler’ was used to identify gene ontologies (GOs) and pathways which might be influenced by miRNAs from the model [51]. For gene ontology enrichment analysis, the ‘enrichGO()’ function was used with an adjusted *p*-value cutoff of 0.05 and a minimum gene set size of 5. R’s ‘enrichR’ package was used to identify pathways affected by miRNAs using the ‘enrichr()’ function against the Reactome pathway database [52,53]. Putative affected pathways were determined to be significant at an adjusted *p*-value threshold of 0.05.

4. Results

This section presents the results in four key aspects that highlight the effectiveness of our proposed approach. First, we present the minimal miRNA biomarker set identified from the ensemble feature selection method. Second, we highlight the best model selected during the training phase and its performance metrics. Third, evaluation of selected best model and its performance on validation sets. Finally, we present the biological pathway analysis based on the selected miRNA features.

4.1. Selected Minimal miRNA Feature Set

We identified a minimal set consisting of 10 miRNAs through the ensemble feature selection method combined with nested cross-validation. These miRNAs significantly contribute to the classification of Usher syndrome when compared to control samples. The selected miRNAs are: hsa-miR-148a-3p, hsa-miR-183-5p, hsa-miR-146a-5p, hsa-miR-28-5p, hsa-miR-30c-5p, hsa-miR-551b-3p, hsa-miR-642a-5p, hsa-miR-181a-5p, hsa-miR-28-3p, hsa-miR-182-5p.

The SHAP summary plot shown in Figure 3, illustrates the contribution of specific miRNAs to the model's output for distinguishing between Usher syndrome and control samples. Each dot represents a SHAP value for a given miRNA in a sample, with color indicating the miRNA's expression level (blue for low and red for high). The position on the x -axis shows the impact of the feature on the model's prediction: positive values push the model towards predicting Usher syndrome, while negative values push towards control. The feature importance plot (in Figure 4) reveals the contribution of individual miRNAs to the classification decision between Usher syndrome patients and control samples.

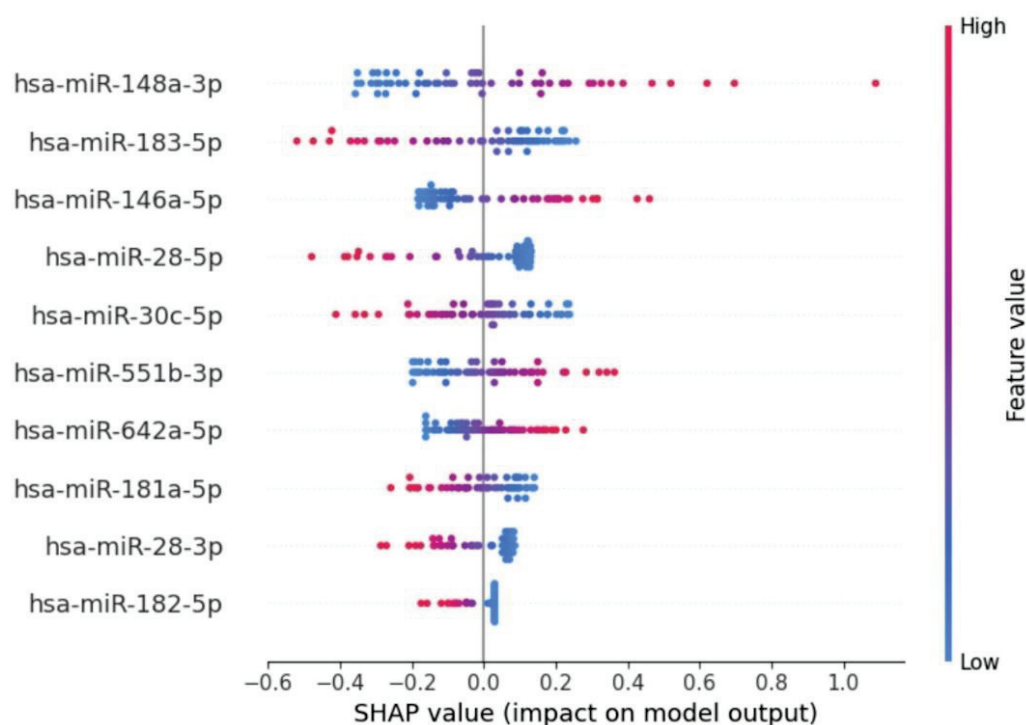


Figure 3. SHAP summary plot with selected miRNA features.

Key observations for each miRNA:

hsa-miR-148a-3p: Highly ranked in SHAP feature importance, contributing significantly toward USH prediction with high SHAP values. In other words, higher expression levels (red) increase the SHAP value positively, strongly contributing to Usher syndrome prediction.

hsa-miR-183-5p: Both high and low expression values are seen across the SHAP value spectrum, indicating variable contributions. However, overall miRNA expression levels are downregulated in usher compared to control.

hsa-miR-146a-5p: Contributes to usher prediction with positive SHAP values. Expression levels are generally upregulated in usher and downregulated in control.

hsa-miR-28-5p, hsa-miR-28-3p, and hsa-miR-182-5p: These miRNAs exhibit similar expression pattern where positive SHAP pushes towards usher prediction and negative SHAP values indicates significant contribution towards control prediction. Moreover, Lower SHAP values indicate that expression levels are significantly downregulated in usher compared to control.

hsa-miR-551b-3p and hsa-miR-642a-5p: Show a consistent pattern of positive SHAP values, indicating their importance in usher prediction. These miRNAs are upregulated in usher and downregulated in control.

hsa-miR-30c-5p and hsa-miR-181a-5p: Both miRNAs were expressed at higher levels in controls and reduced in usher, contributing negatively to SHAP values for control classification.

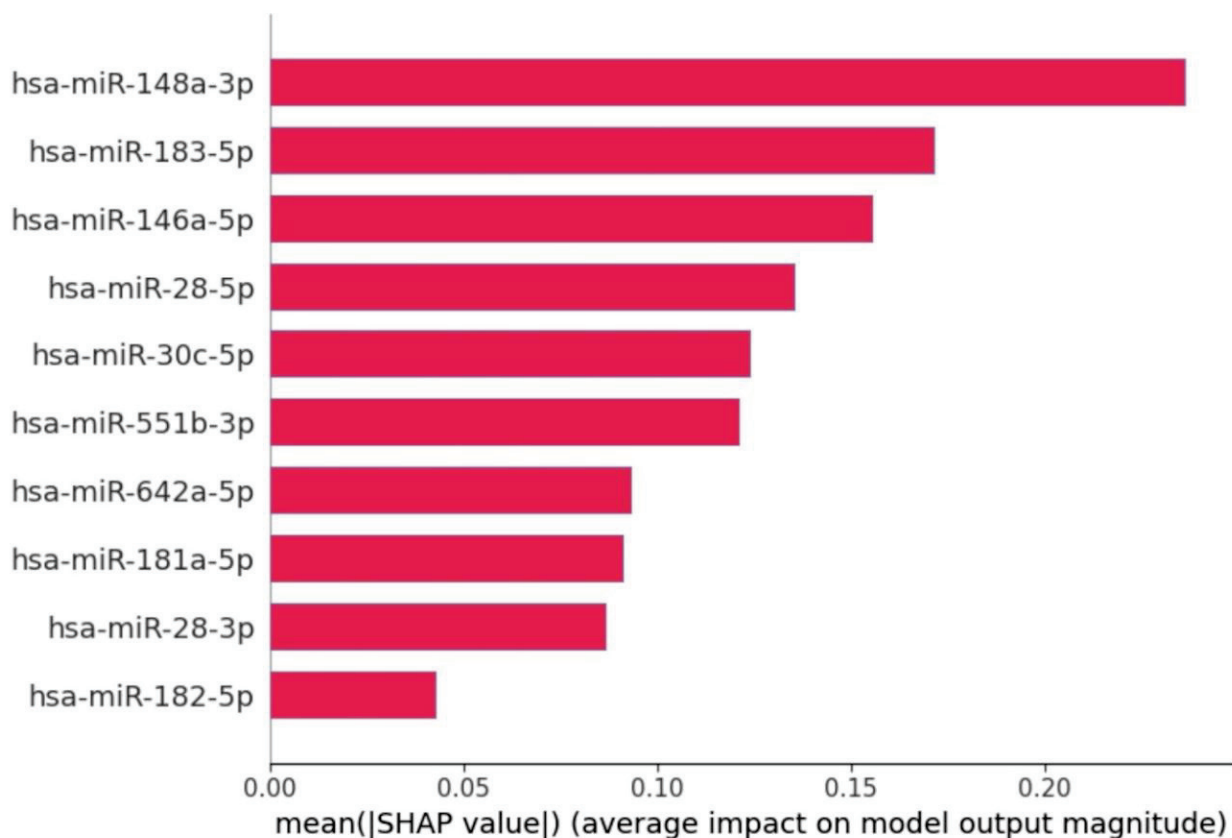


Figure 4. SHAP feature importance plot that shows importance of each selected feature.

4.2. Model Training Results

For model training, five machine learning classifiers were employed for classifying Usher syndrome using miRNA data: Logistic Regression (LR), Random Forest (RF), Support Vector Machine (SVM), XGBoost (XGB), and AdaBoost (ADB). We used the default hyperparameters for all machine learning models as provided by their respective libraries, including the default number of trees for Random Forest and the default kernel function for SVM. No systematic hyperparameter tuning (e.g., grid search or random search) was performed in the current implementation. To ensure robust model evaluation, LPOCV approach with $p = 6$ (10% of total samples) was implemented, resulting in 10 iterations. At each iteration, 54 samples were used for training, and 6 samples were held out for valida-

tion. For the inner training, the 54 training samples were split into 80% for training and 20% for testing using stratified k-fold cross-validation. Important features were selected during model training and testing, with final model performance evaluated using the validation sets. Table 1 illustrates the performance metrics of the five classifiers on training and testing sets, including accuracy, sensitivity, specificity, F1 score, and AUC, while Table 2 displays the evaluation metrics of the best-performing model on the validation sets. They are computed using standard classification metrics [13] derived from the confusion matrix, which are described below.

	Predicted Positive	Predicted Negative
Actual Positive	TP	FN
Actual Negative	FP	TN

Where:

- TP = True Positives (real positives predicted as positives);
- FN = False Negatives (real positives incorrectly predicted as negatives);
- FP = False Positives (real negatives incorrectly predicted as positives);
- TN = True Negatives (real negatives correctly predicted as negatives).

Performance Metric Formulas

1. Accuracy

$$Accuracy = \frac{(TP + TN)}{(TP + TN + FP + FN)}$$

2. Sensitivity (Recall or True positive rate)

$$Sensitivity = \frac{TP}{(TP + FP)}$$

3. Specificity

$$Specificity = \frac{TN}{(TN + FP)}$$

4. F1 score

$$F1\ score = \frac{2 * (Precision * Recall)}{(Precision + Recall)}$$

where $Precision = \frac{TP}{(TP + FP)}$.

5. Area Under the Curve (AUC)

Calculated from the ROC curve plotting True Positive Rate (Sensitivity) against False Positive Rate (FPR), where

$$FPR = \frac{FP}{(FP + TN)}$$

Accuracy

As shown in Table 1, Logistic Regression achieved the highest mean accuracy (0.98 ± 0.05 , CI: 0.95–1.00), and consistently performed well across all iterations, maintaining values between 0.95 and 1.00. SVM also achieved a mean accuracy of 0.98 ± 0.05 (CI: 0.95–1.00), while Random Forest followed closely with 0.97 ± 0.07 (CI: 0.92–1.00). XGBoost (0.93 ± 0.15) and AdaBoost (0.95 ± 0.15) showed greater variability, with AdaBoost

displaying the most instability, sometimes dropping to accuracies as low as 0.84. Despite this, AdaBoost occasionally achieved accuracy as high as 1.0.

Sensitivity

As shown in Table 1, Logistic Regression and Random Forest maintained perfect sensitivity (1.00 ± 0.00) across all iterations, clearly outperforming the other models. SVM showed strong but slightly lower sensitivity (0.95 ± 0.00). XGBoost and AdaBoost were the least stable, with sensitivity ranging from 0.67 to 1.0, and a large standard deviation of ± 0.30 , indicating inconsistency across iterations.

Specificity

Regarding specificity, presented in Table 1, AdaBoost achieved perfect specificity (1.00 ± 0.00), followed by XGBoost (0.97 ± 0.10), Logistic Regression and SVM (0.97 ± 0.10), and Random Forest (0.93 ± 0.13). However, AdaBoost displayed the highest variance, with performance fluctuating significantly. Logistic Regression, in contrast, maintained a consistent specificity with lower variance, reinforcing its reliability.

F1 Score

As shown in Table 1, Logistic Regression exhibited the most consistent performance with F1 scores of 0.99 ± 0.04 (CI: 0.95–1.00), reflecting robust model behavior. SVM and Random Forest also performed well (both at 0.97–0.99) with minor fluctuations. XGBoost had a lower mean F1 score (0.89 ± 0.30), and AdaBoost showed the most variability (0.90 ± 0.30), occasionally dipping below 0.90. Among the classifiers, Logistic Regression was the most stable, making it a potentially more reliable choice for miRNA-based classification tasks.

AUC

As shown in Table 1, AUC values for Logistic Regression remained nearly perfect (0.99 ± 0.10 , CI: 0.99–1.00) across iterations, indicating high and consistent discriminatory power. SVM and Random Forest also showed strong AUC values (0.99 ± 0.03). XGBoost and AdaBoost exhibited more variability, though both reached a maximum AUC of 1.00 in certain iterations.

Overall, Logistic Regression demonstrated the most consistent and robust performance across accuracy, sensitivity, specificity, F1 score, and AUC. In contrast, AdaBoost exhibited the most unstable results, often lagging behind other models in terms of accuracy, sensitivity, and specificity. These results suggest that Logistic Regression may be the most suitable model for miRNA-based Usher syndrome classification, offering reliable performance with minimal fluctuations across different iterations of the LPOCV process.

Table 1. Model training results including mean \pm std with 95% confidence interval).

Model	Accuracy	Sensitivity	Specificity	F1 Score	AUC
Logistic Regression	0.98 ± 0.05 (0.95, 1.00)	1.00 ± 0.00 (1.00, 1.00)	0.97 ± 0.10 (0.89, 1.00)	0.99 ± 0.04 (0.95, 1.00)	0.99 ± 0.10 (0.99, 1.00)
Random Forest	0.97 ± 0.07 (0.92, 1.00)	1.00 ± 0.00 (1.00, 1.00)	0.93 ± 0.13 (0.83, 1.00)	0.97 ± 0.06 (0.93, 1.00)	0.95 ± 0.03 (0.90, 1.00)
SVM	0.98 ± 0.05 (0.94, 1.00)	0.95 ± 0.00 (0.90, 1.00)	0.97 ± 0.10 (0.89, 1.00)	0.99 ± 0.04 (0.95, 1.00)	0.99 ± 0.03 (0.93, 1.00)
XGBoost	0.93 ± 0.15 (0.82, 1.00)	0.90 ± 0.30 (0.67, 1.00)	0.97 ± 0.10 (0.96, 1.00)	0.89 ± 0.30 (0.66, 1.00)	0.99 ± 0.03 (0.96, 1.00)
AdaBoost	0.95 ± 0.15 (0.84, 1.00)	0.90 ± 0.30 (0.67, 1.00)	0.91 ± 0.00 (0.80, 1.00)	0.90 ± 0.30 (0.91, 0.96)	0.94 ± 0.13 (0.90, 1.00))

Table 2. Model validation results including mean \pm std with 95% confidence interval).

Model	Accuracy	Sensitivity	Specificity	F1 Score	AUC
Logistic Regression	0.97 \pm 0.08 (0.93, 1.00)	0.98 \pm 0.07 (0.91, 1.00)	0.93 \pm 0.16 (0.91, 1.00)	0.95 \pm 0.10 (0.91, 1.00)	0.97 \pm 0.08 (0.92, 1.00)
Random Forest	0.60 \pm 0.30 (0.38, 0.81)	0.30 \pm 0.42 (0.00, 0.60)	1.00 \pm 0.00 (1.00, 1.00)	0.93 \pm 0.12 (0.84, 1.00)	0.60 \pm 0.30 (0.38, 0.81)
SVM	0.90 \pm 0.21 (0.82, 1.00)	0.93 \pm 0.12 (0.86, 1.00)	0.97 \pm 0.07 (0.91, 1.00)	0.98 \pm 0.06 (0.93, 1.00)	0.90 \pm 0.21 (0.86, 1.00)
XGBoost	0.50 \pm 0.19 (0.36, 0.63)	0.12 \pm 0.31 (0.10, 0.35)	0.90 \pm 0.31 (0.67, 1.00)	0.71 \pm 0.28 (0.50, 0.91)	0.50 \pm 0.19 (0.36, 0.63)
AdaBoost	0.48 \pm 0.14 (0.37, 0.58)	0.34 \pm 0.20 (0.00, 0.36)	1.00 \pm 0.00 (1.00, 1.00)	0.52 \pm 0.07 (0.46, 0.58)	0.48 \pm 0.14 (0.37, 0.58)

4.3. Model Evaluation

The performance of five models in classifying Usher syndrome samples using miRNA data was evaluated across 10 validation sets over 10 iterations, as shown in Table 2. Logistic regression model demonstrated consistent accuracy with minimal fluctuations, highlighting its ability to generalize effectively in classifying Usher syndrome samples compared to control samples. Specificity remained close to 1.0 across most iterations, indicating the model's reliability in identifying true positives. Sensitivity, although exhibiting slight variability, was stable enough to confirm the model's effectiveness in correctly classifying control samples. The F1 Score, a balance between precision and recall, consistently hovered around 1.00 with minimal variation, reflecting the robust behavior of the model in maintaining a strong balance between true positive rate and precision. Additionally, the Area Under the Curve (AUC) remained nearly perfect, consistently achieving values between 0.97 and 1.0. The overall performance metrics are as follows: mean Accuracy of 97%, mean Sensitivity of 98.0%, mean Specificity of 93%, mean F1 Score of 95%, and mean AUC of 97%. Furthermore, Logistic Regression demonstrated stable and reliable performance across all evaluation metrics, making it a suitable and robust choice for miRNA-based classification of Usher syndrome.

4.4. Pathway Analysis

There were 6115 unique genes predicted to be targeted by the 10 miRNAs derived from the ML model. In total, 572 of these genes were predicted to be strongly suppressed by the miRNAs of interest. There were six GOs flagged to be affected (Figure 5). Of the six GOs, four were associated with neuronal or axon development. The remaining two GOs were signal transduction related, with RAS protein signal transduction, and GTPase mediated signal transduction both predicted to be suppressed by miRNAs of interest. Included are the top 10 pathways associated with the genes targeted by miRNAs from the ML model (Figure 6). The top two pathways in terms of combined scores are both related to neuronal development, corroborating observations from GO enrichment findings (Table 3).

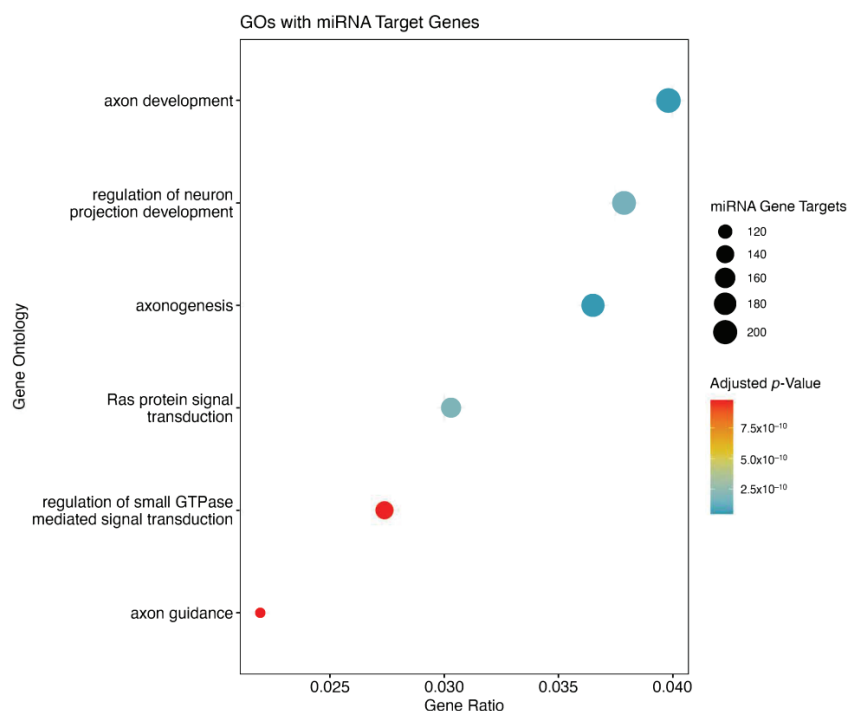


Figure 5. Panel A depicts gene ontologies (GOs) which contain genes targeted by miRNAs from the model. GOs were considered significant at a BH adjusted p -value ≤ 0.05 . The size of the dot in the dot plot corresponds with the number of genes in the GO targeted by miRNAs from the model, while the color corresponds to adjusted p -values.

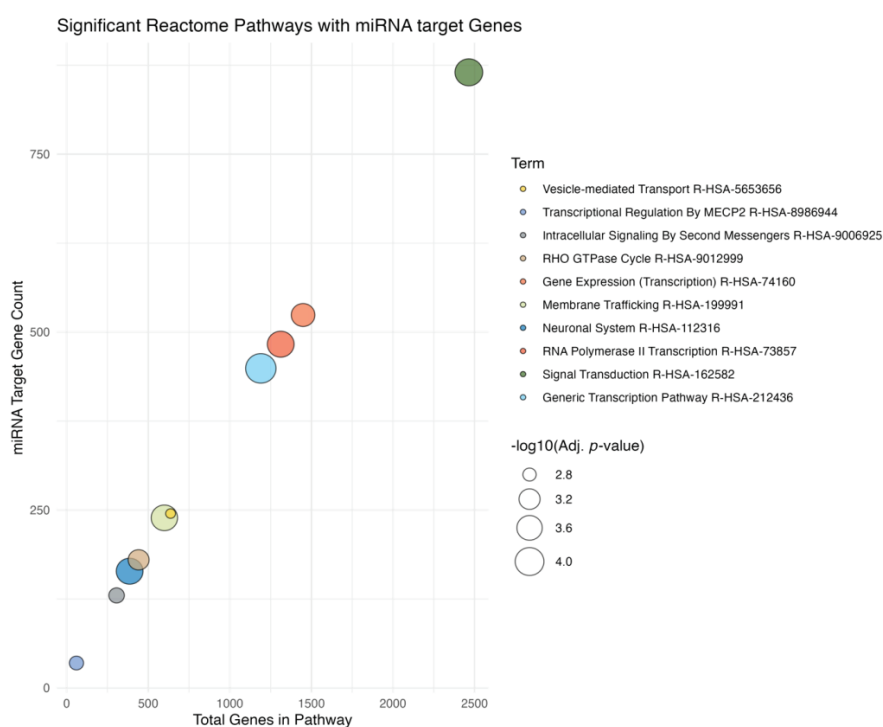


Figure 6. EnrichR analysis of pathways with gene targets corresponding to the miRNAs from the model, using Reactome pathways database. Y-axis is the number of miRNA associated genes, and the x-axis is the total number of genes in the pathway. Dots in the plot represent the different Reactome pathways, which may be influenced by miRNAs from the model. Pathways are colored according to Reactome ID, and the size of the dot reflects the $-\log_{10}(p\text{-value})$.

Table 3. Top 10 pathways predicted to be influenced by miRNAs of interest. Gene-count is the number of genes in the pathway targeted by miRNAs from ML model. Genes in pathway are the total number of genes in the term pathway followed by the percent of affected genes in the pathway. P-values and adjusted p-values are provided from enrichR results, as well as the odds ratios and combined score outputs from enrichR. Significance was determined at an adjusted p -value < 0.05 , and a combined score > 15 .

Term	Gene-Count	Genes in Pathway	Percent of Path	p -Value	Adj. p -Value	Odds-Ratio	Combined-Score
Transcriptional Regulation By MECP2 R-HSA-8986944	35	60	58.33	7.56×10^{-6}	0.0015	3.19	37.64
Neuronal System R-HSA-112316	164	386	42.49	3.91×10^{-7}	0.0002	1.70	25.03
Generic Transcription Pathway R-HSA-212436	449	1190	37.73	3.28×10^{-8}	0.0001	1.41	24.23
Membrane Trafficking R-HSA-199991	239	599	39.90	5.34×10^{-7}	0.0002	1.53	22.07
RHO GTPase Cycle R-HSA-9012999	180	441	40.82	2.55×10^{-6}	0.0007	1.58	20.39
Intracellular Signaling By Second Messengers R-HSA-9006925	130	306	42.48	6.03×10^{-6}	0.0013	1.69	20.34
RNA Polymerase II Transcription R-HSA-73857	483	1312	36.81	3.31×10^{-7}	0.0002	1.35	20.16
Signal Transduction R-HSA-162582	865	2465	35.09	1.47×10^{-7}	0.0001	1.27	19.91
Gene Expression (Transcription) R-HSA-74160	524	1449	36.16	1.27×10^{-6}	0.0004	1.31	17.83
Vesicle-mediated Transport R-HSA-5653656	245	637	38.46	9.82×10^{-6}	0.0018	1.44	16.57

5. Discussion

The proposed methodology for biomarker discovery in Usher syndrome integrates ensemble feature selection with nested cross-validation to identify a minimal set of miRNA biomarkers. This minimal biomarker set represents the smallest subset of miRNAs that can reliably distinguish Usher syndrome from control samples. However, due to the rarity of Usher syndrome, obtaining large sample sizes remains a challenge, which may impact the generalizability of the selected biomarkers. As more data become available, this approach is designed to refine and update the biomarker set, improving its robustness and biological relevance over time. Future validation on larger datasets will enhance its clinical applicability and reliability.

By employing nested cross-validation, this methodology ensures that dataset partitioning into training and validation sets is performed iteratively. The training sets undergo additional stratified cross-validation, where multiple validation sets are tested to generate stable performance metrics. Compared to single-set validation, this process provides more reliable feature selection and performance evaluation. The combination of ensemble feature selection and rigorous cross-validation enhances the stability of identified biomarkers, ensuring their reproducibility, even with limited data.

Future studies incorporating larger sample sizes will be essential to further validate and refine the identified miRNA biomarkers, advancing their potential use in early diagnosis and therapeutic targeting of Usher syndrome. The methodology leverages LPOCV for the outer loop and stratified k-fold cross-validation for the inner loop, ensuring robust biomarker selection.

6. Limitations and Future Directions

While our study demonstrates promising results for the automated detection of Usher syndrome using miRNA expression profiles, there are several limitations. First, the dataset used is small, reflecting the challenges inherent to studying rare genetic disorders like Usher syndrome. A small sample size may limit the robustness and generalizability of the identified minimal feature set, which may not fully capture the variability across a more diverse population. Additionally, the heterogeneity of miRNA expression influenced by various factors (e.g., age, genetic background, and environmental factors) could impact the model's performance when applied to different cohorts. Another limitation lies in miRNA data itself; variations in miRNA extraction and quantification techniques may influence the reproducibility of results across different labs or clinical settings.

Future research should focus on expanding the dataset by including samples from diverse populations and incorporating longitudinal data where possible. This would allow the feature selection method to capture a broader range of genetic variability, enhancing the generalizability of the model. In addition, further studies should explore integrating multi-omics data (e.g., mRNA, protein, and epigenetic data) to improve predictive accuracy and capture complex biological interactions related to Usher syndrome. Another potential avenue is the incorporation of transfer learning or domain adaptation techniques to enable the model trained on miRNA data to be effectively applied to new data sources or patient groups. Validation in a clinical setting is also necessary to assess the practicality and reliability of the proposed framework in real-world diagnostic workflows. Finally, as more samples become available, it would be valuable to explore deep learning-based models that could potentially improve classification performance by capturing non-linear relationships in the data.

7. Conclusions

This study presents a machine learning-based approach incorporating ensemble feature selection and nested cross-validation for the discovery of miRNA biomarkers associated with Usher syndrome. Given the rarity of Usher syndrome, obtaining large datasets is challenging. Our method aims to maximize the reliability of biomarker identification by selecting a minimal set of miRNAs that remain robust against variations in sample size. The ensemble feature selection component integrates results from multiple models, while nested cross-validation ensures rigorous evaluation. This provides more reliable biomarker selection compared to conventional validation methods. The identified minimal biomarker set represents a promising step toward developing miRNA-based biomarkers for Usher syndrome, although its generalizability to larger, more diverse datasets remains a future goal. Our results demonstrate that as more data become available, the biomarker set can be refined and updated, enhancing its clinical applicability. Overall, this study provides a robust framework for feature selection and validation in small, complex datasets, with potential applications beyond Usher syndrome to other rare genetic disorders where data availability is limited.

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

Funding: This research was funded by a grant from the Ryan foundation, 3100 E Willamette Lane, Greenwood Village, CO 80121 to MRF.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The datasets generated and analyzed during the current study are available from the corresponding author upon reasonable request. Additionally, the code used for analysis is also available upon request.

Acknowledgments: We wish to thank Jennifer Bushing, Genomics Core Facility, University of Nebraska Medical Center, Omaha, NE, USA, for her technical assistance in NanoString miRNA microarray experiments. The UNMC Genomics Core Facility receives partial support from the National Institute for General Medical Science (NIGMS) INBRE-P20GM103427-19, as well as the National Cancer Institute and The Fred & Pamela Buffett Cancer Center Support Grant-P30CA036727. This publication's contents are the sole responsibility of the authors and do not necessarily represent the official views of the NIH or NIGMS.

Conflicts of Interest: All authors declare no conflict of interest.

References

- Spandau, U.H.; Rohrschneider, K. Prevalence and geographical distribution of Usher syndrome in Germany. *Graefes Arch. Clin. Exp. Ophthalmol.* **2002**, *240*, 495–498. [CrossRef] [PubMed]
- Castiglione, A.; Möller, C. Usher syndrome. *Audiol. Res.* **2022**, *12*, 42–65. [CrossRef]
- Boughman, J.A.; Vernon, M.; Shaver, K.A. Usher syndrome: Definition and estimate of prevalence from two high-risk populations. *J. Chronic Dis.* **1983**, *36*, 595–603. [CrossRef]
- Toms, M.; Pagarkar, W.; Moosajee, M. Usher syndrome: Clinical features, molecular genetics and advancing therapeutics. *Ther. Adv. Ophthalmol.* **2020**, *12*, 2515841420952194. [CrossRef]
- Rijavec, N.; Novak Grubic, V. Usher syndrome and psychiatric symptoms: A challenge in psychiatric management. *Psychiatr. Danub.* **2009**, *21*, 68–71. [PubMed]
- Tsilou, E.T.; Rubin, B.I.; Caruso, R.C.; Reed, G.F.; Pikus, A.; Hejtmancik, J.F.; Iwata, F.; Redman, J.B.; Kaiser-Kupfer, M.I. Usher syndrome clinical types I and II: Could ocular symptoms and signs differentiate between the two types? *Acta Ophthalmol. Scand.* **2002**, *80*, 196–201. [CrossRef]
- Tom, W.A.; Chandel, D.S.; Jiang, C.; Krzyzanowski, G.; Fernando, N.; Olou, A.; Fernando, M.R. Genotype Characterization and MiRNA Expression Profiling in Usher Syndrome Cell Lines. *Int. J. Mol. Sci.* **2024**, *25*, 9993. [CrossRef] [PubMed]
- Wang, S.; Xu, C.Y.; Zhu, Y.; Ding, W.; Hu, J.; Xu, B.; Guo, Y.; Liu, X. A rare transcript homozygous variants in CLRN1 (USH3A) causes Usher syndrome type 3 in a Chinese family. *Orphanet J. Rare Dis.* **2024**, *19*, 349. [CrossRef]
- Lagana, A.; Forte, S.; Giudice, A.; Arena, M.R.; Puglisi, P.L.; Giugno, R.; Pulvirenti, A.; Shasha, D.; Ferro, A. miRo: A miRNA knowledge base. *Database* **2009**, *2009*, bap008. [CrossRef]
- Andersen, G.B.; Tost, J. Circulating miRNAs as biomarker in cancer. In *Tumor Liquid Biopsies*; Springer: Cham, Switzerland, 2020; pp. 277–298.
- Thelagathoti, R.K.; Tom, W.A.; Jiang, C.; Chandel, D.S.; Krzyzanowski, G.; Olou, A.; Fernando, R.M. A Network Analysis Approach to Detect and Differentiate Usher Syndrome Types Using miRNA Expression Profiles: A Pilot Study. *BioMedInformatics* **2024**, *4*, 2271–2286. [CrossRef]
- Lopez-Rincon, A.; Martinez-Archundia, M.; Martinez-Ruiz, G.U.; Schoenhuth, A.; Tonda, A. Automatic discovery of 100-miRNA signature for cancer classification using ensemble feature selection. *BMC Bioinform.* **2019**, *20*, 480. [CrossRef] [PubMed]
- Torres, R.; Judson-Torres, R.L. Research techniques made simple: Feature selection for biomarker discovery. *J. Investig. Dermatol.* **2019**, *139*, 2068–2074. [CrossRef] [PubMed]
- Khalifa, W.; Yousef, M.; Saçar Demirci, M.D.; Allmer, J. The impact of feature selection on one and two-class classification performance for plant microRNAs. *PeerJ* **2016**, *4*, e2135. [CrossRef]
- Sarkar, J.P.; Saha, I.; Sarkar, A.; Maulik, U. Machine learning integrated ensemble of feature selection methods followed by survival analysis for predicting breast cancer subtype-specific miRNA biomarkers. *Comput. Biol. Med.* **2021**, *131*, 104244. [CrossRef]
- Saeys, Y.; Abeel, T.; Van de Peer, Y. Robust feature selection using ensemble feature selection techniques. In *Machine Learning and Knowledge Discovery in Databases, Proceedings of the European Conference, ECML PKDD 2008, Antwerp, Belgium, 15–19 September 2008, Proceedings, Part II*; Springer: Berlin/Heidelberg, Germany, 2008; pp. 313–325.
- Seijo-Pardo, B.; Porto-Díaz, I.; Bolón-Canedo, V.; Alonso-Betanzos, A. Ensemble feature selection: Homogeneous and heterogeneous approaches. *Knowl.-Based Syst.* **2017**, *118*, 124–139. [CrossRef]

18. Si, H.; Sun, X.; Chen, Y.; Cao, Y.; Chen, S.; Wang, H.; Hu, C. Circulating microRNA-92a and microRNA-21 as novel minimally invasive biomarkers for primary breast cancer. *J. Cancer Res. Clin. Oncol.* **2013**, *139*, 223–229. [CrossRef] [PubMed]
19. Jović, A.; Brkić, K.; Bogunović, N. A review of feature selection methods with applications. In Proceedings of the 2015 38th International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO), Opatija, Croatia, 25–29 May 2015; pp. 1200–1205.
20. Cai, Z.; Xu, D.; Zhang, Q.; Zhang, J.; Ngai, S.M.; Shao, J. Classification of lung cancer using ensemble-based feature selection and machine learning methods. *Mol. Biosyst.* **2015**, *11*, 791–800. [CrossRef]
21. Lopez-Rincon, A.; Mendoza-Maldonado, L.; Martinez-Archundia, M.; Schönhuth, A.; Kraneveld, A.D.; Garssen, J.; Tonda, A. Machine learning-based ensemble recursive feature selection of circulating miRNAs for cancer tumor classification. *Cancers* **2020**, *12*, 1785. [CrossRef]
22. Colombelli, F.; Kowalski, T.W.; Recamonde-Mendoza, M. A hybrid ensemble feature selection design for candidate biomarkers discovery from transcriptome profiles. *Knowl.-Based Syst.* **2022**, *254*, 109655. [CrossRef]
23. Canouil, M.; Bouland, G.A.; Bonnefond, A.; Froguel, P.; 't Hart, L.M.; Sliker, R.C. NACHO: An R package for quality control of NanoString nCounter data. *Bioinformatics* **2020**, *36*, 970–971. [CrossRef]
24. Kotlarchyk, A.; Khoshgoftaar, T.; Pavlovic, M.; Zhuang, H.; Pandya, A.S. Identification of microRNA biomarkers for cancer by combining multiple feature selection techniques. *J. Comput. Methods Sci. Eng.* **2011**, *11*, 283–298. [CrossRef]
25. Bashir, S.; Khan, Z.S.; Khan, F.H.; Anjum, A.; Bashir, K. Improving heart disease prediction using feature selection approaches. In Proceedings of the 2019 16th International Bhurban Conference on Applied Sciences and Technology (IBCAST), Islamabad, Pakistan, 8–12 January 2019; IEEE: Piscataway, NJ, USA, 2019; pp. 619–623.
26. Fu, K.S.; Min, P.J.; Li, T.J. Feature selection in pattern recognition. *IEEE Trans. Syst. Sci. Cybern.* **1970**, *6*, 33–39. [CrossRef]
27. Van Hulse, J.; Khoshgoftaar, T.M.; Napolitano, A.; Wald, R. Threshold-based feature selection techniques for high-dimensional bioinformatics data. *Netw. Model. Anal. Health Inform. Bioinform.* **2012**, *1*, 47–61. [CrossRef]
28. Guyon, I.; Elisseeff, A. An introduction to variable and feature selection. *J. Mach. Learn. Res.* **2003**, *3*, 1157–1182.
29. Kohavi, R.; John, G.H. Wrappers for feature subset selection. *Artif. Intell.* **1997**, *97*, 273–324. [CrossRef]
30. Hastie, T.; Tibshirani, R.; Friedman, J.H.; Friedman, J.H. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*, 2nd ed.; Springer: New York, NY, USA, 2009.
31. Bolón-Canedo, V.; Alonso-Betanzos, A. Ensembles for feature selection: A review and future trends. *Inf. Fusion.* **2019**, *52*, 1–12. [CrossRef]
32. Jeon, H.; Oh, S. Hybrid-recursive feature elimination for efficient feature selection. *Appl. Sci.* **2020**, *10*, 3211. [CrossRef]
33. Liu, W.; Wang, J. Recursive elimination–election algorithms for wrapper feature selection. *Appl. Soft Comput.* **2021**, *113*, 107956. [CrossRef]
34. Nguyen, C.; Wang, Y.; Nguyen, H.N. Random forest classifier combined with feature selection for breast cancer diagnosis and prognostic. *J. Biomed. Sci. Eng.* **2013**, *6*, 551–560. [CrossRef]
35. Wang, H.; Yang, F.; Luo, Z. An experimental study of the intrinsic stability of random forest variable importance measures. *BMC Bioinform.* **2016**, *17*, 60. [CrossRef]
36. Strobl, C.; Boulesteix, A.L.; Zeileis, A.; Hothorn, T. Bias in random forest variable importance measures: Illustrations, sources and a solution. *BMC Bioinform.* **2007**, *8*, 25. [CrossRef] [PubMed]
37. Tibshirani, R. Regression shrinkage and selection via the lasso. *J. R. Stat. Soc. B Stat. Methodol.* **1996**, *58*, 267–288. [CrossRef]
38. Choudhary, R.; Gianey, H.K. Comprehensive review on supervised machine learning algorithms. In Proceedings of the 2017 International Conference on Machine Learning and Data Science (MLDS), Noida, India, 14–15 December 2017; IEEE: Piscataway, NJ, USA, 2017; pp. 37–43.
39. Park, H.A. An introduction to logistic regression: From basic concepts to interpretation with particular attention to nursing domain. *J. Korean Acad. Nurs.* **2013**, *43*, 154–164. [CrossRef]
40. Parmar, A.; Katariya, R.; Patel, V. A review on random forest: An ensemble classifier. In *International Conference on Intelligent Data Communication Technologies and Internet of Things (ICICI) 2018*; Springer International Publishing: Cham, Switzerland, 2019; pp. 758–763.
41. Lewis, M.J.; Spiliopoulou, A.; Goldmann, K.; Pitzalis, C.; McKeigue, P.; Barnes, M.R. nestedcv: An R package for fast implementation of nested cross-validation with embedded feature selection designed for transcriptomics and high-dimensional data. *Bioinform. Adv.* **2023**, *3*, vbad048. [CrossRef] [PubMed]
42. Pian, C.; Mao, S.; Zhang, G.; Du, J.; Li, F.; Leung, S.Y.; Fan, X. Discovering cancer-related miRNAs from miRNA-target interactions by support vector machines. *Mol. Ther. Nucleic Acids.* **2020**, *19*, 1423–1433. [CrossRef]
43. Zhang, X.; Li, T.; Wang, J.; Li, J.; Chen, L.; Liu, C. Identification of cancer-related long non-coding RNAs using XGBoost with high accuracy. *Front. Genet.* **2019**, *10*, 735. [CrossRef]
44. Liu, D.; Huang, Y.; Nie, W.; Zhang, J.; Deng, L. SMALF: miRNA-disease associations prediction based on stacked autoencoder and XGBoost. *BMC Bioinform.* **2021**, *22*, 219. [CrossRef] [PubMed]

45. Crisci, C.; Ghattas, B.; Perera, G. A review of supervised machine learning algorithms and their applications to ecological data. *Ecol. Model.* **2012**, *240*, 113–122. [CrossRef]
46. Guan, D.G.; Liao, J.Y.; Qu, Z.H.; Zhang, Y.; Qu, L.H. mirExplorer: Detecting microRNAs from genome and next generation sequencing data using the AdaBoost method with transition probability matrix and combined features. *RNA Biol.* **2011**, *8*, 922–934. [CrossRef]
47. Zhong, Y.; He, J.; Chalise, P. Nested and repeated cross validation for classification model with high-dimensional data. *Rev. Colomb. Estad.* **2020**, *43*, 103–125. [CrossRef]
48. Browne, M.W. Cross-validation methods. *J. Math. Psychol.* **2000**, *44*, 108–132. [CrossRef] [PubMed]
49. Zhong, Y.; Chalise, P.; He, J. Nested cross-validation with ensemble feature selection and classification model for high-dimensional biological data. *Commun. Stat. Simul. Comput.* **2023**, *52*, 110–125. [CrossRef]
50. McGeary, S.E.; Lin, K.S.; Shi, C.Y.; Pham, T.M.; Bisaria, N.; Kelley, G.M.; Bartel, D.P. The biochemical basis of microRNA targeting efficacy. *Science* **2019**, *366*, eaav1741. [CrossRef]
51. Wu, T.; Hu, E.; Xu, S.; Chen, M.; Guo, P.; Dai, Z.; Feng, T.; Zhou, L.; Tang, W.; Zhan, L.; et al. clusterProfiler 4.0, A universal enrichment tool for interpreting omics data. *Innovation* **2021**, *2*, 100141. [CrossRef] [PubMed]
52. Jassal, B.; Matthews, L.; Viteri, G.; Gong, C.; Lorente, P.; Fabregat, A.; Sidiropoulos, K.; Cook, J.; Gillespie, M.; Haw, R.; et al. The reactome pathway knowledgebase. *Nucleic Acids Res.* **2020**, *48*, D498–D503. [CrossRef]
53. Kuleshov, M.V.; Jones, M.R.; Rouillard, A.D.; Fernandez, N.F.; Duan, Q.; Wang, Z.; Ma'ayan, A. Enrichr: A comprehensive gene set enrichment analysis web server 2016 update. *Nucleic Acids Res.* **2016**, *44*, W90–W97. [CrossRef]

Disclaimer/Publisher’s Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.

MDPI AG
Grosspeteranlage 5
4052 Basel
Switzerland
Tel.: +41 61 683 77 34

Bioengineering Editorial Office
E-mail: bioengineering@mdpi.com
www.mdpi.com/journal/bioengineering



Disclaimer/Publisher's Note: The title and front matter of this reprint are at the discretion of the Guest Editor. The publisher is not responsible for their content or any associated concerns. The statements, opinions and data contained in all individual articles are solely those of the individual Editor and contributors and not of MDPI. MDPI disclaims responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.



Academic Open
Access Publishing

mdpi.com

ISBN 978-3-7258-4564-4