

Review

# A Review of Fuzzy and Pattern-Based Approaches for Class Imbalance Problems

Ismael Lin <sup>1</sup>, Octavio Loyola-González <sup>2,\*</sup>, Raúl Monroy <sup>1</sup> and Miguel Angel Medina-Pérez <sup>1</sup>

<sup>1</sup> School of Engineering and Science, Tecnológico de Monterrey, Carretera al Lago de Guadalupe Km. 3.5, Atizapán 52926, Estado de México, Mexico; A01210802@itesm.mx (I.L.); raulm@tec.mx (R.M.); miguel@tec.mx (M.A.M.-P.)

<sup>2</sup> Altair Management Consultants Corp., 303 Wyman St., Suite 300, Waltham, MA 02451, USA

\* Correspondence: olg@altair.consulting

**Abstract:** The usage of imbalanced databases is a recurrent problem in real-world data such as medical diagnostic, fraud detection, and pattern recognition. Nevertheless, in class imbalance problems, the classifiers are commonly biased by the class with more objects (majority class) and ignore the class with fewer objects (minority class). There are different ways to solve the class imbalance problem, and there has been a trend towards the usage of patterns and fuzzy approaches due to the favorable results. In this paper, we provide an in-depth review of popular methods for imbalanced databases related to patterns and fuzzy approaches. The reviewed papers include classifiers, data preprocessing, and evaluation metrics. We identify different application domains and describe how the methods are used. Finally, we suggest further research directions according to the analysis of the reviewed papers and the trend of the state of the art.

**Keywords:** contrast patterns; data mining; fuzzy set; imbalanced databases



**Citation:** Lin, I.; Loyola-González, O.; Monroy, R.; Medina-Pérez, M.A. A Review of Fuzzy and Pattern-Based Approaches for Class Imbalance Problems. *Appl. Sci.* **2021**, *11*, 6310. <https://doi.org/10.3390/app11146310>

Academic Editor: Federico Divina

Received: 8 June 2021

Accepted: 5 July 2021

Published: 8 July 2021

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



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

## 1. Introduction

In recent years, classification using imbalanced classes has attracted increasing interest because it is a recurrent problem in real-world data, such as financial statement fraud detection [1], bankruptcy prediction [2–4], medical decision making [5], fault diagnosis [6,7], card fraud detection [8,9], pattern recognition [10], cancer gene expression [11,12], and telecommunications fraud [13]. The class imbalance problem is presented when the objects are not distributed equally among classes; usually, the popular classifier is biased to the class that has significantly more objects (a.k.a. majority class) and dismisses the class with significantly fewer objects (a.k.a. minority class) [14,15].

Popular classifiers such as support vector machine (SVM) [16], C4.5 [17], and k nearest neighbour (kNN) [18] were originally created for non-imbalanced databases. If the same techniques are used to handle imbalanced databases, these classifiers could obtain poor classification results for the minority class [19].

Many approaches have been developed in recent years and have addressed the imbalance problem with great interest in the scientific community [15,20–22]. Unfortunately, some results are obtained by black-box models, and there is a lack of understandability in the model itself. For that reason, the international scientific community has a great interest in creating explainable artificial intelligence (XAI) models [23]. Those models need to be accurate and explainable for experts in a specific area [24].

Among these approaches, the usage of patterns and fuzzy logic has shown more accurate results than other classifiers [25,26]. Pattern-based classifiers provide an expression defined in a particular language that describes a set of objects. Nevertheless, these types of patterns are restrictive in their classifications because their features are limited to crisp values [27]. The hard cut of the data can result in an abrupt classification of the objects.

This problem can be solved with fuzzy patterns that have the property of flexibility, and the results are expressed in a language closer to that used by an expert [28,29].

To the best of our knowledge, there is no review of the most recent papers for the class imbalance problem addressed by fuzzy and pattern-based approaches. Hence, in this paper, we provide a systematic review of the state of the art for imbalanced databases that include fuzzy and pattern-based approaches, including theoretical and practical approaches. For the theory, we present common approaches to deal with imbalanced data; these approaches will include data mining, preprocessing, evaluation metrics, and classifiers. For the applications, we gather the most common real-world scenarios where imbalanced data is a significant concern. We cover different research areas so that the reader can have a better understanding of the problem.

The rest of this paper is structured as follows: Section 2 presents the general background. Section 3 describes the research methodology and presents different approaches to deal with class imbalance problems with patterns and fuzzy approaches. Section 4 presents different research areas where class imbalance problems are presented in real-world data. Section 5 presents a taxonomy to group the different approaches. Section 6 explains our thoughts of the current state of the art and presents future directions. Section 7 presents the conclusions of this paper.

## 2. General Background

In this section, we present a general background that explains the class imbalance problem, common ways to deal with imbalanced data, how pattern-based classifiers works, and preliminaries about pattern-based classification and fuzzy logic.

### 2.1. Class Imbalance Problem

As we previously stated, the class imbalance problem is a recurrent issue in real-world data, and it occurs when the objects are not distributed equally among the problem's classes. The objects of the minority class can be described as safe, borderline, rare, and outliers [30], making the classification process even harder for classifiers [31]. In the following list, there is a description for each of them:

- Safe: Data located in the homogeneous regions from one class only (majority or minority).
- Borderline: Data located in nearby decision boundaries between classes. In this scenario, the classifiers need to decide the class of the objects when they are in the decision boundary, which, due to the bias, result in favor of the majority class.
- Rare: Data located inside the majority class is often seen as overlapping. The classifier tends to classify the minority class as part of the majority class. The effect of this has been discussed in different works [32].
- Outliers: Data located far away from the sample space. The minority objects could be treated as noise by the classifier; on the other hand, noise could be treated as minority objects [33]. This happens when there are outlier objects in the database and the data should not be removed because it could be a representation of a minority class.

The presence of small disjuncts and the lack of density in the training data also affect classic classifier performance. These characteristics tend to affect outlier data with a greater impact [19].

### 2.2. Approaches to Deal with Class Imbalance Problems

To deal with the class imbalance problem, three approaches are commonly used: data level, algorithm level, and cost-sensitive [15,31,34].

1. Data level: The objective of this approach is to create a balanced training dataset by preprocessing the data through artificial manipulation. There are three solutions to data sampling: over-sampling, under-sampling, and hybrid-sampling.
  - (a) Over-sampling: New objects are generated for the minority class. The basic version of this solution is the random oversampling of the objects belonging

- to the minority class. However, the main drawback is that it can lead to overfitting [35,36].
- (b) Under-sampling: Objects are removed from the majority class. The goal is to have the same number of objects in each class. The basic solution of this method is random under-sampling. The disadvantage of using this method is that it can exclude a significant amount of the original data.
  - (c) Hybrid-sampling: A combination of over-sampling and under-sampling. This approach generates objects for the minority class while it eliminates objects from the majority class.
2. Algorithm level: The aim of this type of approach is a specific modification of the classifier. This approach is not flexible for different classification problems because it focuses on a specific classifier with a specific type of database. Nevertheless, the results could lead to good classification results for a particular problem. This type of solution can also combine strengths of different solutions as the NeuroFuzzy Model [37], which combines a fuzzy system trained as a neural network [31].
  3. Cost-sensitive: The objective is to create a cost matrix that is built with different misclassification costs. The misclassified objects of the minority class have a higher misclassification cost than the misclassified objects of the majority class. One of the main disadvantages is the cost-sensitive problem, which appears because the cost of misclassification is different for each of the classes. Therefore, this type of problem cannot be compared against non-cost-sensitive problems [38].

### 2.3. Pattern-Based Classifiers

Pattern-based classifiers can explain their results, through patterns, in a language closer to that used by experts, and the results have been shown to be more accurate than other popular classifiers, such as decision trees, nearest neighbor, bagging, boosting, SVM, and naive Bayes [25,26].

The stages of building a pattern-based classifier are:

- Mining stage: At this stage, several patterns are extracted from the training dataset. Mining contrast patterns is a challenging problem due to the high computational cost generated by the exponential number of candidate patterns [12,39–41].
- Filtering stage: At this stage, there are set-based filters and quality measures that need to distinguish between patterns that have a high discriminative ability for supervised classification [41]. The quality is usually established by measuring reliability, novelty, coverage, conciseness, peculiarity, diversity, utility, and actionability [42]. All the previous measurements take into account two parameters: if the pattern covers an object and if the object is representative of the class determined by the pattern.
- Classification stage: The last stage is the classification of query objects. At this stage, the classifier combines the patterns and creates a voting scheme. Finally, it is necessary to evaluate the performance of the classifier to determine its quality.

### 2.4. Fuzzy Logic

Fuzzy logic has been used as a term to refer to degrees of something [43], rather than a Boolean value (i.e., 0 or 1). It is helpful to handle imprecision, and it gives more flexibility in the systems [44]. In 1965, the concept of fuzzy logic theory was introduced by Zadeh [45]. The idea came while working on a problem with natural language, which is complicated to translate to binary values because not everything can be described as absolutely true or absolutely false.

The main purpose of this approach was to overcome many of the problems found in classical logic theory [46], especially the limitations presented with crisp values. With a fuzzy approach, the items have a membership grade value that goes from a lower limit to an upper limit for a given class A; typically, it goes in an interval from [0,1]. The lower limit (membership value zero) represents no membership to class A, while the upper limit (membership value one) represents full membership to class A. The other possible

numbers mean that there is a partial degree of membership of the corresponding class. The membership grade is assigned by a membership function that possesses the quality of shape adaptability according to a specific need. Elements with zero degree of membership are usually not listed.

A fuzzy set is defined as:

$$\tilde{A} = \{(o, \mu_{\tilde{A}}(o)) | o \in O\}.$$

where  $\tilde{A}$  is the fuzzy set,  $O$  is a set of objects,  $o$  is the object, and  $\mu_{\tilde{A}}(o)$  is the membership grade of  $o$  in  $\tilde{A}$  [47].

As we mentioned before, fuzzy logic has the primary purpose of overcome the limitations presented with crisp values. This problem is extended to classifiers where the hard cut of the data can result in an abrupt classification of the objects. A common solution is to apply fuzzy logic to different stages of the classifier.

A fuzzy classifier calculates a membership vector  $M = \{m_1, m_2, m_3, \dots, m_n\}$  for an object  $o$ . The vector element  $m_i \in [0, 1]$  is the membership grade value of  $o$  in the class  $C_i$  [48]. The object  $o$  is represented by a vector of  $t$ -dimensional features from a universe of discourse  $U$  and  $C_1, C_2, C_3, \dots, C_n$  is a set of given classes or classes to be discovered.

Since fuzzy set theory is a broader topic, more information can be consulted in [47]. It details definitions, algebraic operations with fuzzy numbers, fuzzy functions, and applications. We also suggest the work of Orazbayev et al. [49] and Nicolas Werro [50], where they explained in detail the process of developing complex systems in real-world environments with a fuzzy approach.

### 3. Pattern and Fuzzy Approaches for Imbalance Problems

In this section, we explain our research methodology, which includes the number of reviewed papers, the range of covered years, and the aim of the analyzed papers. Then, we list the most recent data-level and algorithm-level approaches for imbalance problems that use patterns or fuzzy approaches. Finally, we finish the section with a discussion of the reviewed papers.

For the reviewed papers, we perform an overview of their work that includes a summary of their method, the databases, the evaluation metrics, the compared methods and their results.

#### 3.1. Research Methodology

For this review, we performed an in-depth review of papers that include fuzzy and/or pattern-based approaches for imbalanced databases. The search includes 62 documents from 2005 to 2021. The range was determined by the first work that covers fuzzy sets, pattern-based, and imbalance problems. We also include documents that have cited the work of Garcia-Borroto et al. [20]. This case was included because it was the first published work that combined fuzzy sets and pattern classification; unfortunately, it does not cover the imbalance problem. The results that cited Garcia-Borroto et al., showed 22 documents from 2011 to 2020. Therefore, a total of 84 documents were relevant in the scope of our research.

In recent decades, there have been many algorithms to address the data classification problems. However, in this paper, we reduce the scope to the approaches that include fuzzy sets, pattern-based, imbalance problems, and documents that have cited [20]. This decision was made so as to avoid confusion due to the number of methods or approaches that exist in the classification process.

#### 3.2. Data-Level Approaches

In this subsection, we present the most recent data-level approaches presented for imbalance problems that use patterns or fuzzy approaches.

In 2017, Liu et al. [51] proposed a method to handle class imbalance and missing values called fuzzy-based information decomposition (FID). FID created artificial data

of the minority class. Their scheme is divided into two parts, weighting, and recovery. The weighting process is made by the fuzzy membership functions. The functions are used to measure how much the data contributes to the missing estimation. For the recovery step, they consider the contribution of the data to determine the missing values. They used a C4.5 [52] classifier with an entropy-based splitting criterion. They used 27 databases combined from different repositories (e.g., PROMISE [53], UCI [54], KEEL [55]). The evaluation metrics used for their experiments are G-mean, AUC, and Matthews correlation coefficients (MCC) [56]. Their results outperform the following crisp methods: Mix [57], KNNI [58], SOM [59], SM [35], CBOS [60], CBUS [61], ROS [62], RUS [62], and MWM [60].

In 2018, Liu et al. [63] proposed a fuzzy rule-based oversampling technique (FRO) to solve the class imbalance problem. FRO uses the training data to create fuzzy rules and determine the weight of each rule. The weight of the rule represents how much the sample belongs to a fuzzy space. After the rules are created, they synthesize the artificial data according to those rules. Another contribution is that FRO can handle missing values with the best matched fuzzy rule, considering the correlation and difference between attributes. They used 55 real-world databases from the UCI [54] and KEEL [55] repository. They used AUC, recall, and F1-measure as their evaluation metrics. Their solution is compared against SMOTE [35], MWMOTE [64], RWO [65], RACOG [66], and the fuzzy method FID [51]. Their results showed that their technique is better or equal to the other compared methods and is robust in cases with overlapping.

In 2020, Pawel Ksieniewicz [67] proposed the construction of the support-domain decision boundary implementing the fuzzy templates method. Their goal was to avoid artificial data created with oversampling techniques. Their approach includes two solutions, one based on fuzzy templates, and the second one includes the standard deviation. They compare the results obtained by RDB, ROS-RDB, FTDB [68], and SDB. In RDB and ROS-RDB, the decision boundary is defined by a line equation. For FDTB, the decision boundary includes the mean support of each class. Finally, for SDB, they included the standard deviation. The databases used for the experiments are from the KEEL [55] repository and have an imbalance degree greater than 1:9 ratio with up to 20 features. They used F1-score, precision, recall, accuracy, and geometric mean to evaluate the performance of their solution. Their results claimed to improve the process of classification against oversampling methods by avoiding an overweight in the minority class.

In 2020, Ren et al. [69] proposed a fuzzy representativeness difference-based oversampling technique called FRDOAC. Their method uses affinity propagation and chromosome theory of inheritance. First, they use an oversampling method, fuzzy representativeness difference (FRD), which finds a representative sample of each class according to the importance of the samples. Then, they used Mahalanobis distance [70] to calculate the representativeness of every sample. In the last step, they created artificial data based on the chromosome theory of inheritance. The comparison was made against other crisp oversampling methods such as: SMOTE [35], MWMOTE [64], RWO [65], RAGOC, and GIR [71]. They used 16 databases from the UCI [54] and KEEL [55] repository. For the evaluation metrics, they used F-measure, G-mean, and AUC. Finally, the results showed better performance than the other methods.

In 2020, Kaur et al. [72] proposed a hybrid data level that overcomes some difficulties presented in different types of data impurities. They define data impurities as data that present noise, different data distributions, or class imbalance. Their approach has three different phases; in the first phase, they clean the data from noise and applied a radial basis kernel for clustering minority and majority classes. In the second phase, the clusters are artificially balanced; they used radial basis kernel fuzzy membership to reduce the majority class and firefly-based SMOTE [73] to increase the minority class. In the last step, they use a DT [52] as their classifier. They tested their method in three synthetic databases and 44 UCI [54] real-world databases. They used AUC to evaluate the performance of the method. They compared their results against ASS [74], ADOMS [74], AHC [75], B\_SMOTE [76], ROS [77], SL\_SMOTE [78], SMOTE [35], SMOTE\_ENN [77], SMOTE\_TL [77], SPIDER [79],

SPIDER2, CNN [80], CNNTL [81], CPM [82], NCL [83], OSS [84], RUS [85], SBC [86], TL [81], and SMOTE\_RSB, which has a fuzzy approach [87]. Finally, their results outperform the other methods especially in databases with a high imbalanced data ratio.

### 3.3. Algorithm Level

In this subsection, we present the most recent algorithm-level approaches presented for imbalance problems that use patterns and fuzzy logic.

In 2011, Garcia-Borroto et al. [20] proposed a fuzzy emerging pattern technique that overcomes the problem presented with continuous attributes using fuzzy discretization. To extract the fuzzy patterns, they used several fuzzy decision trees. For the induction procedure, they implemented linguistic hedges that solved the problem of the initial fuzzy discretization of continuous attributes. They created a way to test the discriminability of a fuzzy pattern named Trust. Finally, they proposed a new classifier (FEPC) that uses a graph-based strategy for pattern organization. The results outperform popular crisp classifiers such as bagging [88], boosting [88], c4.5 [52], random forest [89], and SJEP [90]; however, the results were similar compared to kNN and SVM, which were tested on 16 databases from the UCI repository [54].

In 2014, Buscema et al. [91] proposed K-Contractive Map (K-CM), which is a neural network that uses the variable connection weights to obtain the z-transforms. Then, they combined k-NN classifiers, which are used for class evaluation, and the connection weights are given by the Learning of the Auto-Contractive Map (Auto-CM) strategy [92]. The authors mentioned how K-CM helps bottom-up algorithms to provide a symbolic explanation of their learning. They also claimed that K-CM is not prone to over-fitting because it works with a complex minimization of the energy function with parallel constraints that includes all the input variable. The experiments were performed in 10 databases. Overall, K-CM showed the best performance when it was compared against k-NN [18], RF [93], PLS-DA [94], MLP [95], CART [96], Logistic [97], LDA [98], and SMO [99]. However, it has a similar performance with k-NN because it has the same modeling principles.

In 2016, Fernandez et al. [100] reviewed evolutionary fuzzy systems (EFSs) for imbalance problems. They introduce some concepts about linguistic fuzzy rules systems. They presented a taxonomy of the methodologies of that moment. Then, they present the most popular methods, until that year, to solve the imbalance problem (i.e., data level, algorithm level, cost-sensitive, and methods embedded into ensemble learning). Finally, to demonstrate the performance of EFSs, they implemented the GP-COACH-H [101] algorithm. The evaluation metric was G-mean and they compared their results against crisp C4.5 with SMOTE + ENN [77]. They used 44 databases from the KEEL repository [55]. In the end, the results are favorable for the EFSs and it outperforms the results of the SMOTE + ENN + C4.5.

In 2016, Fan et al. [102] proposed a fuzzy membership evaluation that assigns the membership value according to the class certainty. Their method is EFSVM and is inspired by FSVM [103]. The method is going to pay more attention to the minority class due to the entropy levels. Their method has three main steps: first, the entropy of the samples in the majority class is calculated. Then, the majority class is divided into subsets. Finally, each sample in the minority class is assigned to a larger fuzzy membership to guarantee the importance of the class. They compared their method, in three synthetic databases, against SVM [104] and the fuzzy method FSVM. Then, they compared their method on 64 real-world databases from the KEEL repository [55], against FSVM, SVM-SMOTE, SVM-OSS- SVM-RUS, EasyEnsemble [105], AdaBoost [106], and 1-NN. The results showed that EFSVM has better results than the compared algorithms in terms of AUC.

In 2017, Zhang et al. [107] introduced KRNN (k rare-class nearest neighbour), which is a modification of the k nearest neighbour (kNN) classifier [18]. KRNN creates groups of objects from the minority class and tries to adjust the induction bias of kNN in correspondence to the size and the distribution of the groups. kRNN also uses the Laplace estimate to adjust the posterior probability estimation for query objects.

In 2017, Zhu et al. [108] proposed an algorithm using an entropy-based fuzzy membership evaluation that enhances the importance of patterns. Their algorithm can handle the imbalance problem as well as matrix patterns. The algorithm is an entropy-based matrix pattern-oriented Ho–Kashyap learning machine with regularization learning (EMatMHKS), and it is based in the MatMHKS [109] algorithm. They divide the classification process into two steps: first, they determine the fuzzy membership for each pattern and then they used the obtained values to solve the criterion function. They used SMOTE [35], ADASYN [110], CBSO [111], EE [105], and BC [112] to preprocess the data (separately), and later they used each of the classifiers. For their experiments, they used ten real-world databases. They used accuracy, sensitivity, specificity, precision, F-measure, and G-mean as their evaluation metrics. They compared their algorithm against SVM, MatMHKS, and the three fuzzy methods FSVM, B-FSVM [113], and FSVM-CIL [114]. In the showed results, EMatMHKS has better overall performance in terms of the evaluation metrics, and it also has lower computational complexity than FSVM and B-FSVM.

In 2017, Pruengkarn et al. [115] proposed an approach to deal with imbalanced data by combining complementary fuzzy support vector machine (CMTFSVM) and synthetic minority oversampling technique (SMOTE). To enhance the classification performance, they used a membership function. The experiments were performed in four benchmark databases from KEEL [55] and UCI repositories [54], and one real-world data database. Their results are compared against crisp NN [116], SVM [16], and the fuzzy method, FSVM [114], showing that combining CMT undersampling technique and an oversampling technique presents the best results.

In 2018, Lee et al. [117] proposed a classifier that handles class imbalance and overlapping problems by geometrically separating the data. Their solution is based on fuzzy support vector machine and k-NN. First, they assign an overlap-sensitive weight scheme that used k-NN and different error cost algorithms. Then, they used an overlap-sensitive margin classifier (OSM) that separates the data into soft and hard overlap regions. For the soft region, the data is classified with the decision boundary of the OSM. For the hard region, they classified the data with 1-NN. Finally, it classified the data using the decision boundary of the separated regions. Their method was tested in 12 synthetics databases and 29 real-world databases from the KEEL [55] repository, where they were evaluated with geometric mean and F1 as their evaluation metrics. Their algorithm was compared against normal SVM, under + SVM, SMOTE + SVM, SDC [118], 1-NN [119], and Boosting SVM [120]. They also compared their algorithm against FSVM-CIL cen [114], FSVM-CIL hyp [114], EFSVM [102], and EMatMHKS [108], in which all of them are fuzzy solutions. All the results were favorable for their proposed method.

In 2019, Gupta et al. [121] proposed two variants of entropy based fuzzy SVM (EFSVM) [102], one that uses least squares, called EFLSSVM-CIL, and another that uses twin support called entropy based fuzzy least squares twin support vector machine (EFLSTWSVM-CIL). In both cases, the fuzzy membership is calculated with the entropy values of each sample. Their algorithms use linear equations to find the decision surface, in contrast with other similar methods that use quadratic programming problem. The performance of the algorithms is evaluated with AUC and the databases used are from the KEEL [55] and UCI [54] repository. The results were tested in linear and non-linear cases, where they compared their methods against TWSVM [122], FTWSVM [122], EFSVM, and NFTWSVM [123], where the last three are fuzzy solutions. Their results showed that EFLSTWSVM-CIL outperforms the other classifiers in terms of learning speed.

In 2019, Arafat et al. [124] proposed an under-sampling method with support vectors. They combined support vector decision boundary points from the majority class with an equal number of points from the minority class. After they obtained a balanced dataset, they used SVM [125] as the base learner for new instances, and they tested their method against C4.5 [52], Naive Bayes [126], Random forest [127], and AdaBoost [128]. The used 13 databases were from the KEEL repository [55]. In their results, they showed that their method outperforms the other classifiers.

In 2019, Liu et al. [129] proposed an algorithm based on adjustable fuzzy classification with a multi-objective genetic strategy based on decomposition. Their algorithm is called AFC-MOGD, which generates fuzzy rules and then those rules are optimized to get new fuzzy rules. They used an optimization algorithm based on decomposition with the purpose of designing a new pattern. They consider the class percentage to determine the class label and the rule weight to get more viable rules. They used 11 databases from the KEEL [55] repository and for the evaluation metric, they used area under the ROC convex hull (AUCH). They compared their algorithm against MOGF-CS [130], C4.5 [52], and E-Algorithm [131]. Their results showed they the new implementation outperforms other classifiers.

In 2019, Cho et al. [132] proposed an instance-based entropy fuzzy support vector machine (IEFSVM). The fuzzy membership of each sample is determined by the entropy information of kNN, which focuses on the importance of each sample. They used polar coordinates to determine the entropy function of each sample. They take into consideration the diversity of entropy to increase the size of the neighbors k. They used 35 databases for the UCI [54] repository and 12 real-world databases. They used AUC to evaluate their performance against SVM [16], undersampling SVM, cost-sensitive SVM, fuzzy SVM [114], entropy fuzzy SVM [102], cost-sensitive AdaBoost [133], cost-sensitive RF [93], EasyEnsemble [105], RUSBoost [85], and weighted ELM [134]. In their results, they proved that IEFSVM has a higher AUC value than the other methods and the results are less sensitive to noise.

In 2019, Sakr et al. [135] proposed a multilabel classification for complex activity recognition. They used emerging patterns and fuzzy sets. Their classification process uses a training dataset of simple activities, where they extract strong jumping emerging patterns (SJEPs). Then, their scoring function uses SJEPs and fuzzy membership values, which result in the labels of the existing activities. They assumed that all the activities are linearly separated. They used three evaluation metrics: precision, recall, and F-measure. The results outperform the other compared crisp methods (i.e., SVM [16], k-NN [18], and HMM [136]).

In 2019, Patel et al. [137] proposed a combination of adaptive K-nearest neighbor (ADPTKNN) and fuzzy K-nearest neighbor. The adaptive K-nearest neighbor part is used to overcome the difficulties presented in fuzzy K-nearest neighbor when it deals with imbalanced databases. Their new addition chooses different values of k for different classes according to their sizes. They tested their work in: ten databases, different values of K (i.e., 5-25 in steps of 5), and evaluates F-measure, AUC, and G-mean. They compared their algorithm (Fuzzy ADPTKNN) with NWKNN [138], Adpt-NWKNN [139], and Fuzzy NWKNN [140]. In most of the presented cases, their proposed method performs better than the other approaches.

In 2019, García-Vico et al. [42] proposed an algorithm for big data environments (BD-EFEP). They claimed it is the first multi-objective evolutionary algorithm for pattern mining in big data. BD-EFEP has the objective of extract high-quality emerging patterns that describes the discriminative characteristics of the data. BD-EFEP follows a competitive-cooperative schema, where the patterns compete with each other, but they cooperate to describe the space's greatest possible area. To improve efficiency in evaluating the individual without losing quality, they used a MapReduce-based global approach or a token competition-based procedure. In conclusion, BD-EFEP obtains a set of patterns with an improvement in the trade-off between the generality and the reliability of the results. These patterns are extracted faster than other approaches, which makes the algorithm relevant for big data environments.

In 2019, García-Vico et al. [141] presented the effects of different quality measures that are used as objectives for multi-objectives approaches, focused on the extraction of emerging patterns (EPs) in big data environments. EPs are a type of contrast patterns extracted by using the quality measure Growth Rate, which is computed using the mean of the highest ratio between the support of a certain pattern in one class and the support of the same pattern in the other classes [41]. They used eight combinations of quality measures:

Jac and TPR, G-mean and Jac, Jac and FPR, G-mean and WRacc, Jac and WRacc, SuppDiff and Jac, TPR and FPR, and WRacc and SupDiff [142]. They performed the experiments in six large-scale databases from the UCI repository [54]. The number of instances goes up to 11 million, but the number of variables is low. They used the BD-EFEP [42] algorithm, which is one of the two algorithms for the extraction of EPs (i.e., EvAEFP-Spark [143] and the BD-EFEP). In their results, they showed that the combination of TPR and FPR, G-mean and Jac, and Jac and FPR, are the most suitable quality measures for the descriptive rules in emerging pattern mining. Nevertheless, Jac and FPR is shown as the best trade-off in the descriptive rules for EPs.

In 2019, Luna et al. [144] proposed three different subgroup discovery approaches for mining subgroups in multiple instances problems. The proposed approaches were based on an exhaustive search approach (SD-Map [145]), an evolutionary algorithm based on grammar-guided genetic programming (CGBA-SD [146]), and an evolutionary fuzzy system (NMEEF-SD [147]). Those approaches are adapted to multiple instance problems. The approaches were tested in two different steps: first, they used ten benchmarks representing real-world applications, and then, 20 artificial databases generated by themselves. Their results showed that SD-Map-MI has poor performance with low quality and bad trade-off between sensitivity and confidence. CGBA-SD-MI has a good trade-off between sensitivity and confidence, as well as good performance in the different databases. NMEEF-SD-MI also has rules of low quality and bad results on confidence and low sensitivity; however, some databases present a confidence level higher than 83%. Finally, they conclude that they need to keep performing more tests because it can be a promising approach to medicine or Bioinformatics problems.

In 2020, García-Vico et al. [148] proposed a method based on an adaptive version of the NSGA-II algorithm. They proposed a cooperative-competitive multi-objective evolutionary fuzzy system called E2PAMEA. The algorithm is used to extract emerging patterns in big data environments. E2PAMEA uses an adaptive schema that selects different genetic operators. The genetic operators used for diversity are two-point and HUX crossover. Similar to their previous work [42], they used a cooperative-competitive schema along with a token competition-based procedure for reliable results. Finally, E2PAMEA outperforms previous approaches showing that their results are more reliable, and the processing time and physical memory are also improved.

In 2020, Liu [21] presented an extension of the fuzzy support vector machines for class imbalance learning (FSVM-CIL) method. The idea is to overcome the limitation of FSVM-CIL for imbalanced databases with borderline noise (i.e., between-class borderline). To solve this limitation, a new distance measure and a new fuzzy function are introduced. The distance measure characterizes the distance of different data points to the hyperplane, and the Gaussian fuzzy function reduces the impact of the borderline noise. The experiments were performed on several public imbalanced datasets. The results show an improvement over the original FuzzySVM-CIL [114], SVM [16], CS-SVM [149], and SMOTE-SVM [150].

In 2020, Gu et al. [151] proposed a self-adaptive synthetic over-sampling approach (SASYNO). They synthesize data samples near the real data samples, giving priority to the minority class. Their method balances the classes and boosts the performance. They claimed it could be applied to different base algorithms such as SVM, k-NN, DP, rule-based classifiers, DT, and others. They used four databases for their experiments, and the results showed that their approach enhances the overall performance in terms of specificity, F-measure, and overall ranks. The classifiers where they applied their approach were SONFIS [152], SVM [125], and k-NN [153].

In 2020, García-Vico et al. [154] presented a preliminary many objective algorithm for extracting Emerging Fuzzy Patterns (ManyObjective-EFEP). They combined fuzzy logic, for soft-computing, with NSGA-III [155]; the purpose is to increase the search space in emerging pattern mining. The algorithm is compared against an adaptation of NSGA-II [156] and they used 46 databases from the UCI repository [54]. The quality measures

used are nP, nV, WRACC, CONF, GR, TPR, and FPR [142]. They presented results of a set of patterns with more interpretability and precision. They conclude that the algorithm could be interesting for real-world applications. However, they need to continue the study and analysis due to the complex space and the trade-off among different quality measures.

Finally, in Table 1, we show key merit(s) and disadvantage(s) or possible improvements for each proposal presented in Section 3.3.

**Table 1.** Algorithm-level approaches with their key merit(s) and disadvantage(s).

Year	Ref.	Key Merit(s)	Disadvantage(s)/Improvements
2011	[20]	-Proposed a fuzzy emerging pattern technique. -Proposed the FEPC classifier.	-Fuzzy emerging patterns are highly dependent on the quality measure Growth Rate.
2014	[91]	-Proposed K-Contractive Map (K-CM).	-Similar performance with a classical k-NN classifier.
2016	[100]	-A study in evolutionary fuzzy systems (EFSs) for imbalance problems.	-A taxonomy of the reviewed methods is missing.
2016	[102]	-Presented EFSVM. A fuzzy membership evaluation that assigns the membership value according to the class certainty.	-Unfortunately, they do not mention possible improvements in their work.
2017	[107]	-Proposed KRNN based on a KNN classifier.	-KRNN could be extended to multiple classes.
2017	[108]	-Proposed EMatMHKS. An algorithm with an entropy-based fuzzy membership and based in MatMHKS.	-An improvement in their function to measure the entropy-based fuzzy membership.
2017	[115]	-Proposed a combination of CMTFSVM and SMOTE. entropy-based fuzzy.	-A comparison with other popular classifiers can enrich the results.
2018	[117]	-Proposed a classifier based on OSM, and inspired on FSVM and k-NN. It geometrically separates the data to solve the imbalance problem.	-It uses a 1-NN algorithm for the hard-overlapping regions, which can result in a high generalization error.
2019	[121]	-Proposed two variants of EFSVM. One uses least squares and the other one uses twin SVM.	-The results could improve with the implementation of heuristics solutions in the method for parameter selection.
2019	[124]	-Proposed a method to generate balanced data with support vectors.	-The usage of real-world data can enrich the results of their method.
2019	[129]	-Proposed AFC-MOGD. An algorithm based on adjustable fuzzy classification with a multi-objective genetic strategy.	-The usage of real-world data can enrich the results of their method.
2019	[132]	-Proposed an instance-based EFSVM.	-The selection of different neighborhood sizes could improve the results due to the better knowledge of the distribution of the data.
2019	[135]	-Proposed a multilabel classification for complex activity recognition.	-A comparison against other fuzzy classifiers would enrich the results.
2019	[137]	-Proposed Fuzzy ADPTKNN. It combines ADPTKNN and fuzzy k-NN.	-The method could be extended to feature-based NN.
2019	[42]	-Proposed BD-EFEP. An algorithm for big data environments.	-Additional comparisons against big data environments can enrich the results.
2019	[141]	-Presents the effects of different quality measures in patterns, focused on Big Data Environments.	-New approaches for efficient extraction of patterns in big data environments are needed.

Table 1. Cont.

Year	Ref.	Key Merit(s)	Disadvantage(s)/Improvements
2019	[144]	-Proposed three approaches for mining subgroups in multiple instances problems.	-More tests are needed to improve their results and to determine the imbalance ratio in which the method is more suitable.
2020	[148]	-Proposed an adaptive version of NSGA-II.	-An optimization of the fuzzy sets could improve the results.
2020	[21]	-Proposed an extension of FSVM-CIL with a new distance measure and a new fuzzy function.	-A comparison against more fuzzy classifiers would enrich the results.
2020	[151]	-Proposed SASYNO. A self-adaptive synthetic over-sampling approach.	-The usage of more databases would enrich the results.
2020	[154]	-Presents a preliminary many objective algorithm for extracting Emerging Fuzzy Patterns.	-The usage of real-world databases would enrich the results.

### 3.4. Discussion

In the reviewed documents, the proposed methods with fuzzy pattern-based classifiers have shown better classifier results than other non-fuzzy approaches. Despite the favorable results, fuzzy pattern-based classifiers have not been studied more in detail. To the best of our knowledge, the only classifier based on fuzzy patterns is that proposed in [20]. Additionally, there are no fuzzy pattern-based classifiers specifically designed to deal with class imbalance problems and the approaches that are closer to this type of problem have not been studied enough.

Fuzzy approaches have achieved those results for different reasons; one of them is that the results are similar to the language used by experts. This solution is due to the use of linguistic hedges (e.g., “very”, “less”, “often”) that overcome the problems presented with the discretization of continuous features. Another characteristic is the flexibility the offers the fuzzification of the data. The fuzzification gives a wider window of opportunity in different classification steps (i.e., mining process, filtering process, variable window size).

## 4. Applications Domains

In this next section, we present the papers that the research was used directly into an application domain. As we see, most research was conducted in the field of medicine, followed by that of the financial market.

In 2017, Gerald Schaefer [157] showed strategies to handle the class imbalance problem in digital pathology. In the paper, it is showed how fuzzy rule-based algorithms can use cost-sensitive functions to include misclassification cost and how to use ensemble classification methods for imbalanced data. They evaluate the methods in the AIDPATH project [157] and are waiting to get results to make them public.

In 2018, Jaafar et al. [158] proposed a classifier named Mahalanobis fuzzy k-nearest centroid neighbor (MFkNCN). First, they calculate the k-nearest centroid neighborhood for each class. Then, they used a fuzzy membership to designate the correct class label of the query sample. They used a real-world database from the Shiraz University, where it has the adult heart sound of 112 subjects (79 healthy people and 33 patients). They used accuracy as their performance metric. They compared their classifier against kNN, kNCN, and two fuzzy approaches, FkNN and MFkCNN. Their results showed that MFkNCN outperforms the other classifiers.

In 2018, Kemal Polat [159] proposed a data preprocessing method to classify Parkinson, hepatitis, Pima Indians, single proton emission computed thoracic heart, and thoracic surgery medical databases. Their method is divided into three main steps. First, they used k-means, fuzzy c-means, and mean shift to calculate the cluster centers. Secondly, they calculated the absolute differences between each attribute and the cluster centers; the average of the differences of each attribute is also calculated. In the last step, they proposed three attribute weighting methods to reduce the variance within the class. If the similarity

rate is low, the coefficient will be high, and if the similarity rate is high, the coefficient will be low. Their method is compared against random sampling and they used Linear discriminant analysis, k-NN, SVM [16], and RT as classification methods. The databases are from the UCI [54] repository. The performance is evaluated with accuracy, precision, Recall, AUC, j-value, and F-measure. Finally, their results showed that their approach has better results than random sampling method. They claimed that the three proposed attribute weighting methods could be used in signal and image classification.

In 2019, Cho et al. [160] proposed an investment decision model for a peer-to-peer (P2P) lending market based on the instance-based entropy fuzzy support vector machine (IEFSVM). The idea of this algorithm is based on the necessity of an effective prediction method that supports decisions in the P2P lending market. First, for the training set, they select the desired features and train the loan status with IEFSVM. Secondly, for the test set, they remove the unselected features and predict the loan status using the same IEFSVM. In the end, the method selects the loans classified as fully paid. The database used is from the Lending Club, which is specialized in P2P loans. They used AUC, precision, predicted negative condition rate as their evaluation metrics. They compared their method against Cost-sensitive adaptive boosting [133], cost-sensitive random forest [93], EasyEnsemble [105], random undersampling boosting [85], weighted extreme learning machine [134], cost-sensitive extreme gradient boosting [161], and one fuzzy approach, EFSVM [102]. Their results showed that their method outperforms the other classifiers for loan status classification.

In 2019, Li et al. [162] proposed a Bayesian Possibilistic C-means (BPCM) clustering algorithm that provides a preliminary cervical cancer diagnostic. They explained that the motivation of the research is due the high lethality and morbidity of the disease. The information used in the diagnostic process could be highly private and many patients decided not to provide this information. To solve the lack of attributes for the classification, they look for patterns from the complete data, and then, they estimated the missing values with the closest representative pattern. In the first stage, they used a fuzzy clustering approach that models the noise and uncertainty of the database. During the second stage, they use a Bayesian approach to get the cluster centroids which can be translated as the representative. They used accuracy and sensitivity as their evaluation metrics. They tested their algorithm in a database with 858 patients. The proposed solution can identify the patterns and is able to predict the missing values with an accuracy of 76%

In 2020, Ambika et al. [22] proposed a method that looks in medical databases to predict the risk of developing hypertension; as they mentioned, most medical databases have the class imbalance problem. Their method is divided into four phases: data collection, data preparation, learning model, and data exploration. For the data collection, they used databases from primary health centers. In the preparation phase, they used mean substitution [163] for missing values and interquartile range [164] to detect outliers. The learning model uses SVM with AdaBoost [165]. Finally, for the exploration phase, they used fuzzy-based association mining rule. They used precision, recall, accuracy, and F-measure to evaluate their performance. They compared their results against ANN, IBK, and NaiveBayes. Their results showed an improvement in medical diagnosis, but they said that it needs further work in classification errors.

In 2020, García-Vico et al. [166] proposed an algorithm, called FEPDS, to extract fuzzy emerging patterns in Data Stream environments. The algorithm can be described in three main steps, a recollection, an adaptive learning step, and a learning multi-objective evolutionary algorithm. First, they collected instances by batches and once the batch is collected, the learning method updates the fuzzy patterns model according to the current batch. Then, they used a multi-objective evolutionary learning algorithm that extracts pattern models according to simplicity and reliability. Finally, the algorithm is continuously evaluating the instance against unseen instances to determine the quality of the new information acquired. The algorithm was tested in the profiles of customers of New York City taxis corresponding to the fare amount; it was also included recurrent

behaviors such as payment by card on medium or high fares and the daily travel to work of the citizens. They conclude that FEPDS can adapt the model to abrupt changes without losing performance in terms of execution time. The algorithm also has great scalability and performance in terms of memory consumption, allowing its usage in real-world environments.

## 5. Taxonomy

In the reviewed papers, we classified the type of approach and presented a table (Table 2) with some advantages and disadvantages. The table is divided into theoretical and real-world applications, and subdivided into the three approaches commonly used to deal with class imbalance problems [15,31,34]. This division was used to be concise in the common categories found in the literature and follow the same methodology that we used for the systematic review of the state of the art for class imbalance problems on fuzzy and pattern-based classification (i.e., theoretical and practical approaches). Another advantage is to quickly identify the type of approach with its benefits and drawbacks.

We create a taxonomy that is a summary of the similarities and differences among the different approaches. Most of the papers are focused on the algorithm level due to the possible improvement in their results. However, those approaches are limited to the specific problem to solve. In real-world applications, most of them are related to medicine, with some exceptions in finance and urban planning. Additionally, in real-world applications, it is common to combine different approaches to deal with the problem ahead to get the most of each approach (i.e., data level and algorithm level been to most common combination).

Figure 1 shows the results of the subject area of the 62 papers that cover fuzzy sets, pattern-based, and imbalance problems without real-world applications. Figure 2 shows the number of publications by year.

As is expected, the three main subject areas are Computer Science, Mathematics, and Engineering. Additionally, Figure 2 shows the increase in published documents in the field, where 2018 was the year with the higher number of publications across the 15 reviewed years.

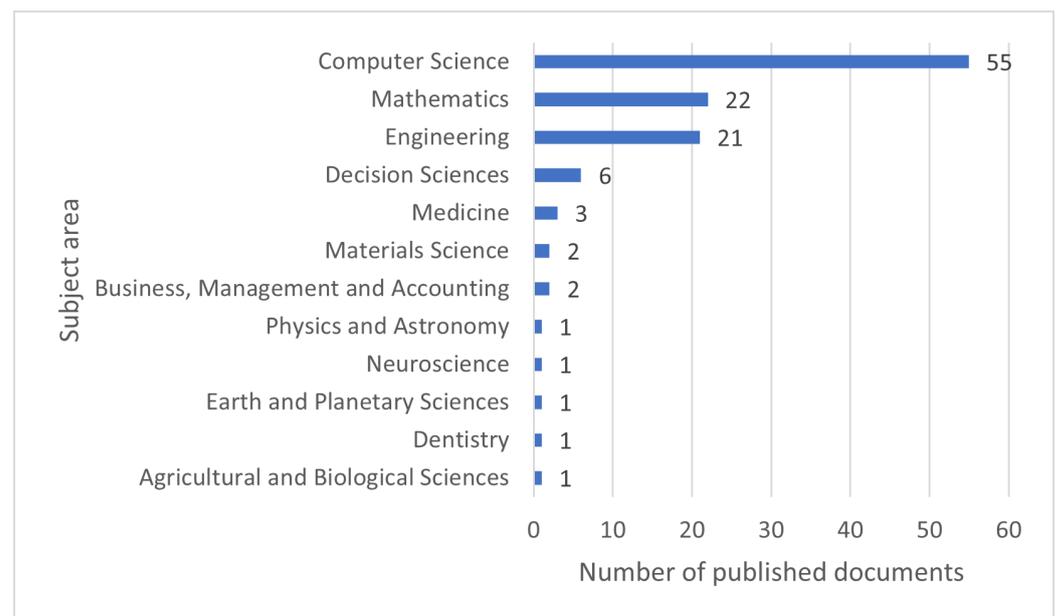
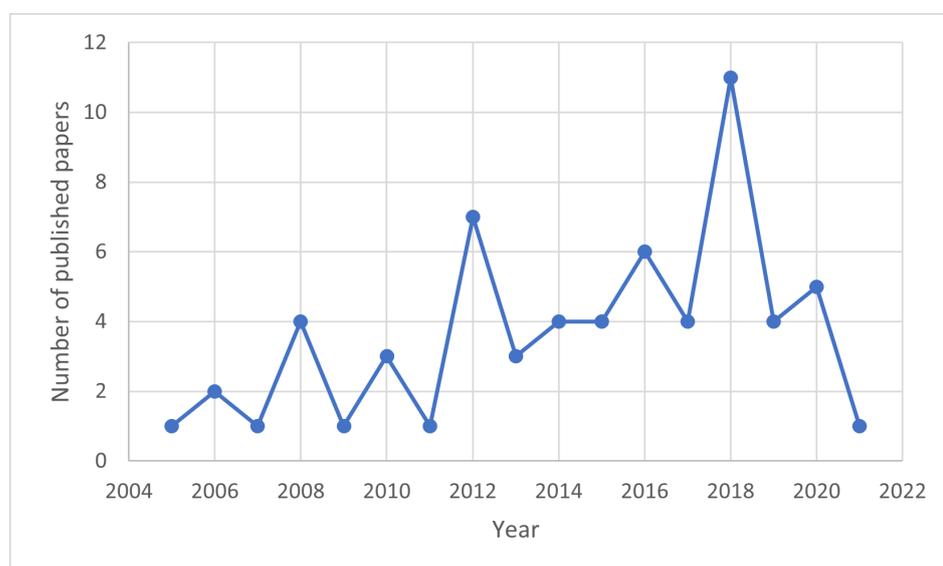


Figure 1. Number of published documents by subject area.

**Table 2.** Taxonomy of similarities and differences among the reviewed papers.

Application	Approach	Refs.	Advantage	Disadvantage
Theoretical	Data level	[51,63,67,69,72]	Data is artificially manipulated to deal with the imbalance problem.	The data created can lead to a bias in the classification process.
	Algorithm level	[20,21,42,91,100,102,107,108,115,117,121,124,129,132,135,137,141,144,148,151,154]	The results from the experiments tend to have positive results due to the fitting process of the problem.	Each solution solves the imbalance problem in their own scenario, which does not present a general solution for most cases.
	Cost-sensitive	-	-	-
Medicine	Data level Algorithm level	[22]	The used methods have been deeply studied and have better results against other learning techniques.	They need to enhance how they handle classification errors.
	Algorithm level	[158]	The usage of Mahalanobis distance in combination with fuzzy kNN, improves the results of normal kNN.	The results are limited to the specific case.
	Data level Algorithm level	[159]	The approach can be used in diseases, signal, and image classification	Optimization could be improved.
	Algorithm level	[162]	The author mentioned that there are room for improvements, and they have promising results.	Accuracy and sensitivity are somehow low, 76% and 79% respectively.
	Data level Cost-sensitive	[157]	Improved robustness and classification performance in contrast of a single lassifier.	The results are not known, but they will report them in future publications.
Financial	Algorithm level	[160]	Robust results for the P2P lending market.	The decision model could be considered too simple.
Urban planning	Algorithm level	[166]	The model can adapt to abrupt changes and has great scalability.	It is unknown how it will perform outside the pilot city.



**Figure 2.** Number of published documents by year.

## 6. Future Directions

From the studies reviewed in this paper, we can see that a prominent approach to solve the class imbalance problem is by using fuzzy classification due to the flexibility that fuzzy approaches provide. Nevertheless, the approaches previously reviewed are extensions of previous solutions. Pattern-based solutions have positive classification results, but they have been limited to use emerging patterns only, which contains several limitations [142].

Analyzing the results of the reviewed work and exploiting the advantages of each approach, we infer that fuzzy pattern-based classifiers specifically designed to deal with class imbalance problems could achieve better results with a statistically significant difference than other popular state-of-the-art fuzzy and pattern-based classifiers.

To carry out this fuzzy pattern-based solution, better algorithms for mining fuzzy patterns in class imbalance problems and new fuzzy pattern-based classifiers for class imbalance problems are still needed. We proposed the following strategies on each stage of the pattern-based classification (i.e., mining, filtering, and classification) to improve the result when dealing with imbalanced data.

In the mining stage, it is needed to extract fuzzy patterns that are representatives of all classes. One approach is by inducing diverse fuzzy decision trees, which has shown better results than other contrast pattern mining approaches [20]. Then it is needed to analyze whether the extracted fuzzy patterns are representatives of all classes. Further, it is important to analyze the sources of bias, mainly focusing on strategies to mitigate fragmentation data and noise.

In the filtering stage, it is needed to obtain a representative subset of patterns for all problem's classes. The subset needs to be, at least, as good to classify as the set containing all the fuzzy patterns. To obtain the subset, the implementation of strategies based on weights assigned to fuzzy patterns, considering the class imbalance ratio, could be used; with this improvement, the problem of high computational cost generated by the exponential number of candidate patterns [12,39–41] could be reduced. Another possible improvement is modifying or developing a quality measure for fuzzy patterns that could improve the extracted patterns.

In the classification stage, the voting scheme needs to be adapted for fuzzy pattern-based classification. To improve this step, we could consider weighting the objects in consequence of the class imbalance ratio presented by the training dataset. For the classifier, the result of each stage is of significant importance due to the interconnected nature of all the stages.

Finally, to the best of our knowledge, fuzzy pattern-based classification has not been studied in detail, and it is a good area of opportunity in the research field.

## 7. Conclusions

In this paper, we review and analyze the current state-of-the-art of classifiers that involved fuzzy techniques and pattern-based classifiers for solving class imbalance problems. This research was motivated by the renewed interest of the machine learning community in creating explainable artificial intelligence models and the wide interest in applying this type of model to the real world.

Therefore, in this paper, we performed an in-depth review of current state-of-the-art techniques. We reviewed 84 papers that include fuzzy and/or pattern-based approaches for imbalanced databases, as well as papers that have cited the work of Garcia-Borroto M et al. We also included the most common applications domains. Next, we present a taxonomy that includes the type of approach with its advantages and disadvantages. Then, we include future directions of the field according to the reviewed papers. As a final part of the review, we provide the conclusions and some advantages and disadvantages of all the reviewed approaches.

- Advantages:
  - Fuzzy and pattern-based approaches attract interest from the research community.
  - Fuzzy logic is widely used for its flexibility and understandability of the results.
  - Medicine is an area where the imbalance problem is constantly presented and uses the newest techniques.
  - Techniques that include fuzzy approaches have shown better classification results in comparison to other classifiers based on non-fuzzy approaches.
  - Fuzzy pattern-based approaches are a promising solution to handle the imbalanced data problem. However, this type of classifier should be studied further.
- Disadvantages:
  - Despite the flexibility of fuzzy approaches, they can lead to repetitive solutions that are small variations of other ones.
  - The quality of fuzzy patterns is highly dependent on the quality of the features of the fuzzification process.
  - Fuzzy emerging patterns are highly dependent on the quality measure Growth Rate, which could not provide good patterns as stated in [142].
  - The combination of fuzzy and pattern-based approaches has not been studied in detail, so some research can lead to a dead end.

In summary, we present a survey that covers an in-depth review of papers from 2005 to 2021 that include fuzzy sets, pattern-based, imbalance problems, and documents that have cited [20]. Then, we present our analysis and show possible future directions. Finally, we finish with a conclusion of the solutions to classification for imbalanced data.

**Author Contributions:** Conceptualization, O.L.-G.; Methodology, I.L. and O.L.-G.; Software, I.L. and O.L.-G.; Validation, O.L.-G., R.M. and M.A.M.-P.; Formal analysis, I.L.; Investigation, I.L.; Resources, O.L.-G. and M.A.M.-P.; Data Curation, O.L.-G. and I.L.; Writing—Original Draft Preparation, I.L.; Writing—Review and Editing, O.L.-G., R.M. and M.A.M.-P.; Visualization, O.L.-G.; Supervision, O.L.-G., R.M. and M.A.M.-P.; Project Administration, I.L. and O.L.-G.; Funding Acquisition, R.M. All authors have read and agreed to the published version of the manuscript.

**Funding:** This work was partly supported by the National Council of Science and Technology of Mexico under the scholarship grant 1008244.

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

**Informed Consent Statement:** Not applicable.

**Data Availability Statement:** Not applicable.

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

## References

1. An, B.; Suh, Y. Identifying financial statement fraud with decision rules obtained from Modified Random Forest. *Data Technol. Appl.* **2020**, *54*, 235–255. [[CrossRef](#)]
2. De Bock, K.W.; Coussement, K.; Lessmann, S. Cost-sensitive business failure prediction when misclassification costs are uncertain: A heterogeneous ensemble selection approach. *Eur. J. Oper. Res.* **2020**, *285*, 612–630. [[CrossRef](#)]
3. Kim, T.; Ahn, H. A hybrid under-sampling approach for better bankruptcy prediction. *J. Intell. Inf. Syst.* **2015**, *21*, 173–190.
4. Zhou, L. Performance of corporate bankruptcy prediction models on imbalanced dataset: The effect of sampling methods. *Knowl. Based Syst.* **2013**, *41*, 16–25. [[CrossRef](#)]
5. Mazurowski, M.A.; Habas, P.A.; Zurada, J.M.; Lo, J.Y.; Baker, J.A.; Tourassi, G.D. Training neural network classifiers for medical decision making: The effects of imbalanced datasets on classification performance. *Neural Netw.* **2008**, *21*, 427–436. [[CrossRef](#)]
6. Goyal, D.; Choudhary, A.; Pabla, B.; Dhama, S. Support vector machines based non-contact fault diagnosis system for bearings. *J. Intell. Manuf.* **2020**, *31*, 1275–1289. [[CrossRef](#)]
7. Zhu, Z.B.; Song, Z.H. Fault diagnosis based on imbalance modified kernel Fisher discriminant analysis. *Chem. Eng. Res. Des.* **2010**, *88*, 936–951. [[CrossRef](#)]
8. Fawcett, T.; Provost, F. Adaptive fraud detection. *Data Min. Knowl. Discov.* **1997**, *1*, 291–316. [[CrossRef](#)]

9. Minastireanu, E.A.; Gabriela, M. Methods of Handling Unbalanced Datasets in Credit Card Fraud Detection. *BRAIN. Broad Res. Artif. Intell. Neurosci.* **2020**, *11*, 131–143. [[CrossRef](#)]
10. Gao, X.; Chen, Z.; Tang, S.; Zhang, Y.; Li, J. Adaptive weighted imbalance learning with application to abnormal activity recognition. *Neurocomputing* **2016**, *173*, 1927–1935. [[CrossRef](#)]
11. Koziarski, M.; Kwolek, B.; Cyganek, B. Convolutional neural network-based classification of histopathological images affected by data imbalance. In *Video Analytics. Face and Facial Expression Recognition*; Springer: Berlin/Heidelberg, Germany, 2018; Volume 11264, pp. 1–11.
12. Yu, H.; Ni, J.; Dan, Y.; Xu, S. Mining and integrating reliable decision rules for imbalanced cancer gene expression data sets. *Tsinghua Sci. Technol.* **2012**, *17*, 666–673. [[CrossRef](#)]
13. Olszewski, D. A probabilistic approach to fraud detection in telecommunications. *Knowl. Based Syst.* **2012**, *26*, 246–258. [[CrossRef](#)]
14. Chen, L.; Dong, G. Using Emerging Patterns in Outlier and Rare-Class Prediction. In *Contrast Data Mining: Concepts, Algorithms, and Applications*; CRC Press: Boca Raton, FL, USA, 2013; pp. 171–186.
15. Loyola-González, O.; Medina-Pérez, M.A.; Martínez-Trinidad, J.F.; Carrasco-Ochoa, J.A.; Monroy, R.; García-Borroto, M. PBC4cip: A new contrast pattern-based classifier for class imbalance problems. *Knowl. Based Syst.* **2017**, *115*, 100–109. [[CrossRef](#)]
16. Cortes, C.; Vapnik, V. Support-vector networks. *Mach. Learn.* **1995**, *20*, 273–297. [[CrossRef](#)]
17. Quinlan, J.R. Bagging, boosting, and C4.5. In *Proceedings of the Conference on Artificial Intelligence*, Portland, OR, USA, 4–8 August 1996; Volume 1, pp. 725–730.
18. Aha, D.W.; Kibler, D.; Albert, M.K. Instance-based learning algorithms. *Mach. Learn.* **1991**, *6*, 37–66. [[CrossRef](#)]
19. López, V.; Fernández, A.; García, S.; Palade, V.; Herrera, F. An insight into classification with imbalanced data: Empirical results and current trends on using data intrinsic characteristics. *Inf. Sci.* **2013**, *250*, 113–141. [[CrossRef](#)]
20. García-Borroto, M.; Martínez-Trinidad, J.F.; Carrasco-Ochoa, J.A. Fuzzy emerging patterns for classifying hard domains. *Knowl. Inf. Syst.* **2011**, *28*, 473–489. [[CrossRef](#)]
21. Liu, J. Fuzzy support vector machine for imbalanced data with borderline noise. *Fuzzy Sets Syst.* **2020**. [[CrossRef](#)]
22. Ambika, M.; Raghuraman, G.; SaiRamesh, L. Enhanced decision support system to predict and prevent hypertension using computational intelligence techniques. *Soft Comput.* **2020**, *24*, 13293–13304. [[CrossRef](#)]
23. Loyola-González, O. Black-box vs. white-box: Understanding their advantages and weaknesses from a practical point of view. *IEEE Access* **2019**, *7*, 154096–154113. [[CrossRef](#)]
24. Loyola-González, O.; Gutierrez-Rodríguez, A.E.; Medina-Pérez, M.A.; Monroy, R.; Martínez-Trinidad, J.F.; Carrasco-Ochoa, J.A.; García-Borroto, M. An Explainable Artificial Intelligence Model for Clustering Numerical Databases. *IEEE Access* **2020**, *8*, 52370–52384. [[CrossRef](#)]
25. García-Borroto, M.; Martínez-Trinidad, J.F.; Carrasco-Ochoa, J.A.; Medina-Pérez, M.A.; Ruiz-Shulcloper, J. LCMine: An efficient algorithm for mining discriminative regularities and its application in supervised classification. *Pattern Recognit.* **2010**, *43*, 3025–3034. [[CrossRef](#)]
26. Zhang, X.; Dong, G. Overview and analysis of contrast pattern based classification. In *Contrast Data Mining: Concepts, Algorithms, and Applications*; Chapman and Hall/CRC: Boca Raton, FL, USA, 2016; Volume 11, pp. 151–170.
27. Liu, C.; Cao, L.; Philip, S.Y. Coupled fuzzy k-nearest neighbors classification of imbalanced non-IID categorical data. In *Proceedings of the 2014 International Joint Conference on Neural Networks*, Beijing, China, 6–11 July 2014; pp. 1122–1129.
28. Dong, G.; Bailey, J. *Contrast Data Mining: Concepts, Algorithms, and Applications*; CRC Press: Boca Raton, FL, USA, 2012.
29. Duan, L.; García-Borroto, M.; Dong, G. More Expressive Contrast Patterns and Their Mining. In *Contrast Data Mining: Concepts, Algorithms, and Applications*; CRC Press: Boca Raton, FL, USA, 2013; pp. 89–108.
30. Napierala, K.; Stefanowski, J. Types of minority class examples and their influence on learning classifiers from imbalanced data. *J. Intell. Inf. Syst.* **2016**, *46*, 563–597. [[CrossRef](#)]
31. Lin, W.C.; Tsai, C.F.; Hu, Y.H.; Jhang, J.S. Clustering-based undersampling in class-imbalanced data. *Inf. Sci.* **2017**, *409–410*, 17–26. [[CrossRef](#)]
32. Denil, M.; Trappenberg, T. Overlap versus imbalance. In *Proceedings of the Canadian Conference on Artificial Intelligence*, Ottawa, ON, Canada, 31 May–2 June 2010; Springer: Berlin/Heidelberg, Germany, 2010; pp. 220–231.
33. Beyan, C.; Fisher, R. Classifying imbalanced data sets using similarity based hierarchical decomposition. *Pattern Recognit.* **2015**, *48*, 1653–1672. [[CrossRef](#)]
34. Haixiang, G.; Yijing, L.; Shang, J.; Mingyun, G.; Yuanyue, H.; Bing, G. Learning from class-imbalanced data: Review of methods and applications. *Expert Syst. Appl.* **2017**, *73*, 220–239. [[CrossRef](#)]
35. Chawla, N.; Bowyer, K.; Hall, L.; Kegelmeyer, P. SMOTE: Synthetic minority over-sampling technique. *J. Artif. Intell. Res.* **2002**, *16*, 321–357. [[CrossRef](#)]
36. Wang, K.J.; Makond, B.; Chen, K.H.; Wang, K.M. A hybrid classifier combining SMOTE with PSO to estimate 5-year survivability of breast cancer patients. *Appl. Soft Comput.* **2014**, *20*, 15–24. [[CrossRef](#)]
37. Gao, M.; Hong, X.; Harris, C.J. Construction of neurofuzzy models for imbalanced data classification. *IEEE Trans. Fuzzy Syst.* **2013**, *22*, 1472–1488. [[CrossRef](#)]
38. Kim, J.; Choi, K.; Kim, G.; Suh, Y. Classification cost: An empirical comparison among traditional classifier, Cost-Sensitive Classifier, and MetaCost. *Expert Syst. Appl.* **2012**, *39*, 4013–4019. [[CrossRef](#)]

39. Dong, G.; Li, J.; Wong, L. The use of emerging patterns in the analysis of gene expression profiles for the diagnosis and understanding of diseases. *New Gener. Data Min. Appl.* **2005**, 331–354. [[CrossRef](#)]
40. Han, J.; Cheng, H.; Xin, D.; Yan, X. Frequent pattern mining: Current status and future directions. *Data Min. Knowl. Discov.* **2007**, *15*, 55–86. [[CrossRef](#)]
41. Gonzalez, O.L. Supervised Classifiers Based on Emerging Patterns for Class Imbalance Problems. Ph.D. Thesis, Coordinación de Ciencias Computacionales, National, Puebla, Mexico, 2017.
42. García-Vico, Á.M.; González, P.; Carmona, C.J.; del Jesus, M.J. A Big Data Approach for the Extraction of Fuzzy Emerging Patterns. *Cogn. Comput.* **2019**, *11*, 400–417. [[CrossRef](#)]
43. Nguyen, H.T.; Walker, C.L.; Walker, E.A. *A First Course in Fuzzy Logic*; CRC Press: Boca Raton, FL, USA, 2018.
44. Ross, T.J. *Fuzzy Logic with Engineering Applications*; Wiley Online Library: Hoboken, NJ, USA, 2004; Volume 2.
45. Zadeh, L.A. Fuzzy sets. *Inf. Control* **1965**, *8*, 338–353. [[CrossRef](#)]
46. Lior, R.; Oded, M. *Data Mining with Decision Trees: Theory and Applications*; World Scientific: Singapore, 2014; Volume 81.
47. Zimmermann, H.J. *Fuzzy Set Theory and Its Applications*; Springer Science & Business Media: Berlin/Heidelberg, Germany, 2011.
48. Gramann, K.D.M. Fuzzy classification: An overview. *Fuzzy-Syst. Comput. Sci.* **1994**, 277–294. [[CrossRef](#)]
49. Orazbayev, B.; Ospanov, E.; Orazbayeva, K.; Kurmangazieva, L. A hybrid method for the development of mathematical models of a chemical engineering system in ambiguous conditions. *Math. Model. Comput. Simulations* **2018**, *10*, 748–758. [[CrossRef](#)]
50. Werro, N. *Fuzzy Classification of Online Customers*; Springer: Berlin/Heidelberg, Germany, 2015.
51. Liu, S.; Zhang, J.; Xiang, Y.; Zhou, W. Fuzzy-based information decomposition for incomplete and imbalanced data learning. *IEEE Trans. Fuzzy Syst.* **2017**, *25*, 1476–1490. [[CrossRef](#)]
52. Quinlan, J.R. *C4.5: Programs for Machine Learning*; Elsevier: Amsterdam, The Netherlands, 2014.
53. Shirabad, J.S.; Menzies, T.J. The PROMISE repository of software engineering databases. *Sch. Inf. Technol. Eng. Univ.* **2005**, 24
54. Dua, D.; Graff, C. *UCI Machine Learning Repository*; UCI: Irvine, CA, USA, 2017
55. Alcalá-Fdez, J.; Fernández, A.; Luengo, J.; Derrac, J.; García, S.; Sánchez, L.; Herrera, F. Keel data-mining software tool: Data set repository, integration of algorithms and experimental analysis framework. *J. Mult. Valued Log. Soft Comput.* **2011**, *17*, 255–287
56. Bekkar, M.; Djemaa, H.K.; Alitouche, T.A. Evaluation measures for models assessment over imbalanced data sets. *J. Inf. Eng. Appl.* **2013**, *3*, 27–39.
57. Zhu, X.; Zhang, S.; Jin, Z.; Zhang, Z.; Xu, Z. Missing value estimation for mixed-attribute data sets. *IEEE Trans. Knowl. Data Eng.* **2010**, *23*, 110–121. [[CrossRef](#)]
58. Pan, R.; Yang, T.; Cao, J.; Lu, K.; Zhang, Z. Missing data imputation by K nearest neighbours based on grey relational structure and mutual information. *Appl. Intell.* **2015**, *43*, 614–632. [[CrossRef](#)]
59. Folguera, L.; Zupan, J.; Cicerone, D.; Magallanes, J.F. Self-organizing maps for imputation of missing data in incomplete data matrices. *Chemom. Intell. Lab. Syst.* **2015**, *143*, 146–151. [[CrossRef](#)]
60. Jo, T.; Japkowicz, N. Class imbalances versus small disjuncts. *ACM Sigkdd Explor. Newsl.* **2004**, *6*, 40–49. [[CrossRef](#)]
61. Rahman, M.M.; Davis, D. Cluster based under-sampling for unbalanced cardiovascular data. In Proceedings of the World Congress on Engineering, San Francisco, CA, USA, 23–25 October 2013; Volume 3, pp. 3–5.
62. He, H.; Garcia, E.A. Learning from imbalanced data. *IEEE Trans. Knowl. Data Eng.* **2009**, *21*, 1263–1284.
63. Liu, G.; Yang, Y.; Li, B. Fuzzy rule-based oversampling technique for imbalanced and incomplete data learning. *Knowl. Based Syst.* **2018**, *158*, 154–174. [[CrossRef](#)]
64. Barua, S.; Islam, M.M.; Yao, X.; Murase, K. MWMOTE—majority weighted minority oversampling technique for imbalanced data set learning. *IEEE Trans. Knowl. Data Eng.* **2012**, *26*, 405–425. [[CrossRef](#)]
65. Zhang, H.; Li, M. RWO-Sampling: A random walk over-sampling approach to imbalanced data classification. *Inf. Fusion* **2014**, *20*, 99–116. [[CrossRef](#)]
66. Das, B.; Krishnan, N.C.; Cook, D.J. RACOG and wRACOG: Two probabilistic oversampling techniques. *IEEE Trans. Knowl. Data Eng.* **2014**, *27*, 222–234. [[CrossRef](#)]
67. Ksieniewicz, P. Standard Decision Boundary in a Support-Domain of Fuzzy Classifier Prediction for the Task of Imbalanced Data Classification. In Proceedings of the International Conference on Computational Science, Amsterdam, The Netherlands, 3–5 June 2020; Springer: Berlin/Heidelberg, Germany, 2020; pp. 103–116.
68. Kuncheva, L.; Bezdek, J.C.; Sutton, M.A. On combining multiple classifiers by fuzzy templates. In Proceedings of the 1998 Conference of the North American Fuzzy Information Processing Society-NAFIPS (Cat. No. 98TH8353), Pensacola Beach, FL, USA, 20–21 August 1998; pp. 193–197.
69. Ren, R.; Yang, Y.; Sun, L. Oversampling technique based on fuzzy representativeness difference for classifying imbalanced data. *Appl. Intell.* **2020**, *50*, 2465–2487. [[CrossRef](#)]
70. Mahalanobis, P.C. *On the Generalized Distance in Statistics*; National Institute of Science of India: Bengaluru, India, 1936.
71. Tang, B.; He, H. GIR-based ensemble sampling approaches for imbalanced learning. *Pattern Recognit.* **2017**, *71*, 306–319. [[CrossRef](#)]
72. Kaur, P.; Gosain, A. Robust hybrid data-level sampling approach to handle imbalanced data during classification. *Soft Comput.* **2020**, *24*, 15715–15732. [[CrossRef](#)]
73. Kaur, P.; Gosain, A. FF-SMOTE: A metaheuristic approach to combat class imbalance in binary classification. *Appl. Artif. Intell.* **2019**, *33*, 420–439. [[CrossRef](#)]

74. Tang, S.; Chen, S.P. The generation mechanism of synthetic minority class examples. In Proceedings of the 2008 International Conference on Information Technology and Applications in Biomedicine, Shenzhen, China, 30–31 May 2008; pp. 444–447.
75. Feng, L.; Qiu, M.H.; Wang, Y.X.; Xiang, Q.L.; Yang, Y.F.; Liu, K. A fast divisive clustering algorithm using an improved discrete particle swarm optimizer. *Pattern Recognit. Lett.* **2010**, *31*, 1216–1225. [[CrossRef](#)]
76. Han, H.; Wang, W.Y.; Mao, B.H. Borderline-SMOTE: A new over-sampling method in imbalanced data sets learning. In Proceedings of the International Conference on Intelligent Computing, Hefei, China, 23–26 August 2005; Springer: Berlin/Heidelberg, Germany, 2005; pp. 878–887.
77. Batista, G.E.; Prati, R.C.; Monard, M.C. A study of the behavior of several methods for balancing machine learning training data. *ACM SIGKDD Explor. Newsl.* **2004**, *6*, 20–29. [[CrossRef](#)]
78. Bunkhumpornpat, C.; Sinapiromsaran, K.; Lursinsap, C. Safe-level-smote: Safe-level-synthetic minority over-sampling technique for handling the class imbalanced problem. In Proceedings of the Pacific-Asia Conference on Knowledge Discovery and Data Mining, Bangkok, Thailand, 27–30 April 2009; Springer: Berlin/Heidelberg, Germany, 2009; pp. 475–482.
79. Stefanowski, J.; Wilk, S. Selective pre-processing of imbalanced data for improving classification performance. In Proceedings of the International Conference on Data Warehousing and Knowledge Discovery, Turin, Italy, 1–5 September 2008; Springer: Berlin/Heidelberg, Germany, 2008; pp. 283–292.
80. Hart, P. The condensed nearest neighbor rule (Corresp.). *IEEE Trans. Inf. Theory* **1968**, *14*, 515–516. [[CrossRef](#)]
81. Tomek, I. Two modifications of CNN. *IEEE Trans. Syst. Man, Cybern. Syst.* **1976**, *6*, 769–772.
82. Yoon, K.; Kwek, S. An unsupervised learning approach to resolving the data imbalanced issue in supervised learning problems in functional genomics. In Proceedings of the Fifth International Conference on Hybrid Intelligent Systems (HIS'05), Rio de Janeiro, Brazil, 6–9 November 2005; pp. 6–11
83. Laurikkala, J. Improving identification of difficult small classes by balancing class distribution. In Proceedings of the Conference on Artificial Intelligence in Medicine in Europe, Cascais, Portugal, 1–5 July 2001; Springer: Berlin/Heidelberg, Germany, 2001; pp. 63–66.
84. Kubat, M.; Matwin, S. Addressing the curse of imbalanced training sets: One-sided selection. In Proceedings of the Fourteenth International Conference on Machine Learning, Nashville, TN, USA, 8–12 July 1997; Citeseer: Pittsburgh, PA, USA 1997; Volume 97, pp. 179–186.
85. Seiffert, C.; Khoshgoftaar, T.M.; Van Hulse, J.; Napolitano, A. RUSBoost: A hybrid approach to alleviating class imbalance. *IEEE Trans. Syst. Man Cybern. Part A Syst. Hum.* **2009**, *40*, 185–197. [[CrossRef](#)]
86. Yen, S.J.; Lee, Y.S. Cluster-based under-sampling approaches for imbalanced data distributions. *Expert Syst. Appl.* **2009**, *36*, 5718–5727. [[CrossRef](#)]
87. Ramentol, E.; Caballero, Y.; Bello, R.; Herrera, F. SMOTE-RS B\*: A hybrid preprocessing approach based on oversampling and undersampling for high imbalanced data-sets using SMOTE and rough sets theory. *Knowl. Inf. Syst.* **2012**, *33*, 245–265. [[CrossRef](#)]
88. Kuncheva, L.I. *Combining Pattern Classifiers: Methods and Algorithms*; John Wiley & Sons: Hoboken, NJ, USA, 2014.
89. Ho, T.K. The random subspace method for constructing decision forests. *IEEE Trans. Pattern Anal. Mach. Intell.* **1998**, *20*, 832–844.
90. Fan, H.; Ramamohanarao, K. Fast discovery and the generalization of strong jumping emerging patterns for building compact and accurate classifiers. *IEEE Trans. Knowl. Data Eng.* **2006**, *18*, 721–737. [[CrossRef](#)]
91. Buscema, M.; Consonni, V.; Ballabio, D.; Mauri, A.; Massini, G.; Breda, M.; Todeschini, R. K-CM: A new artificial neural network. Application to supervised pattern recognition. *Chemom. Intell. Lab. Syst.* **2014**, *138*, 110–119. [[CrossRef](#)]
92. Buscema, M.; Grossi, E. The semantic connectivity map: An adapting self-organising knowledge discovery method in data bases. Experience in gastro-oesophageal reflux disease. *Int. J. Data Min. Bioinform.* **2008**, *2*, 362–404. [[CrossRef](#)] [[PubMed](#)]
93. Breiman, L. Random forests. *Mach. Learn.* **2001**, *45*, 5–32. [[CrossRef](#)]
94. Stähle, L.; Wold, S. Partial least squares analysis with cross-validation for the two-class problem: A Monte Carlo study. *J. Chemom.* **1987**, *1*, 185–196. [[CrossRef](#)]
95. Collobert, R.; Bengio, S. Links between perceptrons, MLPs and SVMs. In Proceedings of the Twenty-First International Conference on Machine Learning, Banff, AB, Canada, 4–8 July 2004; p. 23.
96. Breiman, L.; Friedman, J.; Stone, C.J.; Olshen, R.A. *Classification and Regression Trees*; CRC Press: Boca Raton, FL, USA, 1984.
97. Hosmer, D.W.; Lemeshow, S. *Applied Logistic Regression*; Wiley: New York, NY, USA, 2000.
98. McLachlan, G.J. *Discriminant Analysis and Statistical Pattern Recognition*; John Wiley & Sons: Hoboken, NJ, USA, 2004; Volume 544.
99. Platt, J.C. Fast training of support vector machines using sequential minimal optimization. In *Advances in Kernel Methods-Support Vector Learning*; Schoelkopf, B., Burges, C., Smola, A., Eds.; MIT Press: Cambridge, MA, USA, 1998
100. Fernández, A.; Herrera, F. Evolutionary Fuzzy Systems: A Case Study in Imbalanced Classification. In *Fuzzy Logic and Information Fusion*; Springer: Berlin/Heidelberg, Germany, 2016; pp. 169–200.
101. López, V.; Fernández, A.; Del Jesus, M.J.; Herrera, F. A hierarchical genetic fuzzy system based on genetic programming for addressing classification with highly imbalanced and borderline data-sets. *Knowl. Based Syst.* **2013**, *38*, 85–104. [[CrossRef](#)]
102. Fan, Q.; Wang, Z.; Li, D.; Gao, D.; Zha, H. Entropy-based fuzzy support vector machine for imbalanced datasets. *Knowl. Based Syst.* **2017**, *115*, 87–99. [[CrossRef](#)]
103. Hsu, C.W.; Lin, C.J. A comparison of methods for multiclass support vector machines. *IEEE Trans. Neural Netw.* **2002**, *13*, 415–425.

104. Platt, J.C. Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. *Adv. Large Margin Classif.* **1999**, *10*, 61–74.
105. Liu, X.Y.; Wu, J.; Zhou, Z.H. Exploratory undersampling for class-imbalance learning. *IEEE Trans. Syst. Man Cybern. Part B (Cybern.)* **2008**, *39*, 539–550.
106. Galar, M.; Fernandez, A.; Barrenechea, E.; Bustince, H.; Herrera, F. A review on ensembles for the class imbalance problem: Bagging-, boosting-, and hybrid-based approaches. *IEEE Trans. Syst. Man. Cybern. Part C Appl. Rev.* **2011**, *42*, 463–484. [[CrossRef](#)]
107. Zhang, X.; Li, Y.; Kotagiri, R.; Wu, L.; Tari, Z.; Cheriet, M. KRNN: K Rare-class Nearest Neighbour classification. *Pattern Recognit.* **2017**, *62*, 33–44. [[CrossRef](#)]
108. Zhu, C.; Wang, Z. Entropy-based matrix learning machine for imbalanced data sets. *Pattern Recognit. Lett.* **2017**, *88*, 72–80. [[CrossRef](#)]
109. Chen, S.; Wang, Z.; Tian, Y. Matrix-pattern-oriented Ho–Kashyap classifier with regularization learning. *Pattern Recognit.* **2007**, *40*, 1533–1543. [[CrossRef](#)]
110. He, H.; Bai, Y.; Garcia, E.A.; Li, S. ADASYN: Adaptive synthetic sampling approach for imbalanced learning. In Proceedings of the 2008 IEEE International Joint Conference on Neural Networks (IEEE World Congress on Computational Intelligence), Hong Kong, China, 1–8 June 2008; pp. 1322–1328.
111. Barua, S.; Islam, M.M.; Murase, K. A novel synthetic minority oversampling technique for imbalanced data set learning. In Proceedings of the International Conference on Neural Information Processing, Shanghai, China, 13–17 November 2011; Springer: Berlin/Heidelberg, Germany, 2011; pp. 735–744.
112. Viola, P.; Jones, M. Rapid object detection using a boosted cascade of simple features. In Proceedings of the 2001 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, CVPR 2001, Kauai, HI, USA, 8–14 December 2001; Volume 1, pp. 511–518
113. Wang, Y.; Wang, S.; Lai, K.K. A new fuzzy support vector machine to evaluate credit risk. *IEEE Trans. Fuzzy Syst.* **2005**, *13*, 820–831. [[CrossRef](#)]
114. Batuwita, R.; Palade, V. FSVM-CIL: Fuzzy support vector machines for class imbalance learning. *IEEE Trans. Fuzzy Syst.* **2010**, *18*, 558–571. [[CrossRef](#)]
115. Pruengkarn, R.; Wong, K.W.; Fung, C.C. Imbalanced data classification using complementary fuzzy support vector machine techniques and smote. In Proceedings of the 2017 IEEE International Conference on Systems, Man, and Cybernetics (SMC), Banff, AB, Canada, 5–8 October 2017; pp. 978–983.
116. Tu, J.V. Advantages and disadvantages of using artificial neural networks versus logistic regression for predicting medical outcomes. *J. Clin. Epidemiol.* **1996**, *49*, 1225–1231. [[CrossRef](#)]
117. Lee, H.K.; Kim, S.B. An overlap-sensitive margin classifier for imbalanced and overlapping data. *Expert Syst. Appl.* **2018**, *98*, 72–83. [[CrossRef](#)]
118. Akbani, R.; Kwek, S.; Japkowicz, N. Applying support vector machines to imbalanced datasets. In Proceedings of the European Conference on Machine Learning, Pisa, Italy, 20–24 September 2004; Springer: Berlin/Heidelberg, Germany, 2004; pp. 39–50.
119. García, V.; Mollineda, R.A.; Sánchez, J.S. On the k-NN performance in a challenging scenario of imbalance and overlapping. *Pattern Anal. Appl.* **2008**, *11*, 269–280. [[CrossRef](#)]
120. Wang, B.X.; Japkowicz, N. Boosting support vector machines for imbalanced data sets. *Knowl. Inf. Syst.* **2010**, *25*, 1–20. [[CrossRef](#)]
121. Gupta, D.; Richhariya, B. Entropy based fuzzy least squares twin support vector machine for class imbalance learning. *Appl. Intell.* **2018**, *48*, 4212–4231. [[CrossRef](#)]
122. Shao, Y.H.; Chen, W.J.; Zhang, J.J.; Wang, Z.; Deng, N.Y. An efficient weighted Lagrangian twin support vector machine for imbalanced data classification. *Pattern Recognit.* **2014**, *47*, 3158–3167. [[CrossRef](#)]
123. Chen, S.G.; Wu, X.J. A new fuzzy twin support vector machine for pattern classification. *Int. J. Mach. Learn. Cybern.* **2018**, *9*, 1553–1564. [[CrossRef](#)]
124. Arafat, M.Y.; Hoque, S.; Xu, S.; Farid, D.M. An Under-Sampling Method with Support Vectors in Multi-class Imbalanced Data Classification. In Proceedings of the 2019 13th International Conference on Software, Knowledge, Information Management and Applications, Island of Ulkulhas, Maldives, 26–28 August 2019; pp. 1–6.
125. Cristianini, N.; Shawe-Taylor, J. *An Introduction to Support Vector Machines and Other Kernel-Based Learning Methods*; Cambridge University Press: Cambridge, UK, 2000.
126. Farid, D.M.; Zhang, L.; Rahman, C.M.; Hossain, M.A.; Strachan, R. Hybrid decision tree and naïve Bayes classifiers for multi-class classification tasks. *Expert Syst. Appl.* **2014**, *41*, 1937–1946. [[CrossRef](#)]
127. Arafat, M.Y.; Hoque, S.; Farid, D.M. Cluster-based under-sampling with random forest for multi-class imbalanced classification. In Proceedings of the 2017 11th International Conference on Software, Knowledge, Information Management and Applications, Malabe, Sri Lanka, 6–8 December 2017; pp. 1–6.
128. Schapire, R.E. Explaining adaboost. In *Empirical Inference*; Springer: Berlin/Heidelberg, Germany, 2013; pp. 37–52.
129. Liu, R.; Wang, F.; He, M.; Jiao, L. An adjustable fuzzy classification algorithm using an improved multi-objective genetic strategy based on decomposition for imbalance dataset. *Knowl. Inf. Syst.* **2019**, *61*, 1583–1605. [[CrossRef](#)]
130. Ducange, P.; Lazzarini, B.; Marcelloni, F. Multi-objective genetic fuzzy classifiers for imbalanced and cost-sensitive datasets. *Soft Comput.* **2010**, *14*, 713–728. [[CrossRef](#)]

131. Xu, L.; Chow, M.Y.; Taylor, L.S. Power distribution fault cause identification with imbalanced data using the data mining-based fuzzy classification E-algorithm. *IEEE Trans. Power Syst.* **2007**, *22*, 164–171. [[CrossRef](#)]
132. Cho, P.; Lee, M.; Chang, W. Instance-based entropy fuzzy support vector machine for imbalanced data. *Pattern Anal. Appl.* **2020**, *23*, 1183–1202. [[CrossRef](#)]
133. Freund, Y.; Schapire, R.E. Experiments with a new boosting algorithm. In Proceedings of the Thirteenth International Conference on Machine Learning, Bari, Italy, 3–6 July 1996; Volume 96, pp. 148–156.
134. Zong, W.; Huang, G.B.; Chen, Y. Weighted extreme learning machine for imbalance learning. *Neurocomputing* **2013**, *101*, 229–242. [[CrossRef](#)]
135. Sakr, N.A.; Abu-Elkheir, M.; Atwan, A.; Soliman, H. A multilabel classification approach for complex human activities using a combination of emerging patterns and fuzzy sets. *Int. J. Electr. Comput. Eng.* **2019**, *9*, 2993–3001. [[CrossRef](#)]
136. Modayil, J.; Bai, T.; Kautz, H. Improving the recognition of interleaved activities. In Proceedings of the 10th International Conference on Ubiquitous Computing, Seoul, Korea, 21–24 September 2008; pp. 40–43.
137. Patel, H.; Thakur, G. An improved fuzzy k-nearest neighbor algorithm for imbalanced data using adaptive approach. *IETE J. Res.* **2019**, *65*, 780–789. [[CrossRef](#)]
138. Tan, S. Neighbor-weighted k-nearest neighbor for unbalanced text corpus. *Expert Syst. Appl.* **2005**, *28*, 667–671. [[CrossRef](#)]
139. Patel, H.; Thakur, G. A hybrid weighted nearest neighbor approach to mine imbalanced data. In Proceedings of the International Conference on Data Mining (DMIN), The Steering Committee of The World Congress in Computer Science, Las Vegas, NV, USA, 25–28 July 2016; p. 106.
140. Patel, H.; Thakur, G.S. Classification of imbalanced data using a modified fuzzy-neighbor weighted approach. *Int. J. Intell. Eng. Syst.* **2017**, *10*, 56–64.
141. García-Vico, Á.M.; González, P.; Carmona, C.J.; del Jesus, M.J. Study on the use of different quality measures within a multi-objective evolutionary algorithm approach for emerging pattern mining in big data environments. *Big Data Anal.* **2019**, *4*, 1–15. [[CrossRef](#)]
142. García-Borroto, M.; Loyola-González, O.; Martínez-Trinidad, J.F.; Carrasco-Ochoa, J.A. Evaluation of quality measures for contrast patterns by using unseen objects. *Expert Syst. Appl.* **2017**, *83*, 104–113. [[CrossRef](#)]
143. García-Vico, A.M.; González, P.; del Jesus, M.J.; Carmona, C.J. A first approach to handle fuzzy emerging patterns mining on big data problems: The EvAEFP-spark algorithm. In Proceedings of the 2017 IEEE International Conference on Fuzzy Systems (FUZZ-IEEE), Naples, Italy, 9–12 July 2017; pp. 1–6. [[CrossRef](#)]
144. Luna, J.M.; Carmona, C.J.; García-Vico, A.M.; del Jesus, M.J.; Ventura, S. Subgroup Discovery on Multiple Instance Data. *Int. J. Comput. Intell. Syst.* **2019**, *12*, 1602–1612. [[CrossRef](#)]
145. Atzmueller, M.; Puppe, F. SD-Map-A fast algorithm for exhaustive subgroup discovery. In Proceedings of the European Conference on Principles of Data Mining and Knowledge Discovery, Freiburg, Germany, 3–5 September 2006; Springer: Berlin/Heidelberg, Germany, 2006; pp. 6–17.
146. Luna, J.M.; Romero, J.R.; Romero, C.; Ventura, S. On the Use of Genetic Programming for Mining Comprehensible Rules in Subgroup Discovery. *IEEE Trans. Cybern.* **2014**, *44*, 2329–2341. [[CrossRef](#)]
147. Carmona, C.J.; González, P.; del Jesus, M.J.; Herrera, F. NMEEF-SD: Non-dominated Multiobjective Evolutionary Algorithm for Extracting Fuzzy Rules in Subgroup Discovery. *IEEE Trans. Fuzzy Syst.* **2010**, *18*, 958–970. [[CrossRef](#)]
148. Garcia-Vico, A.M.; Chartea, F.; González, P.; Elizondob, D.; Carmona, C.J. E2PAMEA: A fast evolutionary algorithm for extracting fuzzy emerging patterns in big data environments. *Neurocomputing* **2020**, *415*, 60–73. [[CrossRef](#)]
149. Zhang, Y.; Zhou, Z.H. Cost-sensitive face recognition. *IEEE Trans. Pattern Anal. Mach. Intell.* **2009**, *32*, 1758–1769. [[CrossRef](#)]
150. Tang, Y.; Zhang, Y.Q.; Chawla, N.V.; Krasser, S. SVMs modeling for highly imbalanced classification. *IEEE Trans. Syst. Man. Cybern. Part B (Cybern.)* **2008**, *39*, 281–288. [[CrossRef](#)] [[PubMed](#)]
151. Gu, X.; Angelov, P.P.; Soares, E.A. A self-adaptive synthetic over-sampling technique for imbalanced classification. *Int. J. Intell. Syst.* **2020**, *35*, 923–943. [[CrossRef](#)]
152. Gu, X.; Angelov, P.; Rong, H.J. Local optimality of self-organising neuro-fuzzy inference systems. *Inf. Sci.* **2019**, *503*, 351–380. [[CrossRef](#)]
153. Cunningham, P.; Delany, S.J. k-Nearest Neighbour Classifiers: 2nd Edition (with Python examples). *arXiv* **2020**, arXiv:2004.04523.
154. Garcia-Vico, A.M.; Carmona, C.J.; Gonzalez, P.; del Jesus, M.J. A Preliminary Many Objective Approach for Extracting Fuzzy Emerging Patterns. In Proceedings of the 15th International Conference on Soft Computing Models in Industrial and Environmental Applications (SOCO 2020), Burgos, Spain, 16–18 September 2020; Herrero, Á., Cambra, C., Urda, D., Sedano, J., Quintián, H., Corchado, E., Eds.; Springer International Publishing: Cham, Switzerland, 2021; pp. 100–110.
155. Deb, K.; Jain, H. An Evolutionary Many-Objective Optimization Algorithm Using Reference-Point-Based Nondominated Sorting Approach, Part I: Solving Problems With Box Constraints. *IEEE Trans. Evol. Comput.* **2014**, *18*, 577–601. [[CrossRef](#)]
156. Deb, K.; Sundar, J. Reference point based multi-objective optimization using evolutionary algorithms. In Proceedings of the 8th Annual Conference on Genetic and Evolutionary Computation, Seattle, WA, USA, 8–11 July 2006; pp. 635–642.
157. Schaefer, G. Strategies for imbalanced pattern classification for digital pathology. In Proceedings of the 2017 6th International Conference on Informatics, Electronics and Vision & 2017 7th International Symposium in Computational Medical and Health Technology, Himeji, Japan, 1–3 September 2017; pp. 1–4.

158. Jaafar, H.; Ramli, N.H.; Nasir, A.S.A. *An Improvement to The k-Nearest Neighbor Classifier for ECG Database*; IOP Conference Series: Materials Science and Engineering; IOP Publishing: Bristol, UK, 2018; Volume 318, pp. 1–10
159. Polat, K. Similarity-based attribute weighting methods via clustering algorithms in the classification of imbalanced medical datasets. *Neural Comput. Appl.* **2018**, *30*, 987–1013. [[CrossRef](#)]
160. Cho, P.; Chang, W.; Song, J.W. Application of instance-based entropy fuzzy support vector machine in peer-to-peer lending investment decision. *IEEE Access* **2019**, *7*, 16925–16939. [[CrossRef](#)]
161. Xia, Y.; Liu, C.; Liu, N. Cost-sensitive boosted tree for loan evaluation in peer-to-peer lending. *Electron. Commer. Res. Appl.* **2017**, *24*, 30–49. [[CrossRef](#)]
162. Li, F.Q.; Wang, S.L.; Liu, G.S. A Bayesian Possibilistic C-Means clustering approach for cervical cancer screening. *Inf. Sci.* **2019**, *501*, 495–510. [[CrossRef](#)]
163. Grzymala-Busse, J.W.; Hu, M. A comparison of several approaches to missing attribute values in data mining. In Proceedings of the International Conference on Rough Sets and Current Trends in Computing, Banff, AB, Canada, 16–19 October 2000; Springer: Berlin/Heidelberg, Germany, 2000; pp. 378–385.
164. Jassim, F.A. Image Denoising Using Interquartile Range Filter with Local Averaging. *arXiv* **2013**, arXiv:1302.1007.
165. Jain, D.; Singh, V. A two-phase hybrid approach using feature selection and Adaptive SVM for chronic disease classification. *Int. J. Comput. Appl.* **2019**, *2*, 1–13 [[CrossRef](#)]
166. García-Vico, Á.M.; Carmona, C.J.; González, P.; Seker, H.; del Jesus, M.J. FEPDS: A Proposal for the Extraction of Fuzzy Emerging Patterns in Data Streams. *IEEE Trans. Fuzzy Syst.* **2020**, *28*, 3193–3203. [[CrossRef](#)]

### Short Biography of Authors



**Ismael Lin** obtained his bachelor's degree in Electrical Engineering from Tecnológico de Monterrey, and his master's degree in Robotics from KTH Royal Institute of Technology. He is a Ph.D. student in Computer Science at Tecnológico de Monterrey, Campus Estado de Mexico, in the Machine Learning research group. His current research focuses on supervised learning, class imbalance problem, and pattern-based classification.



**Octavio Loyola-González** received his PhD degree in Computer Science from the National Institute for Astrophysics, Optics, and Electronics, Mexico, in 2017. He has won several awards from different institutions due to his research work on applied projects; consequently, he is a Member of the National System of Researchers in Mexico (Rank1). He worked as a distinguished professor and researcher at Tecnológico de Monterrey, Campus Puebla, for undergraduate and graduate programs of Computer Sciences. Currently, he is responsible for running Machine Learning & Artificial Intelligence practice inside Altair Management Consultants Corp., where he is involved in the development and implementation using analytics and data mining in the Altair Compass department. He has outstanding experience in the fields of big data & pattern recognition, cloud computing, IoT, and analytical tools to apply them in sectors where he has worked for as Banking & Insurance, Retail, Oil&Gas, Agriculture, Cybersecurity, Biotechnology, and Dactyloscopy. From these applied projects, Dr. Loyola-González has published several books and papers in well-known journals, and he has several ongoing patents as manager and researcher in Altair Compass.



**Raúl Monroy** obtained a Ph.D. degree in Artificial Intelligence from Edinburgh University, in 1998, under the supervision of Prof. Alan Bundy. He has been in Computing at Tecnológico de Monterrey, Campus Estado de México, since 1985. In 2010, he was promoted to (full) Professor in Computer Science. Since 1998, he is a member of the CONACYT-SNI National Research System, rank three. Together with his students and members of his group, Machine Learning Models (GIEE – MAC), Prof. Monroy studies the discovery and application of novel model machine learning models, which he often applies to cybersecurity problems. At Tecnológico de Monterrey, he is also Head of the graduate programme in computing, at region CDMX.



**Miguel Angel Medina-Pérez** received a Ph.D. in Computer Science from the National Institute of Astrophysics, Optics, and Electronics, Mexico, in 2014. He is currently a Research Professor with the Tecnológico de Monterrey, Campus Estado de Mexico, where he is also a member of the GIEE-ML (Machine Learning) Research Group. He has rank 1 in the Mexican Research System. His research interests include Pattern Recognition, Data Visualization, Explainable Artificial Intelligence, Fingerprint Recognition, and Palmprint Recognition. He has published tens of papers in referenced journals, such as "Information Fusion," "IEEE Transactions on Affective Computing," "Pattern Recognition," "IEEE Transactions on Information Forensics and Security," "Knowledge-Based Systems," "Information Sciences," and "Expert Systems with Applications." He has extensive experience developing software to solve Pattern Recognition problems. A successful example is a fingerprint and palmprint recognition framework which has more than 1.3 million visits and 135 thousand downloads.