Survey on Highly Imbalanced Multi-class Data

Mohd Hakim Abdul Hamid[1]
INSFORNET, C-ACT and Fakulti Teknologi Maklumat dan Komunikasi (FTMK), Universiti Teknikal Malaysia Melaka (UTeM),, Hang Tuah Jaya, Melaka, Malaysia

Marina Yusoff[2], Azlinah Mohamed[3]
Institute for Big Data Analytics and Artificial Intelligence
Universiti Teknologi MARA (UiTM)
Shah Alam, Selangor, Malaysia

***Abstract*—Machine learning technology has a massive impact on society because it offers solutions to solve many complicated problems like classification, clustering analysis, and predictions, especially during the COVID-19 pandemic. Data distribution in machine learning has been an essential aspect in providing unbiased solutions. From the earliest literatures published on highly imbalanced data until recently, machine learning research has focused mostly on binary classification data problems. Research on highly imbalanced multi-class data is still greatly unexplored when the need for better analysis and predictions in handling Big Data is required. This study focuses on reviews related to the models or techniques in handling highly imbalanced multi-class data, along with their strengths and weaknesses and related domains. Furthermore, the paper uses the statistical method to explore a case study with a severely imbalanced dataset. This article aims to (1) understand the trend of highly imbalanced multi-class data through analysis of related literatures; (2) analyze the previous and current methods of handling highly imbalanced multi-class data; (3) construct a framework of highly imbalanced multi-class data. The chosen highly imbalanced multi-class dataset analysis will also be performed and adapted to the current methods or techniques in machine learning, followed by discussions on open challenges and the future direction of highly imbalanced multi-class data. Finally, for highly imbalanced multi-class data, this paper presents a novel framework. We hope this research can provide insights on the potential development of better methods or techniques to handle and manipulate highly imbalanced multi-class data.**

Keywords*—*Imbalanced data; highly imbalanced data; highly imbalanced multi-class; data strategies

I. Introduction

Every single piece of information in this world is data. The nature of data is that it is not absolutely balanced [1], [2]. It is either slightly imbalanced or highly imbalanced. Highly imbalanced is a situation where the ratio among classes is elevated. For example, 80:20 or 90:10 or 95:5 (majority: minority). While the example of slightly imbalanced is 60:40 or 55:45 or 70:30 (majority: minority). According to Bellinger et al., the imbalance ratio (IR) for highly imbalanced is 1000:1 (majority: minority) [3]. A well-balanced dataset is only possible in a controlled environment in which variables are preset, and data undergoes proper preprocessing.

The advancement of machine learning (ML) facilitates and alleviates problems in data analysis. The Internet of Things (IoT) and big data technology have accelerated the utilization of big data. During the Covid-19 pandemic, ML analysis helps in many ways like in vaccine development in which the cure is required to restrain the virus spread concisely. With advanced computers, machine learning algorithms can process trillions of data within a short period of time and make accurate predictions, or classification of the problem.

To solve imbalanced data problems, researchers employ three strategies: 1) Data Level or DL 2) Algorithm Level or AL, and 3) Combination Level or CL. DL strategy involves data manipulation activities during the pre-processing phase. Most of the DL strategies involve approaches on data before it enters classifiers. Examples of such methods are undersampling and oversampling and their variants. AL strategy involves classifier-related activities. The main objective of the AL strategy is to apply algorithms or classifiers to manage dataset and handle imbalanced data problems. There are a lot of examples for AL such as Deep Learning variants, and Support Vector Machine (SVM) and their variants. CL strategy is a combination of both DL and AL strategy. This strategy applies to both hybrid and ensemble methods, even though ensemble can also work alone in AL. Ensemble is listed under CL because in most literatures about highly imbalanced data, the ensemble algorithm is combined with some other method(s) to achieve better performance. Some literatures only mention hybrid level, without the ensemble method as part of CL strategy [4], [5]. In some other literature, the ensemble method is mentioned along with DL and AL without the hybrid method [6].

The ensemble method is an interesting area of research. Ensemble algorithms can apply ensemble algorithm to other algorithms, either supervised, semi-supervised or unsupervised (or combination of all) to produce a better algorithm. The hybrid method is also a very interesting method in ML. It combines any method in DL with AL, combines DL method with another DL algorithm(s), or combines AL with another AL method(s). Therefore, this study proposes a new term "Combination Level" for both the hybrid and ensemble methods in highly imbalanced multi-class (HIMC) data. This article then will propose a novel framework for highly imbalanced multi-class data (HIMC). This is crucial so that the structure of the framework and the relevance to future of ML can be explored. It is hoped that this research will pave the road for other researchers to excavate deeper into the nature of HIMC data and various ways to handle it.

This paper is organized into several sections. Section I presents the definition of imbalanced data, highly imbalanced data, multi-class data and highly imbalanced multi-class data. Section II provides the explanations on the research gap, research questions and research objectives. Section III presents the descriptions on the current and previous solutions in handling imbalanced data, previous solutions in handling multi-class data and previous solutions in handling HIMC data. Section IV provides the descriptions on the prominent validation metrics for HIMC data, data type and dataset behavior along with discussion on the case study dataset. Finally, Section V presents the explanations on the proposed and refined framework of HIMC data and will be followed by Section VI, discussions on open issues. Finally, Section VII provides the conclusion.

A. Imbalanced Data

Imbalanced data is currently a prominent research topic and acts as a relatively new research interest in machine learning [1], [7]. A circumstance in which the total number of the majority class is significantly greater than the data of the minority class is known as data imbalanced [8]. Most classifiers are designed to work with a balanced dataset in which the majority class is comparable to or equal to the minority class, or the ratio between the two classes is 50:50. A balanced environment is essential to make sure the classifier can perform at the best level of accuracy. Therefore, when unequal data exists, an imbalance data problem transpires [9], [10]. Imbalanced data also occurs because of the existence of minority classes that are lowly represented [11]. It also can happen when the dataset is skewed [12]. Most classifiers in a balanced class are biased toward the majority class [13], [14], [15]. In real life, all real-world data is imbalanced [16], [17]. Real-world data can have a high chance to fall in the category of highly imbalanced data [18]–[20] or slightly imbalanced data [21].

Imbalanced data exists in numerous disciplines for example, wind turbine fault prediction [11], network diagnosis, wireless sensor application [22], acid amino detection [23], medical diagnosis [24], Internet of Things [25], fraud detection [26], and other domains. Many approaches to overcoming the problem have been offered with regard to data imbalanced problem using different solutions from DL to AL and CL strategy, such as those found in [5], [8], [11], [26]–[29].

B. Highly Imbalanced Data

To deal with the problem of data imbalance, a variety of approaches and methodologies have been proposed. DL and AL strategies focus on ways to reduce biases of classes in a dataset. CL strategy involves an ensemble or a hybrid of several algorithms to achieve the best results. Nevertheless, the problem with imbalance data is far from settled, especially in highly imbalanced data scenarios. The problem with highly imbalanced data is that the ratio is extremely high. The solution that works in slightly imbalanced data might not work in highly imbalanced data. Therefore, highly imbalanced data needs more consideration and investigation. Normal graph distribution does not present in highly imbalanced data as the graph is highly skewed.

In a slightly imbalanced data environment, a conventional method such as Synthetic Minority Over Sampling Technique (SMOTE) can help solve the problem of imbalanced data. Unfortunately, DL method like SMOTE can worsen the classification performance in highly imbalanced data environment. Using randomized methods like Random Undersampling (RUS) is also not effective due to a high variance created in the IR. To overcome the problem of noisy data and overlapping classes, a new approach for data preprocessing is needed to boost classifiers' performance. A new approach to handle detection and filtering noisy data is needed in the scenario of relabeling classes. The problem can increase the imbalance among classes. It could result in the classifier rebalancing the wrong classes.

A highly imbalanced data problem arises when the IR is too high compared to slightly imbalanced data. For example, the ratio of minority to majority less than 50:1 can be considered slightly imbalanced while imbalance ratio (IR) more than 50:1 can be considered highly imbalanced. According to Triguero et al., IR for highly imbalanced data is 50:1 (majority: minority) [30], [31]. Another well-known IR is 100:1 up to 10000:1 [32]. The same IR (100:1) was suggested by Sharma et al. [33]. Sharma et al. suggest 100:1 as highly imbalanced while 1000:1 was categorized as extreme imbalance [33]. Table I shows benchmark of IR.

TABLE I. BENCHMARK OF IR

No.	Reference	IR	Year
1.	He & Garcia	100: 1 up 10,000:1	2009
2.	Triguero et al., Leevy et al.	50:1	2015, 2018
3.	Sharma et al.	>1000:1	2018
4.	Bellinger et al.	1000:1	2019

More practical IR was found among bioinformatics and biotechnology domains, and it was 50:1 [30], [31]. In other literatures, highly imbalanced are also known as rare events in which researchers and scholars stated that the minority data that was from 0.1% to less than 10% , can be considered as rare events [34], [35]. In the real-world, IR ranging from 1000:1 up to 5000:1 is possible in fraud detection and medical science [4], [36], [37].

This research has chosen the latest literatures as the benchmark for highly imbalanced data ratio. The latest found in literatures on IR is suggested by Bellinger et al. which is 1000:1 [3]. Therefore, for the purpose of this research, the IR stated by Bellinger et al. will be used. The ratio from Bellinger et al. has also been chosen because the dataset used in this research matches with the stated IR.

C. Highly Imbalanced Multi-Class (HIMC) Data

Problems in highly imbalanced data originate from problems in slightly imbalanced data, which are alleviated due to the nature of severe IR [3], [33]. Thus, problems in imbalanced multi-class data and highly imbalanced multi-class data can be considered similar in nature, with the difference laying in the IR.

To overcome the HIMC data issue, a new approach for data pre-processing is needed to boost classifiers' performance in HIMC data. Another solution to overcome the problem is by creating synthetic data to move overlapping data to new spaces [38]. Another method is to remove excessive samples and maintain the quality of the data [2], [6] [1], [6].

Relationship among classes is an issue in multi-class data. It is a complicated situation as each group of classes presents different problems to the data [1], [2], [4], [38], [39]. This problem is elevated in HIMC data, and affects classifiers' performance [40], [41]. Leevy et al. and Rendon et al. suggest that more flexible methods such like the heuristic-based method should be explored to solve multi-class data problems [31], [42]. The HIMC data has multiple skewed classes which reduce classifiers' performance as it is challenging to normalize skewness [31], [43], [44], [45]. Due to skewness, it is also difficult to define borders of the overlap classes [4], [46], [47].

D. HIMC Data Research Gap

Among the major challenges with highly unbalanced data are the accuracy of classifying highly imbalanced multi-class data, training efficiency for large data, and sensitivity to high imbalance ratio (IR) [48]. In highly imbalanced data, classifiers are prone to a strong bias toward the majority class, which cannot accurately represent the true problem or convey essential information. The minority class were treated as noisy data at the pre-processed level and will cause the loss of crucial information [1], [21]. This creates new challenges to data level strategy in handling biasness [49], [43], [45].

A dataset with multiple target classes is skewed in distribution in imbalanced multi-class data, and this has a substantial effect on classifier performance [38]. HIMC data has multiple skewed classes which significantly reduce classifiers' performance as it is difficult to normalize skewness [31], [43], [44] due to the difficulty to define borders of the overlap classes [4], [47].

At algorithms level, existing classifiers are modified to remove the biases toward the majority classes. One of the methods is the cost-sensitive method. The cost of misclassification for minority samples is higher than for majority samples in the cost-sensitive method. Determining the cost values of trained data is complex since they are dependent on multiple aspects that have trade-off relationships, such as high-dimension, high noise, small sample size, and others [1], [21]. In financial data, biasness causes highly imbalanced distribution [50].

Therefore, based on arguments regarding HIMC data, this research addresses three research questions and three research objectives.

The developed research questions for this study are:

1) What is the current trend in handling highly imbalanced multi-class data?

2) How to handle highly imbalanced multi-class data?

3) How to develop a framework of highly imbalanced multi-class data?

The following objectives are developed based on the research problem:

1) To understand the trend of highly imbalanced multi-class data through analysis of all related literatures.

2) To analyze the previous and current method of handling highly imbalanced multi-class data.

3) To construct a framework handling highly imbalanced multi-class data through research and literature study.

II. Strategies in Handling Imbalanced Data

Based on previous studies on HIMC data, it is imperative to understand these related sub-topics: (1) Strategies in handling imbalanced data; (2) Previous solutions in handling imbalanced data, imbalanced multi-class data and HIMC data; (3) Related method or technique used in solving HIMC data problems.

The same three strategies in handling slightly imbalanced data can be used in handling highly imbalanced multi-class data. The details of these strategies and their methods are put under Appendix 1. There are three types of DL strategies which are oversampling, undersampling and hybrid strategy. AL strategy can be divided further into four methods which are cost-sensitive learning, skewed learning function, sampling-based and other methods. The CL strategy can be divided into two main methods: hybrid and ensemble. The Hybrid method can be divided further into MTD-based, SVM-variants and other hybrid methods. While ensemble method can be divided further into four methods which are integration with data level, integration with cost-sensitive, bagging variants and boost variants.

DL strategy involves data manipulation activities during the pre-processing phase in machine learning. An example of oversampling-related method is Synthetic Minority Oversampling Technique (SMOTE) [51], while an example of undersampling related method is Random Under sampling (RUS) [52]. From the literature review conducted, several literatures related to HIMC data have been found. However, the strategy or method proposed at DL in HIMC data is hardly mentioned. Therefore, this can be a promising area for future research in HIMC data.

AL strategy involves classifier related activities such as Support Vector Machine (SVM) [53], Deep Learning method using Convolutional Neural Network model (CNN) [54], and K-Nearest Neighbor [55]. Convolutional Neural Network (CNN) model was used to predict Chlorophyll-A concentration in Algal Bloom in managing data imbalance and skewness [56], and a Deep Self-Organizing Map (DSOM) was proposed to detect a well-known pre-miRNA protein as compared to a genome's hundreds of thousands of potential sequences [18].

The CL strategy is a combination of both DL and AL strategies, or combination of DL strategy with another DL, and a combination of AL with another AL strategy. In addition, it can also be a combination of ensemble method with AL strategy, or combination of ensemble method with DL strategy, or it can be a combination of both hybrid and ensemble methods [57], [58]–[60]. For example, based on a combination of data rebalancing and Extreme Gradient Boosting (XGBoost), a unique form of malicious synchrophasors detector is developed [61].

Oversampling and under sampling were combined with SVM in solar flare prediction [7]. Fujiwara et al. proposed a heuristic undersampling and distribution-based sampling with boosting method (HUSDOS-Boost) in handling data problems in health record analysis [44]. The strategies and methods used in handling imbalance data are shown in Fig. 1.

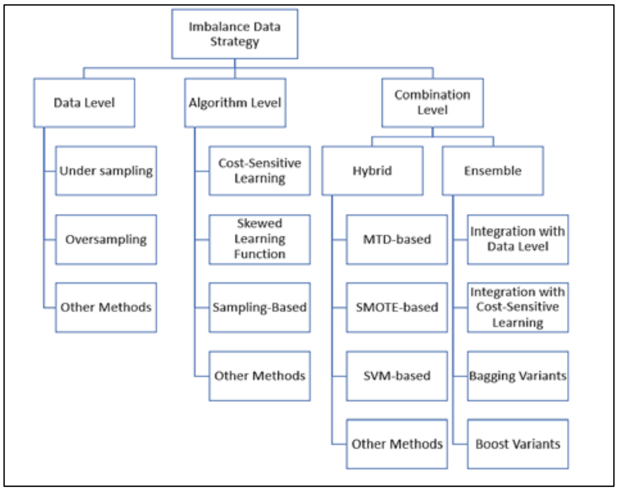


Fig. 1. List of Strategies in Handling Imbalanced Data.

Fig. 1 shows a list of strategies in handling imbalanced data. Imbalanced data can be divided into three strategies namely DL, AL, and CL strategy. Literatures related to DL and AL can be found in many studies while CL concept is still fresh. In literatures such as Kaur et al. and Johnson and Koshgoftaar, DL and AL have been mentioned along with Hybrid Method (HM). The logic behind this concept is that HM is a combination of both DL and AL [29], [42]. In other literatures such as in Sleeman & Krawczyk, DL and AL have been mentioned along with Ensemble Method (EM), while HM has not been specifically mentioned as their work was more focused on EM [6]. Some might argue on the reason to categorize EM into CL, as EM might also fall into AL strategy.

It is imperative to understand that in recent highly imbalanced data research, EM is usually not working alone and is combined with another method except for performance comparison or proposal of a new framework [62], [63]. In highly imbalanced classification, research of supervised and unsupervised fuzzy measure approaches was conducted by Uriz et al. The authors integrated EM with fuzzy integrals and their synergy with various fuzzy measures [64]. Another study was by Liu et al. where they established a unique framework for imbalance classification that was intended at building a strong ensemble by self-paced harmonizing data hardness by under-sampling, in which a classifier was combined with self-paced EM. [15].

Ghorbani et al. proposed a new hybrid model based on a highly imbalanced dataset to predict early mortality risk in intensive care units (ICU). The authors developed an SVM and SMOTE hybrid strategy (SVM-SMOTE) that included several methods, including a Genetic Algorithm (GA) for feature selection (FS) and Stacking and Boosting (EM) for prediction. SVM-SMOTE was used to tackle imbalanced data problems [65]. Using clustering, weighted scoring, SVM, and EM, Ksieniewicz et al. suggested a hybrid method for managing severely imbalanced data categorization in geometric space [66]. Using a combination of data rebalancing, bagging-based ensemble learning, and the Extreme Gradient Boosting (XGBoost) algorithm, a unique form of malicious synchrophasors detector was developed to address the highly imbalanced data problem. Even if malicious synchrophasors occur seldom in practice, a detector trained on a highly imbalanced dataset drawn straight from previous operational data is biased toward the majority class. [61]. Tran et al. proposed a combination of K-Segments, under sampling and bagging EM as experimental research to approach extreme imbalanced data classification [64]. From the examples mentioned in this study, EM was combined with other methods to achieve better performance. Appendix 1 until Appendix 4 will entail strategies and methods involved in handling imbalanced data.

III. Previous Solutions in Handling Highly Imbalanced Multi-Class Data

Ahmadzadeh et al. is one of the most current approaches for extremely unbalanced data that has been proposed [7]. The reviewed literatures discussed solar flare forecasts using under-sampling, over-sampling, and Support Vector Machine (SVM) to handle highly or extremely imbalanced solar flare data. For future development, these works suggested exploring hyperparameter tuning of the proposed method to enhance the model. A multi-class dataset was initially used but then it was converted to binary classification data to make predictions.

Fujiwara et al. published an article in the medical field. In this research, oversampling and undersampling methodologies for highly imbalanced data in health records analysis were presented. When minority samples are too tiny, undersampling and oversampling, or a mix of the two via hybrid and ensemble, did not give satisfactory results. The authors developed HUSDOS-Boost, which stands for Heuristic Under-sampling and Distribution Based Sampling paired with a Boosting ensemble, to cope with the extreme imbalanced and small minority (EISM) problem. When compared to other ensemble approaches, the result was superior. The authors proposed that a hierarchical Bayes model be used to estimate the distribution parameter in future work to improve over-sampling performance [44].

Managing cyber-attacks such as in detecting malicious synchrophasors is very important especially among energy-based companies. Performance of the detectors might be deteriorated severely due to quality of extremely imbalanced data. The authors developed a malicious synchrophasors detector based on data rebalancing, ensemble learning with bagging, and Extreme Gradient Boosting (XGBoost). The proposed method can detect malicious synchrophasors even though only a minimum number of malicious instances were provided [61]. There are several other methods or solutions proposed by different researchers involving different kinds of algorithms or solutions in different kinds of extreme imbalance dataset. Despite all the proposed methods, an extreme or highly imbalance dataset is still a challenging area to explore [44].

A. Related Method/Technique In Handling Highly Imbalanced Multi-Class Data

This section presents the related methods or techniques that have been used by researchers to handle highly imbalanced multi-class data. In the context of this study, the methods described are ensemble, deep learning, and cost-sensitive method.

1) Ensemble method: Research works on highly imbalanced data using ensemble method have become more prevalent since the emergence of Big Data technology [14]. Ensemble method along with hybrid method is steadily gaining attention from researchers around the world. From 2010 to 2021, there are many literatures regarding research works on highly imbalanced data using ensemble method [15], [61], [62], [63], [64] [65], [66], [67], [68]–[74]. The ensemble technique is popular because it combines many algorithm approaches with the ensemble algorithm in machine learning to improve performance [60], [75]–[77]. Compared to a single classifier, the ensemble's total performance was improved by combining several approaches or algorithms. Ensemble modelling is a set of models that work collaboratively to provide a more efficient predictive model. Different modelling techniques, such as Decision Tree (DT) [78]–[82], Neural Networks (NN) [83]–[86], Random Forests (RF) [87], [88]–[91], Support Vector Machines (SVM) [43], [92]–[95] and others can be integrated with ensemble.

Bagging, boosting, and stacking are examples of ensemble techniques that are used to improve the performance of a model or reduce the likelihood of selecting a bad one. There are many literatures on the performance of ensemble method. Among more prevalent methods are Bagging (Bootstrap aggregation) [6], [28] [96]–[99], [100], Stacking (Stacked Generalization) [98], [101], [102], Random Forest (RF) [87], [91], [103], [13], [104]–[109] and Mixtures of Experts [8], [42], [110], [111], [112], [113].

In 1996, Bagging or Bootstrap Aggregating was established as one of the first ensemble approaches. To reduce variance error, this technique trains and picks strong classifiers on subsets of data. Robust performance on outliers, decrease of variance to minimize over-fitting, which requires minimum further parameter adjustment, and the ability to accept high nonlinear interactions are just a few of the advantages. One of the drawbacks of bootstrap aggregation is that the more complicated the model becomes, the less visible and interpretable it becomes [114]–[117].

Boosting is like bagging, but it gives weak classifiers more weight. The weaker classifiers are given additional weight in the following classification phase with each iteration of classifications, increasing their chances of being categorized properly until a stopping point is reached. This can be thought of as course-correcting by re-energizing the data weights that require it. Over-fitting, outlier influence, revision on iteration ending point, and lack of transparency owing to complexity are some of the flaws of this technique, which optimize the cost function. [96], [118]–[120], [121].

Stacking, sometimes known as the least understood ensemble technique, produces ensembles by combining a variety of powerful classifiers. When developing ensemble models, diversity is crucial because it allows stronger learners from various regions to combine their abilities to lower the chance of misclassification. Stacking employs various levels of classification training [98], [102], [122]–[124]. Tier two (2) will use the misclassified regions to adjust the behavior in the next phase if tier one has feature spaces that are misclassified. The biggest flaw in this form of ensemble is that it lacks transparency when it comes to determining a metadata classifier that adjusts for errors to improve prediction accuracy [59], [125]–[128].

When compared to the implementation in a single classifier, the ensemble's total performance is improved by combining several approaches or algorithms. There is a lot of literature stating about the performance regarding ensemble method [1], [2], [125], [129]–[132].

An effective and popular tool for optimizing ensembles of classifiers is the genetic algorithm (GA), belonging to the family of evolutionary algorithms [133]–[135]. The inspiration to study evolutionary computation (EC) was the imitation of nature in its mechanism of natural selection, inheritance, and functioning. Evolutionary computation is used to demonstrate and unravel complex tasks, primarily for optimization. It is trained based on species, not on an entity, that extends across the lifespan of numerous generations of entities. As a result, generations that are produced progressively meet the conditions of the task, and this would improve the adjustments made to the environment.

It is also worth observing the combination of evolutionary computation with ensemble methods. Examples of such combination can be found in several areas of research like in model-based ensemble [136], micro genetic algorithm, parallel genetic algorithms [137], GA with ensemble method [138], [139], and stacking ensemble [58], [124]. The algorithms mentioned are examples of hybrid and ensemble algorithm that falls into the CL strategy.

Ensemble approaches have been widely applied across several disciplines in the domain of credit scoring and bankruptcy prediction [99], [139], [140], [141] including the latest on personal bankruptcy prediction on imbalanced dataset can be found in several literatures [79], [88], [89], [141], [142].

2) Extreme Gradient Boosting (XGBoost): The consequences of noisy data and redundant features, which contributed to the unbalanced data scenario, are mitigated by feature selection methods. In boosting approaches, such as extreme gradient boosting (XGBoost), distributed learning, and multi-core computation, which fully employ the computer's capabilities to speed learning, are possible. As a result, more investigation into boosting is strongly recommended in the highly or extreme imbalanced data research. Chen and Guestrin developed XGBoost which is an advance gradient boosting (GB) [164]. It is a fast, scalable, and efficient algorithm that won Kaggle machine learning competition and was applied in many applications and is used by many corporations.

Base classifiers are known as weak learners. In boosting, models are added concurrently until there is no further change. Boosting is an additive ensemble method that combines new models with existing models to reduce errors. Boosting is a technique for integrating many base classifiers to produce classification accuracy that is considerably better than any single base classifier's performance. A boosted model will produce a good result even if the base classifiers have a marginally better accuracy than random. XGBoost is an open-source library providing a gradient boosting platform for Python, R, C++, and Java. It employs a gradient-boosting technique to generate a prediction model in the form of an ensemble of weak prediction models, most commonly decision trees.

Gradient boosting (GB), stochastic GB, and regularized GB are the three major types of gradients boosting that XGBoost can perform. The XGBoost approach is flexible in its implementation of distributed and parallel computing and can handle sparse data [143]. It is strong enough to handle hyper parameter fine tweaking and regularization parameter addition. It's been put to the test on large-scale challenges and can handle most regression, classification, and ranking problems, as well as custom objective functions. XGBoost is also portable and compatible, allowing it to run on any operating system. It works with AWS, Azure, and GCE, as well as other distributed cloud platforms.

XGBoost is easily coupled to large-scale cloud data-flow systems like Flink and Spark, which were designed specifically for model performance and computing speed. Model tuning, computational environments, and algorithm enhancement are all available in XGBoost. The algorithm was created with the goal of reducing computation time and allocating memory resources efficiently. XGBoost improves classification accuracy and performs calculations 10 times faster than commercial software. It can prevent over-fitness and dealing with missing values. With learning, XGBoost can figure out the dividing path for the test with incomplete eigenvalues. [60].

3) Deep learning method: Research of deep learning in imbalanced data was overwhelming, however research of deep learning in highly imbalanced data still has much room for expansion. The length of the majority data or class is substantially longer than the length of the minority data component in highly imbalanced data settings. To put it another way, the minority data has essentially been ignored by the classifier and most of the time considered as a noisy data [4]. The net gradient, which is responsible for updating the classifier weights, is dominated by the majority data. During early iterations, this reduces the error of the dominant majority quickly, but it often raises the error of the minority group, trapping the algorithm in a slow convergence state. [5].

In the fields of image identification [144], speech recognition [145] and natural language processing [146], [147], deep learning has been widely utilized. However, there have been few studies on the use of deep learning in highly imbalanced data. As the RNN is ideal for time series analysis, one popular application is the deployment of a recurrent neural network (RNN) to investigate network intrusion detection on an imbalanced dataset [148].

Convolutional Neural Network (CNN) is another deep learning model that has been used to forecast bankruptcy. Hosaka et al. took financial statement data from Japanese publicly traded firms and converted the numerical financial ratio data into a grayscale image that was tailored to CNN's characteristics and could be evaluated directly by CNN. To cope with bankruptcy prediction difficulties, Hosaka suggested a CNN framework, and this model beat comparable conventional solutions, including most of the established machine learning techniques [149]. Mai et al. used layers of neural networks to extract attributes from textual data from over 10000 public corporations in the United States to incorporate deep learning into the prediction of bankruptcy. [150].

It has been revealed that when textual data (e.g., news, public company reports) is combined with classical numerical data (e.g., financial ratio data), deep learning performs better in imbalanced data study using textual disclosures, improving prediction accuracy even more. These intriguing findings open up new avenues for research in the field of bankruptcy prediction on imbalanced datasets, providing new insights and ideas. [151].

4) Cost-Sensitive learning: When training a model, cost-sensitive learning considers the costs of prediction errors as well as any additional cost that may be necessary. It is related to classification on datasets that are imbalanced or have skewed class distribution. As a result, a variety of cost-sensitive learning approaches and strategies can be used to solve problems with imbalanced data. [50], [58], [152].

The goal of cost-sensitive learning for imbalanced classification is to assign different costs to different types of misclassification errors, then utilize specific algorithms to compensate for those costs. The concept of a cost matrix facilitates to understand the varied costs of misclassification. A confusion matrix is a list of a model's predictions on classification tasks. It is a table that lists the number of predictions made for each class, separated by the actual class [153].

It is easiest to understand using a classification issue with negative and positive classes, which are commonly labelled with 0 and 1 class labels. Although the meanings of rows and columns can be and often are interchanged with no loss of meaning, the columns in a matrix table indicate the actual class to which the instances belong, and the rows represent the anticipated class. A cell is the number of samples that fulfil the row and column's requirements, and each cell has a unique common name. A confusion matrix for a classification problem is shown in Table II.

The confusion matrix's cost matrix is a matrix that allocates a cost to each cell. The focus of the research on the unbalanced data problem, in relation to the confusion matrix, is on errors, hence 'False Positive' and 'False Negative' will be the primary focal areas. In an imbalanced classification task or a challenge with imbalanced data, the latter is more common than the former.

TABLE II. A CONFUSION MATRIX FOR A CLASSIFICATION TASK

	Actual Negative	**Actual Positive**
Predicted Negative	True Negative (TN)	False Negative (FN)
Predicted Positive	False Positive (FP)	True Positive (TP)

IV. Prominent Validation Metrics For Highly Imbalanced Multi-Class Data

The precision metric, defined as Equation (1), measures the accurately categorized positive class samples.

$$\text{Precision} = \frac{\text{TP}}{\text{TP+FP}} \quad (1)$$

where TP and FP stand for true-positive and false-positive counts, respectively.

The fraction of accurately identified true positive samples is measured by recall, which is calculated using Equation (2):

$$Recall = \frac{TP}{TP+FN} \quad (2)$$

G-mean and AUC have been designed for class imbalanced problem-measurement. As shown in Equation (3), the F-measure or F-1 score is the harmonic mean of precision and recall:

$$f - measure = \frac{(1+\beta)^2 \; x \; recall \; x \; precision}{\beta \; x \; recall + precision} = \frac{(1+\beta) x \frac{TP}{P} x \frac{TP}{PP}}{\frac{TP}{P} x \; \beta + \frac{TP}{PP}} \quad (3)$$

The geometric mean, or G-mean, is a metric that assesses the balanced performance of a classifier, as shown in Equation (4):

$$G - mean = \sqrt{\frac{TP}{TP+FN} x \frac{TN}{TN+FP}} \quad (4)$$

The AUC stands for Area under the ROC Curve, which is used to assess the model's performance [21] and can be used to estimate it, as demonstrated in Equation (5):

$$\text{AUC} = \left(\frac{TP}{TP+FN} + \frac{TN}{TN+FP}\right)/2 \quad (5)$$

As a result, the performance evaluators of the classifier employed in most imbalanced data research include F-measure, G-mean, AUC, and Accuracy [1], [154]. However, in a highly imbalanced data situation, accuracy, also known as balanced accuracy, is not a feasible validation metric due to the nature of the classifier, which is bias towards the majority class and ignores the minority class [21], [63], [155].

A. Related Study on HIMC Data Framework

Before a new framework on highly imbalanced multi-class data can be proposed, related studies on both highly imbalanced binary data and highly imbalanced multi-class data must be discussed thoroughly. This section will present the descriptions and discussions on the differences between highly imbalanced binary data and highly imbalanced multi-class data in terms of literatures, technologies, and domain in real world.

Table III shows literature related to highly imbalanced binary (HIB) data and highly imbalanced multi-class (HIMC) data. Before further explanation on this table is given, a clarification on “Data Mining” domain will be given. It is known as “Data Mining” domain due to the nature of the research which uses multiple different datasets ranging from five datasets to 35 datasets. Each of the datasets are different in terms of discipline and areas of interest such as Abalone, Glasses, Cars, and Credits etc. Therefore, each dataset is a unique domain and thus it is called “Data Mining” domain.

From 2006 to 2021, there were many literatures published about highly imbalanced data. Some of the literatures were on HIB data and the rest were on HIMC data. The related HIB data was from the data mining domain, and this data was grouped based on these algorithms: SR 0-1 LOSS [15], F-BFR [156], T-BFS [157], K-SUB [64], PSU [102], C4.5N [158], DWCE [159], ECSM [72], GP-COACH [160], Fuzzy [161], [162], OSPREY [163], Chi Method, CNN [164], WL-Norm SVM [165], EAIS + Fuzzy [166], GA + Fuzzy [167], US + Ensemble (Data Hardness) [15], Clustering + WS [66], DBE-DCR [73], EUBoost [168], GA-FS-GL [169], GSVM-RU [170], GA-GL+FRBC [171], K-Means + HFS [172], REPMAC k-Means +SVM + DT [173], SwitchingNED [70], SVM-US [174], B-BFS [67].

For the HIMC data, the related domain was the data mining: FMFS [186], DM-UCML [188], WRSEW + RDL [189], aerial imaging: RF-MML [187], bioinformatics: DBNN (DEEP SOM) [18], CNN [56], cyber-attack detection: SAE [190], facial recognition: VFSG [191], fetal aneuploidies: ANN [192], medical: CRF [40], CNN [144], pathology: CNN [193], power synchropasor detection: RXGBOOST [61] and flare forecast: US + OS + SVM [7].

In HIB data, specifically in the data mining domain, there are many published literatures between 2006 and 2021. For data mining domain, several novel methods were developed. For example, Calvert et al. who focused on severely imbalanced big dataset found that C4.5N was by far the best learner in terms of Slow POST attack data. Nevertheless, C4.5 Decision Tree and Chi Algorithm were found to be less effective compared GP-COACH in highly imbalanced dataset [158]. Among the more recent literatures in HIB in data mining domain come from Bringer et al. The authors found that data labeling is difficult in highly imbalanced data environment. Therefore, the authors have proposed OSPREY which is a system to cater data labeling in highly imbalanced data. The system was developed on top of Snorkel framework [163].

The CL is the most popular strategy in HIB data. It can be seen by the number of literatures found especially in the data mining domain. For example, Fernandez et al. studied the behavior of GP-COACH algorithm in highly imbalanced data scenario [160]. Liu et al. developed a novel framework for unbalanced classification that focuses on generating a strong ensemble by self-paced harmonizing data hardness using a mix of under sampling and self-paced ensemble in CL for HIB data [15].

For HIMC data, less literatures were found from 2006 to 2021. Unlike HIB data, there is less research in data mining domains where there is only a single literature at DL [219] and few literatures at AL [18], [56], [40], [190], [144], [187], [121], [191]–[193], and some literature in CL [61], [63], [188], [189], [7].

Domains such imaging [187], bioinformatics [18], [56], cyber-attack [190], facial recognition [191], fetal aneuploidies [192], medical [40], [144] and pathology [193] concentrated on AL strategy while domains such as medical [63], power [61] and solar forecast [7] employed CL strategy. Nonetheless, the only research using DL strategy focused on data mining domain.

TABLE III. HIGHLY IMBALANCED BINARY DATA VS HIGH IMBALANCED MULTI-CLASS DATA

Class	**Data Level**	**Domain**	**Algorithm Level**	**Domain**	**Combination Level**	**Domain**
Binary class	SR 0-1 LOSS [15], F-BFR [156], T-BFS [157], K-SUB [64], PSU [102]	Data Mining	C4.5N [158], DWCE [159], ECSM [72], GP-COACH [160], Fuzzy [161], [162], OSPREY [163], Chi Method, CNN[164], WL-Norm SVM [165]	Data Mining	EAIS + Fuzzy [166], GA + Fuzzy [167], US + Ensemble (Data Hardness) [15], Clustering + WS [66], DBE-DCR [73], EUBoost [168], GA-FS-GL [169], GSVM-RU [170], GA-GL+FRBC [171], K-Means + HFS [172], REPMAC k-Means +SVM + DT [173], SwitchingNED [70], SVM-US [174], B-BFS [67]	Data Mining
	NRA [175]	Fraud Detection	BERT [19]	Malware Detection	Ensemble + RF [68]	Disease Prediction
	SSFS [176]	Phishing Detection	CNN [177], ECDL [178]	Medical Imaging	DS + ST [179]	Hospital Admission
	BPFs [180], MPRM [181]	Speech Recognition	AL [168]	Social Media	PCA + DA + CNN [145]	Imaging
			GSVM-BA [182]	Spam Detection	SMOTE-tBPSO-SVM [183]	Malware Detection
					Boosting + Sampling [184], SMOTE + RF [185]	Medical
					GA + SVM-SMOTE [65]	Mortality Prediction
					Ensemble + SMOTE [71]	Sentiment Analysis
Multi-class	FMFS [186]	Data Mining	RF-MML [187]	Imaging	DM-UCML [188], WRSEW + RDL [189]	Data Mining
			DBNN (DEEP SOM) [18], CNN [56]	Bioinformatics	ELF [63]	Medical
			SAE [190]	Cyber Attack	RXGBOOST [61]	Power
			UCML [121]	Data Mining	US + OS + SVM [7]	Solar Forecast
			VFSG [191]	Facial Recognition		
			ANN [192]	Fetal Aneuploidies		
			CRF [40], CNN [144]	Medical		
			CNN [193]	Pathology		

Therefore, it clear that HIB data is more prominent than HIMC data and there is more room for future research to be conducted in highly imbalanced data. HIB data alone dominates the research with 72.7% while HIMC data with only 27.3%.

Table III demonstrates highly imbalanced data categorization based on techniques used. There are four categories of technique that have been used in HIB data and HIMC data research works. They are statistical, semi-supervised, supervised, and unsupervised. This consists of 81.6% from the overall literatures in CL and almost half which is 46.3% from overall literatures in highly imbalanced data.

Combination Level (CL) dominated in terms of number of literatures in both HIB and HIMC data. Supervised type of research in HIB using ensemble were published between 2006 and 2021. In the same technique category (supervised in HIB data), hybrid method is also not less popular with quite several literatures published within the same period while several literatures for unsupervised type of research have been found in binary class data using hybrid technique. Overall, there have been quite satisfying number of literatures published on CL in term of HIB data.

For HIMC data, number of research in ensemble and hybrid was not as many as HIB data, there are only few literatures published between 2006 until 2021 which was literatures in statistical area using hybrid, and in unsupervised area and other literatures in the area were using ensemble technique. Research in HIMC data area is still quite new.

In terms of feature selection methods in both HIB data and HIMC data, there were several techniques involved in the list of literatures. The technique used was Repetitive Feature Selection, Threshold-based Feature Selection, Binary to Multi-class Feature Selection, Program to Detect Phishing and Feature Maximization for Feature Selection.

The rise of Big Data is one of the most prominent justifications for the adoption of both ensemble and hybrid in all three data techniques (DL, AL, and CL) in both HIB and HIMC data. Big data processing, as well as hybridization and algorithm ensembles have attracted much interest in the research community. Referring to the literatures that have been analyzed, the earliest literature on highly imbalanced data was published in 2006 for HIB data and 2013 for HIMC data respectively, after the emergence of Big Data technology.

V. Proposed Novel HIMC Data Framework

This article presents the descriptions on the technologies used and gathered from previous studies and the domains of the literatures published between 2013 and 2021. The literatures were divided further into four categories which were Supervised, Semi-supervised, Unsupervised, and Statistical. In terms of types of algorithms, the publications were segregated based on several groups which were feature selection, artificial neural network, deep learning, case-sensitive learning, ensemble, and hybrid, including two specific algorithms developed for aerial scene imaging and facial expression recognition.

Fig. 2 illustrates the proposed framework of HIMC data. The framework has been developed based on important elements like data category, data behavior, data characteristics, data strategy, model or technique type, and algorithm or model or technique or framework or classifier used in HIMC data research. The HIMC data framework can be used as the underlying basis of highly imbalanced data research both binary data and multi-class data.

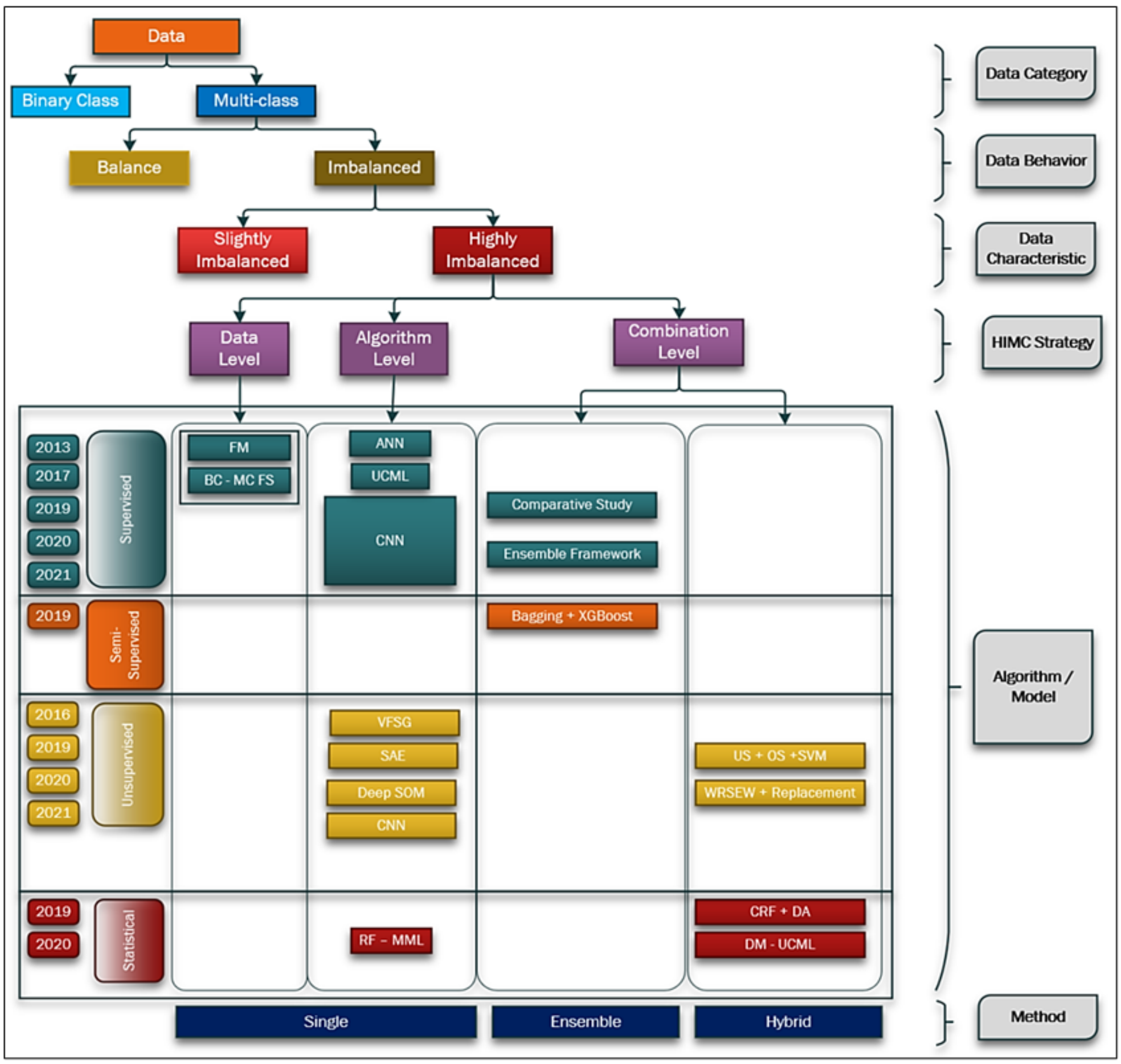


Fig. 2. HIMC Data Framework.

Based on the framework, data can be categorized into binary and multi-class. There are other categories of data such as streamed data and big data, but this research only focuses on these two categories of data. Moreover, both streamed data and big data research are higher-level research when the underlying data categories will come back to multi-class data and binary class data as the base category. Multi-class data can be further divided into balanced and imbalanced data. Balanced data is only present in controlled environment where both majority and minority classes and data are divided to 50-50 percent ratio.

Algorithms or techniques or methods can be divided into different categories which are Supervised, Semi-Supervised, Unsupervised and Statistical. In this framework, year of publication has also been added for easy referencing. Finally, there is technique or model group. It can be categorized into single method, ensemble, and hybrid group. Single method group lists only research works that use a single algorithm or technique or method. Ensemble group is a list of studies that employ ensemble method and hybrid group is a list of research that employ hybrid method. Finally, it is worth mentioning that in this framework, domains for each literature are also stated next to the method used. Multiple domains mean that the referred articles mentioned the use of many benchmarks in many disciplines.

In the framework, highly imbalanced data is further divided into DL, AL, and HM [4], [21]. Since ensemble can work independently or combined with other methods in DL or AL, HM and EM are combined to employ CL strategy. In HIMC data framework, there have been only a couple of literatures related to DL strategy. Using Feature Maximization in Feature Selection, Jean-Charles Lamirel devised a strategy for coping with severely imbalanced textual data grouped into similar classes. [194]. Another study came from Kubler et al. where FS was used to handle problem with how to extend FS from binary classification data to multi-class data [195]. Both literatures used multiple domains in their research.

There are several literatures published between 2016 and 2021 on HIMC data in AL. Domains such as in the study of fetal aneuploidies, medical, application development of facial expression recognition, cyber-attack, aerial scene imaging and bioinformatics. Several articles stated the use of supervised method [56], [144], [121], [192], [196] and others stated the use of unsupervised method [191], [18], [83], [193] and one article mentioned the use of statistical method. All research related in DL and AL employed single method and thus, they were categorized into single method group [187].

CL gained more popularity in recent times. Even though there have been only little literatures related to CL strategy, most of these articles have the best performance benchmark. Two articles mentioned the use of supervised method in medical and malware detection domains [62], [63]. One article mentioned the employment of semi-supervised in power synchrophasors detector domain [61]. Other articles mentioned the use of unsupervised method in rare event of flare forecast and multiple domains [7], [189] and finally the statistical method was used in medical imaging and multiple domains based on several articles [40], [121].

VI. Discussion

Multi-class data issues are more difficult to solve compared to binary classification data [22]. Most of the articles published recently focus on handling problems in binary classes such as bankruptcy prediction (bankrupt vs. non-bankrupt), computer security (normal activity vs. malicious activity), medical (healthy vs. infected). Because multi-class data can be dissected and handled using binary class methods, many literatures focus on binary class data [21], [31].

Multi-class data have different challenges. If the output for binary classification is binary, multi-class output will be multiplied (more than two targets or results) [1]. Several examples of real-world cases involving highly imbalanced data are like hospital readmission [197], feature pattern of Thoracolumbar spine fracture [87], solar flare forecast [7], modeling of Chlorophyll concentration in Algal Bloom [56], semantic segmentation [198], [199], medical imaging [40], malware detection [19], cyber-physical strike discovery [190] and Bioinformatics [18].

In highly imbalanced multi-class data, a multiple skewed distribution is an issue which affects classifiers' performance as it is difficult to decide the boundaries in highly imbalanced multi-class data. A distance-based algorithm such as Hellinger distance has been used to handle the problem. However, there are issues on overlapping class and noisy data [21], [33], [155]. A good solution for the problem is to combine a method that can minimize overlapping class and reduce noisy data [2], [31], [200]. A promising solution that might handle both problems is by using ensemble methods. This is because this method is capable of rectifying imbalance class and improving weak classifiers [201]. However, ensembles suffer from lack of interpretability and are usually computationally expensive [26].

Problems in highly imbalanced data originate from problems in slightly imbalanced data which are alleviated due to the nature of severe imbalance ratio (IR) [3], [33]. Thus, problems in imbalanced multi-class data and highly imbalanced multi-class data can be considered similar in nature with high IR. In ML, multinomial or multi-class classification is the problem of classifying instances into one of three or more classes [21]. A multi-class classification problem is not as developed as the binary classification problem in imbalanced data [4], [21], [110].

High IR has negative impact toward the minority class and overall performance of data may result in information loss [102], [202]. In multi-class data, performance of each class needs to be focused specifically because each class is distinctive. A classifier might obtain good performance in some classes while unsatisfactory results in other classes [4], [18], [21].

Relationship among classes is another issue in multi-class data. It is a complicated situation because each group of classes presents different problems [4]. Two or more classes might overlap in some group. While other classes might have a normal borderline and considered as normal classes. This problem is elevated in highly imbalanced multi-class data, and it affects classifiers' performance [159], [40].

Leevy et al. and Rendon et al. suggest that heuristic-based and more flexible methods have been less developed to solve multi-class data problems [31], [203]. There are literatures in the analysis of relationships among classes in multi-class data that have produced satisfactory results [155]. However, there is a need to improve on highly imbalanced multi-class data domains as the same method has not produced good results in a highly imbalanced environment.

In multi-class data, class overlapping is an issue because it may happen anywhere within the dataset with different groups of overlap classes. The problem is more complicated in highly imbalanced multi-class data scenarios, as it is hard to define borders of the overlapped classes. The problem will affect the classification or prediction performance of the classifier. The data needed to be properly pre-processed and appropriate sampling procedure be applied [4], [47], [47]. The challenge is to develop solutions that consider the different features within the overlapped classes and at the same time also showing good classification or prediction performance [204].

The presence of noisy data in a dataset can increase the imbalance among classes and eventually could result in the classifier rebalancing the wrong classes [205]. To overcome the issue, a new approach for data cleaning is needed to boost classifier performance in multi-class data [46]. Another way to solve the problem is to create synthetic data, which allows overlapping data to be transferred to other locations. Another method is to remove excessive samples. New approaches should consider removing excessive samples while maintaining the quality of the data [2], [6].

In summary, in highly imbalanced multi-class data, minority classes are treated as noisy data. However, the minority data might have crucial information [1], [21]. Another issue is the misclassification cost between the majority and minority data. The cost-sensitive strategy is a well-known method for dealing with data imbalances. Minority samples, on the other hand, had a higher cost of misclassification than majority samples. [50]. Finally, a highly imbalanced multi-class data has multiple skewed classes which reduces classifier performance as it is difficult to normalize the skewness [26], [31], [44] and difficult to define borders of the overlap classes [4], [46]. Future research in highly imbalanced multi-class data should focus on these issues.

Future direction of HIMC data is clear, that is to have more research focusing on the issues presented in this article. One of the main issues of HIMC was the high IR between data classes. This was because the presence of multiple classes with high IR caused multiple skewed distribution and would affect the minority class as it was ignored by the classifiers because they tended to be biased toward the majority class. Another issue was the overlapping classes. This involved the difficulty to define borders of multiple classes. Finally, the issue on the presence of noisy data especially in big dataset. In this issue, classifiers tended to treat the minority data as noisy data due to the high IR.

VII. Conclusion

Data-related research has evolved as more essential, exciting, and beneficial due to the rapid rise of Big Data. In this research, machine learning algorithms at different levels have been addressed by using different strategies which are DL, AL, and CL to properly manage highly imbalanced multi-class data problems. It has proposed a novel framework for HIMC data in the wake of issues concerning HIMC data. However, due to the dynamism and uniqueness of each dataset and the heterogeneity of data, developing a proper method or technique in handling highly imbalanced multi-class data remains a challenge. As a result, the current state-of-the-art algorithms were found and categorized in four different types in this study: supervised, semi-supervised, unsupervised, and statistical. Finally, the performance of various machine learning techniques used to handle HIMC data is compared in this research. Hence, based on the analysis performed, there is a need for a novel framework of HIMC data to be designed. Finally, open issues, challenges, and future direction of HIMC data have been discussed and presented in this article to pave the road for extended research works in HIMC data to be conducted in the future.

Acknowledgment

Ts. Mohd Hakim bin Abdul Hamid, a PhD candidate, would like to express his gratitude to the Information Security, Digital Forensic, and Computer Networking (INSFORNET) research group, Center for Advanced Computing Technology (C-ACT), Fakulti Teknologi Maklumat dan Komunikasi (FTMK), and Universiti Teknikal Malaysia Melaka (UTeM) for funding his PhD scholarship program. He would also want to thank his supervisor and co-supervisor, both of whom are co-authors on this study. He would like to express his gratitude to the Institute for Big Data Analytics and Artificial Intelligence (IBDAAI) and Universiti Teknologi MARA (UiTM) for providing him with the chance to pursue his Ph.D. at UiTM.

References

[1] J. Tanha, Y. Abdi, N. Samadi, N. Razzaghi, and M. Asadpour, “Boosting methods for multi-class imbalanced data classification: an experimental review,” J. Big Data, vol. 7, no. 1, 2020, doi: 10.1186/s40537-020-00349-y.

[2] D. Devi, S. K. Biswas, and B. Purkayastha, “A Review on Solution to Class Imbalance Problem: Undersampling Approaches,” 2020 Int. Conf. Comput. Perform. Eval. ComPE 2020, pp. 626–631, 2020, doi: 10.1109/ComPE49325.2020.9200087.

[3] C. Bellinger, S. Sharma, N. Japkowicz, and O. R. Zaïane, “Framework for extreme imbalance classification: SWIM—sampling with the majority class,” Knowl. Inf. Syst., vol. 62, no. 3, pp. 841–866, 2020, doi: 10.1007/s10115-019-01380-z.

[4] B. Krawczyk, “Learning from imbalanced data: open challenges and future directions,” Prog. Artif. Intell., vol. 5, no. 4, pp. 221–232, 2016, doi: 10.1007/s13748-016-0094-0.

[5] J. M. Johnson and T. M. Khoshgoftaar, “Survey on deep learning with class imbalance,” J. Big Data, vol. 6, no. 1, 2019, doi: 10.1186/s40537-019-0192-5.

[6] W. C. Sleeman and B. Krawczyk, “Bagging Using Instance-Level Difficulty for Multi-Class Imbalanced Big Data Classification on Spark,” Proc. - 2019 IEEE Int. Conf. Big Data, Big Data 2019, pp. 2484–2493, 2019, doi: 10.1109/BigData47090.2019.9006058.

[7] A. Ahmadzadeh et al., “Challenges with extreme class-imbalance and temporal coherence: A study on solar flare data,” arXiv, 2019.

[8] B. Mirzaei, B. Nikpour, and H. Nezamabadi-pour, “CDBH: A clustering and density-based hybrid approach for imbalanced data classification,” Expert Syst. Appl., vol. 164, no. April 2020, p. 114035, 2021, doi: 10.1016/j.eswa.2020.114035.

[9] J. Jedrzejowicz and P. Jedrzejowicz, "GEP-based classifier with drift detection for mining imbalanced data streams," Procedia Comput. Sci., vol. 176, pp. 41–49, 2020, doi: 10.1016/j.procs.2020.08.005.

[10] J. Jedrzejowicz and P. Jedrzejowicz, "GEP-based classifier for mining imbalanced data," Expert Syst. Appl., vol. 164, no. May 2020, p. 114058, 2021, doi: 10.1016/j.eswa.2020.114058.

[11] N. Jiang and N. Li, "A wind turbine frequent principal fault detection and localization approach with imbalanced data using an improved synthetic oversampling technique," Int. J. Electr. Power Energy Syst., vol. 126, no. PA, p. 106595, 2021, doi: 10.1016/j.ijepes.2020.106595.

[12] S. Wang and L. L. Minku, "AUC Estimation and Concept Drift Detection for Imbalanced Data Streams with Multiple Classes," Proc. Int. Jt. Conf. Neural Networks, vol. 2, no. Section V, 2020, doi: 10.1109/IJCNN48605.2020.9207377.

[13] B. Mirzaei, B. Nikpour, and H. Nezamabadi-Pour, "An under-sampling technique for imbalanced data classification based on DBSCAN algorithm," pp. 21–26, 2020, doi: 10.1109/cfis49607.2020.9238718.

[14] A. Goyal and J. Khiari, "Diversity-aware weighted majority vote classifier for imbalanced data," arXiv, 2020.

[15] Z. Liu et al., "Self-paced ensemble for highly imbalanced massive data classification," Proc. - Int. Conf. Data Eng., vol. 2020-April, pp. 841–852, 2020, doi: 10.1109/ICDE48307.2020.00078.

[16] M. E. Khoda, "Mobile Malware Detection with Imbalanced Data using a Novel Synthetic Oversampling Strategy and Deep Learning," 2020, doi: 10.1109/WiMob50308.2020.9253433.

[17] J. Zhao, J. Jin, S. Chen, R. Zhang, B. Yu, and Q. Liu, "A weighted hybrid ensemble method for classifying imbalanced data," Knowledge-Based Syst., vol. 203, p. 106087, 2020, doi: 10.1016/j.knosys.2020.106087.

[18] L. A. Bugnon, C. Yones, D. H. Milone, and G. Stegmayer, "Deep neural architectures for highly imbalanced data in bioinformatics," IEEE Trans. Neural Networks Learn. Syst., vol. 31, no. 8, pp. 2857–2867, 2020, doi: 10.1109/TNNLS.2019.2914471.

[19] R. Oak, M. Du, D. Yan, H. Takawale, and I. Amit, "Malware detection on highly imbalanced data through sequence modeling," Proc. ACM Conf. Comput. Commun. Secur., pp. 37–48, 2019, doi: 10.1145/3338501.3357374.

[20] R. Zhu, Y. Guo, and J. H. Xue, "Adjusting the imbalance ratio by the dimensionality of imbalanced data," Pattern Recognit. Lett., vol. 133, pp. 217–223, 2020, doi: 10.1016/j.patrec.2020.03.004.

[21] H. Ali, M. N. M. Salleh, R. Saedudin, K. Hussain, and M. F. Mushtaq, "Imbalance class problems in data mining: A review," Indones. J. Electr. Eng. Comput. Sci., vol. 14, no. 3, pp. 1552–1563, 2019, doi: 10.11591/ijeecs.v14.i3.pp1552-1563.

[22] H. Patel, D. Singh Rajput, G. Thippa Reddy, C. Iwendi, A. Kashif Bashir, and O. Jo, "A review on classification of imbalanced data for wireless sensor networks," Int. J. Distrib. Sens. Networks, vol. 16, no. 4, 2020, doi: 10.1177/1550147720916404.

[23] Q. Ya-Guan et al., "EMSGD: An Improved Learning Algorithm of Neural Networks with Imbalanced Data," IEEE Access, vol. 8, pp. 64086–64098, 2020, doi: 10.1109/ACCESS.2020.2985097.

[24] D. Gan, J. Shen, B. An, M. Xu, and N. Liu, "Integrating TANBN with cost sensitive classification algorithm for imbalanced data in medical diagnosis," Comput. Ind. Eng., vol. 140, no. June 2019, p. 106266, 2020, doi: 10.1016/j.cie.2019.106266.

[25] C. C. Lin, D. J. Deng, C. H. Kuo, and L. Chen, "Concept drift detection and adaption in big imbalance industrial IoT data using an ensemble learning method of offline classifiers," IEEE Access, vol. 7, pp. 56198–56207, 2019, doi: 10.1109/ACCESS.2019.2912631.

[26] Z. Chen, J. Duan, L. Kang, and G. Qiu, "A hybrid data-level ensemble to enable learning from highly imbalanced dataset," Inf. Sci. (Ny)., vol. 554, pp. 157–176, 2021, doi: 10.1016/j.ins.2020.12.023.

[27] B. Zhao and Q. Yuan, "Improved generative adversarial network for vibration-based fault diagnosis with imbalanced data," Meas. J. Int. Meas. Confed., vol. 169, no. July 2020, p. 108522, 2021, doi: 10.1016/j.measurement.2020.108522.

[28] P. Zyblewski, R. Sabourin, and M. Woźniak, "Preprocessed dynamic classifier ensemble selection for highly imbalanced drifted data streams," Inf. Fusion, vol. 66, no. June 2020, pp. 138–154, 2021, doi: 10.1016/j.inffus.2020.09.004.

[29] H. Kaur, H. S. Pannu, and A. K. Malhi, "A systematic review on imbalanced data challenges in machine learning: Applications and solutions," ACM Comput. Surv., vol. 52, no. 4, 2019, doi: 10.1145/3343440.

[30] I. Triguero, S. Del Río, V. López, J. Bacardit, J. M. Benítez, and F. Herrera, "ROSEFW-RF: The winner algorithm for the ECBDL'14 big data competition: An extremely imbalanced big data bioinformatics problem," Knowledge-Based Syst., vol. 87, pp. 69–79, 2015, doi: 10.1016/j.knosys.2015.05.027.

[31] J. L. Leevy, T. M. Khoshgoftaar, R. A. Bauder, and N. Seliya, "A survey on addressing high-class imbalance in big data," J. Big Data, vol. 5, no. 1, 2018, doi: 10.1186/s40537-018-0151-6.

[32] H. He and E. A. Garcia, Learning from imbalanced data, vol. 21, no. 9. 2009.

[33] S. Sharma, C. Bellinger, B. Krawczyk, O. Zaiane, and N. Japkowicz, "Synthetic Oversampling with the Majority Class: A New Perspective on Handling Extreme Imbalance," Proc. - IEEE Int. Conf. Data Mining, ICDM, vol. 2018-Novem, pp. 447–456, 2018, doi: 10.1109/ICDM.2018.00060.

[34] A. Lazarevic, J. Srivastava, and V. Kumar, "Data Mining for Analysis of Rare Events: A Case Study in Security, Financial and Medical Applications," Pacific-Asia Conf. Knowl. Discov. Data Min., no. December 2014, 2004.

[35] J. Roland, "How negative sampling provides class balance to rare event case data using a vehicular accident prediction project as a use case scenario," 2020, [Online]. Available: https://scholar.utc.edu/theses/681/.

[36] F. Carcillo, Y.-A. Le Borgne, O. Caelen, Y. Kessaci, F. Oblé, and G. Bontempi, "Combining unsupervised and supervised learning in credit card fraud detection," Inf. Sci. (Ny)., 2019, doi: https://doi.org/10.1016/j.ins.2019.05.042.

[37] W. Wei, J. Li, L. Cao, Y. Ou, and J. Chen, "Effective detection of sophisticated online banking fraud on extremely imbalanced data," World Wide Web, vol. 16, no. 4, pp. 449–475, 2013, doi: 10.1007/s11280-012-0178-0.

[38] B. Krawczyk, "Cost-sensitive one-vs-one ensemble for multi-class imbalanced data," Proc. Int. Jt. Conf. Neural Networks, vol. 2016-Octob, pp. 2447–2452, 2016, doi: 10.1109/IJCNN.2016.7727503.

[39] M. S. Sainin, R. Alfred, and F. Ahmad, "Ensemble Meta Classifier with Sampling and Feature Selection for Data with Multiclass Imbalance Problem," J. Inf. Commun. Technol., vol. 20, no. 2, pp. 103–133, 2021, doi: 10.32890/jict2021.20.2.1.

[40] J. Wang and S. Valaee, "From whole to parts: Medical imaging semantic segmentation with very imbalanced data," 2019 IEEE Glob. Commun. Conf. GLOBECOM 2019 - Proc., pp. 1–6, 2019, doi: 10.1109/GLOBECOM38437.2019.9014112.

[41] Y. Lu, "HKBU Institutional Repository Advances in imbalanced data learning Doctor of Philosophy," 2019.

[42] J. M. Johnson and T. M. Khoshgoftaar, "Deep learning and thresholding with class-imbalanced big data," Proc. - 18th IEEE Int. Conf. Mach. Learn. Appl. ICMLA 2019, pp. 755–762, 2019, doi: 10.1109/ICMLA.2019.00134.

[43] Z. Chen, W. Chen, and Y. Shi, "Ensemble learning with label proportions for bankruptcy prediction," Expert Syst. Appl., vol. 146, p. 113155, 2020, doi: 10.1016/j.eswa.2019.113155.

[44] K. Fujiwara et al., "Over- and Under-sampling Approach for Extremely Imbalanced and Small Minority Data Problem in Health Record Analysis," Front. Public Heal., vol. 8, no. May, pp. 1–15, 2020, doi: 10.3389/fpubh.2020.00178.

[45] X. Huang, C. Z. Zhang, and J. Yuan, "Predicting Extreme Financial Risks on Imbalanced Dataset: A Combined Kernel FCM and Kernel SMOTE Based SVM Classifier," Comput. Econ., vol. 56, no. 1, pp. 187–216, 2020, doi: 10.1007/s10614-020-09975-3.

[46] M. Koziarski, "Radial-Based Undersampling Algorithm for Classification of Breast Cancer Histopathological Images Affected by Data Imbalance," Proc. - 2019 12th Int. Congr. Image Signal Process. Biomed. Eng. Informatics, CISP-BMEI 2019, no. 1, pp. 2–6, 2019, doi: 10.1109/CISP-BMEI48845.2019.8966010.

[47] M. Koziarski, B. Krawczyk, and M. Woźniak, "Radial-Based oversampling for noisy imbalanced data classification," Neurocomputing, vol. 343, pp. 19–33, 2019, doi: 10.1016/j.neucom.2018.04.089.

[48] C. M. Vong and J. Du, "Accurate and efficient sequential ensemble learning for highly imbalanced multi-class data," Neural Networks, vol. 128, pp. 268–278, 2020, doi: 10.1016/j.neunet.2020.05.010.

[49] J. M. Johnson and T. M. Khoshgoftaar, "Deep learning and data sampling with imbalanced big data," Proc. - 2019 IEEE 20th Int. Conf. Inf. Reuse Integr. Data Sci. IRI 2019, pp. 175–183, 2019, doi: 10.1109/IRI.2019.00038.

[50] N. Ghatasheh et al., "Cost-sensitive ensemble methods for bankruptcy prediction in a highly imbalanced data distribution: a real case from the Spanish market," Prog. Artif. Intell., vol. 9, no. 4, pp. 361–375, 2020, doi: 10.1007/s13748-020-00219-x.

[51] A. D. Amirruddin, F. M. Muharam, M. H. Ismail, N. P. Tan, and M. F. Ismail, "Hyperspectral spectroscopy and imbalance data approaches for classification of oil palm's macronutrients observed from frond 9 and 17," Comput. Electron. Agric., vol. 178, no. August, p. 105768, 2020, doi: 10.1016/j.compag.2020.105768.

[52] M. Moniruzzaman, A. Bagirov, and I. Gondal, "Partial Undersampling of Imbalanced Data for Cyber Threats Detection," ACM Int. Conf. Proceeding Ser., pp. 2–5, 2020, doi: 10.1145/3373017.3373026.

[53] J. Novakovic and S. Markovic, "Performance of Support Vector Machine in Imbalanced Data Set," 2020 19th Int. Symp. INFOTEH-JAHORINA, INFOTEH 2020 - Proc., no. March, pp. 1–5, 2020, doi: 10.1109/INFOTEH48170.2020.9066276.

[54] P. B. C. Castro, B. Krohling, A. G. C. Pacheco, and R. A. Krohling, "An app to detect melanoma using deep learning: An approach to handle imbalanced data based on evolutionary algorithms," Proc. Int. Jt. Conf. Neural Networks, pp. 2–7, 2020, doi: 10.1109/IJCNN48605.2020.9207552.

[55] J. Ahammad, N. Hossain, and M. S. Alam, "Credit card fraud detection using data pre-processing on imbalanced data - Both oversampling and undersampling," ACM Int. Conf. Proceeding Ser., pp. 18–21, 2020, doi: 10.1145/3377049.3377113.

[56] J. H. Choi, J. Kim, J. Won, and O. Min, "Modelling Chlorophyll-a Concentration using Deep Neural Networks considering Extreme Data Imbalance and Skewness," Int. Conf. Adv. Commun. Technol. ICACT, vol. 2019-Febru, pp. 631–634, 2019, doi: 10.23919/ICACT.2019.8702027.

[57] K. Cheng, S. Gao, W. Dong, X. Yang, Q. Wang, and H. Yu, "Boosting label weighted extreme learning machine for classifying multi-label imbalanced data," Neurocomputing, vol. 403, pp. 360–370, 2020, doi: 10.1016/j.neucom.2020.04.098.

[58] L. Loezer, F. Enembreck, J. P. Barddal, and A. De Souza Britto, "Cost-sensitive learning for imbalanced data streams," Proc. ACM Symp. Appl. Comput., pp. 498–504, 2020, doi: 10.1145/3341105.3373949.

[59] Q. Liu, W. Luo, and T. Shi, "Classification method for imbalanced data set based on EKCStacking algorithm," ACM Int. Conf. Proceeding Ser., pp. 51–56, 2019, doi: 10.1145/3375998.3376002.

[60] H. Li, Y. Cao, S. Li, J. Zhao, and Y. Sun, "XGBoost Model and Its Application to Personal Credit Evaluation," IEEE Intell. Syst., vol. 35, no. 3, pp. 52–61, 2020, doi: 10.1109/MIS.2020.2972533.

[61] J. Wang, Z. Sun, B. Bao, and D. Shi, "Malicious synchrophasor detection based on highly imbalanced historical operational data," CSEE J. Power Energy Syst., vol. 5, no. 1, pp. 11–20, 2019, doi: 10.17775/cseejpes.2018.00200.

[62] Y. Pang et al., "Finding Android Malware Trace from Highly Imbalanced Network Traffic," Proc. - 2017 IEEE Int. Conf. Comput. Sci. Eng. IEEE/IFIP Int. Conf. Embed. Ubiquitous Comput. CSE EUC 2017, vol. 1, pp. 588–595, 2017, doi: 10.1109/CSE-EUC.2017.108.

[63] J. Li, B. Xin, Z. Yang, J. Xu, S. Song, and X. Wang, "Harmonization centered ensemble for small and highly imbalanced medical data classification," Proc. - Int. Symp. Biomed. Imaging, vol. 2021-April, pp. 1742–1745, 2021, doi: 10.1109/ISBI48211.2021.9433824.

[64] T. Tran, L. Tran, and A. Mai, "K-Segments under Bagging approach: An experimental Study on Extremely Imbalanced Data Classification," Proc. - 2019 19th Int. Symp. Commun. Inf. Technol. Isc. 2019, pp. 492–495, 2019, doi: 10.1109/ISCIT.2019.8905145.

[65] R. Ghorbani, R. Ghousi, A. Makui, and A. Atashi, "A New Hybrid Predictive Model to Predict the Early Mortality Risk in Intensive Care Units on a Highly Imbalanced Dataset," IEEE Access, vol. 8, pp. 141066–141079, 2020, doi: 10.1109/ACCESS.2020.3013320.

[66] P. Ksieniewicz and R. Burduk, "Clustering and weighted scoring in geometric space support vector machine ensemble for highly imbalanced data classification," Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics), vol. 12140 LNCS, pp. 128–140, 2020, doi: 10.1007/978-3-030-50423-6_10.

[67] M. Alshawabkeh, M. Moffie, F. Azmandian, J. A. Aslam, J. Dy, and D. Kaeli, "Effective virtual machine monitor intrusion detection using feature selection on highly imbalanced data," Proc. - 9th Int. Conf. Mach. Learn. Appl. ICMLA 2010, pp. 823–827, 2010, doi: 10.1109/ICMLA.2010.127.

[68] M. Khalilia, S. Chakraborty, and M. Popescu, "Predicting disease risks from highly imbalanced data using random forest," BMC Med. Inform. Decis. Mak., vol. 11, no. 1, 2011, doi: 10.1186/1472-6947-11-51.

[69] M. Galar, A. Fernández, E. Barrenechea, and F. Herrera, "EUSBoost: Enhancing ensembles for highly imbalanced data-sets by evolutionary undersampling," Pattern Recognit., vol. 46, no. 12, pp. 3460–3471, 2013, doi: 10.1016/j.patcog.2013.05.006.

[70] S. Gónzalez, S. García, M. Lázaro, A. R. Figueiras-Vidal, and F. Herrera, "Class Switching according to Nearest Enemy Distance for learning from highly imbalanced data-sets," Pattern Recognit., vol. 70, pp. 12–24, 2017, doi: 10.1016/j.patcog.2017.04.028.

[71] S. Al-Azani and E. S. M. El-Alfy, "Using Word Embedding and Ensemble Learning for Highly Imbalanced Data Sentiment Analysis in Short Arabic Text," Procedia Comput. Sci., vol. 109, pp. 359–366, 2017, doi: 10.1016/j.procs.2017.05.365.

[72] P. Zyblewski, P. Ksieniewicz, and M. Woźniak, "Classifier Selection for Highly Imbalanced Data Streams with Minority Driven Ensemble," Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics), vol. 11508 LNAI, no. May, pp. 626–635, 2019, doi: 10.1007/978-3-030-20912-4_57.

[73] D. Chen, X. J. Wang, C. Zhou, and B. Wang, "The Distance-Based Balancing Ensemble Method for Data With a High Imbalance Ratio," IEEE Access, vol. 7, pp. 68940–68956, 2019, doi: 10.1109/ACCESS.2019.2917920.

[74] M. Abouelenien, Xiaohui Yuan, P. Duraisamy, and Xiaojing Yuan, "Improving classification performance for the minority class in highly imbalanced dataset using boosting," 2013, doi: 10.1109/icccnt.2012.6477850.

[75] N. Liu, X. Li, E. Qi, M. Xu, L. Li, and B. Gao, "A Novel Ensemble Learning Paradigm for Medical Diagnosis With Imbalanced Data," IEEE Access, vol. 8, pp. 171263–171280, 2020, doi: 10.1109/access.2020.3014362.

[76] P. Pławiak, M. Abdar, J. Pławiak, V. Makarenkov, and U. R. Acharya, "DGHNL: A new deep genetic hierarchical network of learners for prediction of credit scoring," Inf. Sci. (Ny)., vol. 516, no. April, pp. 401–418, 2020, doi: 10.1016/j.ins.2019.12.045.

[77] P. Pławiak, M. Abdar, and U. Rajendra Acharya, "Application of new deep genetic cascade ensemble of SVM classifiers to predict the Australian credit scoring," Appl. Soft Comput., vol. 84, p. 105740, 2019, doi: https://doi.org/10.1016/j.asoc.2019.105740.

[78] E. I. Altman, M. Iwanicz-Drozdowska, E. K. Laitinen, and A. Suvas, "A Race for Long Horizon Bankruptcy Prediction," Appl. Econ., vol. 52, no. 37, pp. 4092–4111, 2020, doi: 10.1080/00036846.2020.1730762.

[79] M. Soui, S. Smiti, M. W. Mkaouer, and R. Ejbali, "Bankruptcy Prediction Using Stacked Auto-Encoders," Appl. Artif. Intell., vol. 34, no. 1, pp. 80–100, 2020, doi: 10.1080/08839514.2019.1691849.

[80] Q. Zhu, H. Liu, J. Wang, S. Chen, P. Wen, and S. Wang, "Research of System Fault Diagnosis Method Based on Imbalanced Data," 2019 Progn. Syst. Heal. Manag. Conf. PHM-Qingdao 2019, pp. 1–5, 2019, doi: 10.1109/PHM-Qingdao46334.2019.8943068.

[81] E. Kaya, S. Korkmaz, M. A. Sahman, and A. C. Cinar, "DEBOHID: A differential evolution based oversampling approach for highly

imbalanced datasets," Expert Syst. Appl., vol. 169, p. 114482, 2021, doi: 10.1016/j.eswa.2020.114482.

[82] H. Faris et al., "Improving financial bankruptcy prediction in a highly imbalanced class distribution using oversampling and ensemble learning: a case from the Spanish market," Prog. Artif. Intell., vol. 9, no. 1, pp. 31–53, 2020, doi: 10.1007/s13748-019-00197-9.

[83] Z. Moti, S. Hashemi, and A. N. Jahromi, "A Deep Learning-based Malware Hunting Technique to Handle Imbalanced Data," pp. 48–53, 2020, doi: 10.1109/ISCISC51277.2020.9261913.

[84] Y. Qu, P. Quan, M. Lei, and Y. Shi, "Review of bankruptcy prediction using machine learning and deep learning techniques," Procedia Comput. Sci., vol. 162, no. Itqm 2019, pp. 895–899, 2019, doi: 10.1016/j.procs.2019.12.065.

[85] M. Duan, D. Liu, X. Chen, R. Liu, Y. Tan, and L. Liang, "Self-Balancing Federated Learning with Global Imbalanced Data in Mobile Systems," IEEE Trans. Parallel Distrib. Syst., vol. 32, no. 1, pp. 59–71, 2021, doi: 10.1109/TPDS.2020.3009406.

[86] P. du Jardin, "Forecasting bankruptcy using biclustering and neural network-based ensembles," Ann. Oper. Res., vol. 299, no. 1–2, pp. 531–566, 2021, doi: 10.1007/s10479-019-03283-2.

[87] P. S. Nitu, P. Madiraju, and F. A. Pintar, "Identifying Feature Pattern for Weighted Imbalance Data: A Feature Selection Study for Thoracolumbar Spine Fractures in Crash Injury Research," Proc. - 2020 IEEE 21st Int. Conf. Inf. Reuse Integr. Data Sci. IRI 2020, pp. 142–147, 2020, doi: 10.1109/IRI49571.2020.00028.

[88] U. Mahapatra, S. M. Nayak, and M. Rout, "A Systematic Approach to Enhance the Forecasting of Bankruptcy Data," Adv. Intell. Syst. Comput., vol. 1119, no. January, pp. 641–650, 2020, doi: 10.1007/978-981-15-2414-1_64.

[89] M. Zoričák, P. Gnip, P. Drotár, and V. Gazda, "Bankruptcy prediction for small- and medium-sized companies using severely imbalanced datasets," Econ. Model., vol. 84, no. April, pp. 165–176, 2020, doi: 10.1016/j.econmod.2019.04.003.

[90] T. Fitzpatrick and C. Mues, "An empirical comparison of classification algorithms for mortgage default prediction: Evidence from a distressed mortgage market," Eur. J. Oper. Res., vol. 249, no. 2, pp. 427–439, 2016, doi: 10.1016/j.ejor.2015.09.014.

[91] L. E. Boiko Ferreira, H. Murilo Gomes, A. Bifet, and L. S. Oliveira, "Adaptive Random Forests with Resampling for Imbalanced data Streams," Proc. Int. Jt. Conf. Neural Networks, vol. 2019-July, no. July, pp. 1–6, 2019, doi: 10.1109/IJCNN.2019.8852027.

[92] E. I. Altman, M. Iwanicz-Drozdowska, E. K. Laitinen, and A. Suvas, "A Race for Long Horizon Bankruptcy Prediction," Appl. Econ., vol. 00, no. 00, pp. 1–20, 2020, doi: 10.1080/00036846.2020.1730762.

[93] J. Sun, H. Li, H. Fujita, B. Fu, and W. Ai, "Class-imbalanced dynamic financial distress prediction based on Adaboost-SVM ensemble combined with SMOTE and time weighting," Inf. Fusion, vol. 54, no. July 2019, pp. 128–144, 2020, doi: 10.1016/j.inffus.2019.07.006.

[94] L. Yanze and O. Harutoshi, "Quasi-Linear SVM with Local Offsets for High-dimensional Imbalanced Data Classification," 2020.

[95] Y. Chen, "A selective under-sampling based bagging SVM for imbalanced data learning in biomedical event trigger recognition," ACM Int. Conf. Proceeding Ser., pp. 112–119, 2018, doi: 10.1145/3278198.3278221.

[96] Z. Luo, X. Zeng, Z. Bao, and M. Xu, "Deep learning-based strategy for macromolecules classification with imbalanced data from cellular electron cryotomography," arXiv, no. July, pp. 1–8, 2019.

[97] T. Handhika, A. Fahrurozi, R. I. M. Zen, D. P. Lestari, I. Sari, and Murni, "Modified average of the base-level models in the hill-climbing bagged ensemble selection algorithm for credit scoring," Procedia Comput. Sci., vol. 157, pp. 229–237, 2019, doi: 10.1016/j.procs.2019.08.162.

[98] Y. Xia, C. Liu, B. Da, and F. Xie, "A novel heterogeneous ensemble credit scoring model based on bstacking approach," Expert Syst. Appl., vol. 93, pp. 182–199, 2018, doi: https://doi.org/10.1016/j.eswa.2017.10.022.

[99] V. García, A. I. Marqués, and J. S. Sánchez, "Exploring the synergetic effects of sample types on the performance of ensembles for credit risk and corporate bankruptcy prediction," Inf. Fusion, vol. 47, no. June 2018, pp. 88–101, 2019, doi: 10.1016/j.inffus.2018.07.004.

[100] M. Schlögl, "A multivariate analysis of environmental effects on road accident occurrence using a balanced bagging approach," Accid. Anal. Prev., vol. 136, no. December 2019, p. 105398, 2020, doi: 10.1016/j.aap.2019.105398.

[101] L. Wang et al., "Classifying 2-year recurrence in patients with dlbcl using clinical variables with imbalanced data and machine learning methods," Comput. Methods Programs Biomed., vol. 196, p. 105567, 2020, doi: 10.1016/j.cmpb.2020.105567.

[102] Y. S. Jeon and D. J. Lim, "PSU: Particle Stacking Undersampling Method for Highly Imbalanced Big Data," IEEE Access, vol. 8, pp. 131920–131927, 2020, doi: 10.1109/ACCESS.2020.3009753.

[103] W. Wang, M. Zhang, L. Zhang, and Q. Bai, "Imbalanced Data Classification for Multi-Source Heterogenous Sensor Networks," IEEE Access, vol. 8, pp. 27406–27413, 2020, doi: 10.1109/ACCESS.2020.2966324.

[104] P. Branco and L. Torgo, "A study on the impact of data characteristics in imbalanced regression tasks," Proc. - 2019 IEEE Int. Conf. Data Sci. Adv. Anal. DSAA 2019, pp. 193–202, 2019, doi: 10.1109/DSAA.2019.00034.

[105] W. Feng et al., "Dynamic synthetic minority over-sampling technique-based rotation forest for the classification of imbalanced hyperspectral data," IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens., vol. 12, no. 7, pp. 2159–2169, 2019, doi: 10.1109/JSTARS.2019.2922297.

[106] S. Lyra, S. Leonhardt, and C. Hoog Antink, "Early Prediction of Sepsis Using Random Forest Classification for Imbalanced Clinical Data," 2019 Comput. Cardiol. Conf., vol. 45, pp. 1–4, 2019, doi: 10.22489/cinc.2019.276.

[107] M. Yahaya, X. Jiang, C. Fu, K. Bashir, and W. Fan, "Enhancing Crash Injury Severity Prediction on Imbalanced Crash Data by Sampling Technique with Variable Selection," 2019 IEEE Intell. Transp. Syst. Conf. ITSC 2019, pp. 363–368, 2019, doi: 10.1109/ITSC.2019.8917223.

[108] R. O'Brien and H. Ishwaran, "A random forests quantile classifier for class imbalanced data," Pattern Recognit., vol. 90, pp. 232–249, 2019, doi: 10.1016/j.patcog.2019.01.036.

[109] T. M. Shearman, J. M. Varner, S. M. Hood, C. A. Cansler, and J. K. Hiers, "Modelling post-fire tree mortality: Can random forest improve discrimination of imbalanced data?," Ecol. Modell., vol. 414, no. November, p. 108855, 2019, doi: 10.1016/j.ecolmodel.2019.108855.

[110] W. hui Hou, X. kang Wang, H. yu Zhang, J. qiang Wang, and L. Li, "A novel dynamic ensemble selection classifier for an imbalanced data set: An application for credit risk assessment," Knowledge-Based Syst., vol. 208, p. 106462, 2020, doi: 10.1016/j.knosys.2020.106462.

[111] Y. Deng, P. Dolog, J. M. Gass, and K. Denecke, "Obesity entity extraction from real outpatient records: When learning-based methods meet small imbalanced medical data sets," Proc. - IEEE Symp. Comput. Med. Syst., vol. 2019-June, pp. 411–416, 2019, doi: 10.1109/CBMS.2019.00087.

[112] J. M. Melo Neto, H. S. Bernardino, and H. J. C. Barbosa, "On the Impact of the Objective Function on Imbalanced Data using Cartesian Genetic Programming Neuroevolutionary Approaches," 2019 IEEE Congr. Evol. Comput. CEC 2019 - Proc., pp. 1860–1867, 2019, doi: 10.1109/CEC.2019.8789947.

[113] G. Wang, T. Zhou, K.-S. Choi, and J. Lu, "A Deep-Ensemble-Level-Based Interpretable Takagi-Sugeno-Kang Fuzzy Classifier for Imbalanced Data," IEEE Trans. Cybern., pp. 1–14, 2020, doi: 10.1109/tcyb.2020.3016972.

[114] B. Efron and R. J. Tibshirani, An Introduction to the Bootstrap. 1993.

[115] L. Breiman, "Bagging predictions," Mach. Learn., vol. 24, no. 2, pp. 123–140, 1996.

[116] G. Rekha, V. K. Reddy, A. K. Tyagi, and M. M. Nair, "Distance-based Bootstrap Sampling in Bagging for Imbalanced Data-Set," Int. Conf. Emerg. Trends Inf. Technol. Eng. ic-ETITE 2020, pp. 1–6, 2020, doi: 10.1109/ic-ETITE47903.2020.345.

[117] X. Wang, J. Xu, T. Zeng, and L. Jing, "Local distribution-based adaptive minority oversampling for imbalanced data classification," Neurocomputing, vol. 422, pp. 200–213, 2021, doi: 10.1016/j.neucom.2020.05.030.

[118] L. G. Valiant, “A theory of the learnable,” Proc. Annu. ACM Symp. Theory Comput., vol. 27, no. 11, pp. 436–445, 1984, doi: 10.1145/800057.808710.

[119] M. J. Kearns, R. E. Schapire, and L. M. Sellie, “Toward efficient agnostic learning,” Proc. Fifth Annu. ACM Work. Comput. Learn. Theory, vol. 141, pp. 341–352, 1992, doi: 10.1007/bf00993468.

[120] Y. Wang, L. L. Sun, and Q. Jin, “Enhanced Diagnosis of Pneumothorax with an Improved Real-time Augmentation for Imbalanced Chest X-rays Data Based on DCNN,” IEEE/ACM Trans. Comput. Biol. Bioinforma., vol. 14, no. 8, pp. 1–1, 2019, doi: 10.1109/tcbb.2019.2911947.

[121] X. Y. Jing et al., “Multiset Feature Learning for Highly Imbalanced Data Classification,” IEEE Trans. Pattern Anal. Mach. Intell., vol. 43, no. 1, pp. 139–156, 2021, doi: 10.1109/TPAMI.2019.2929166.

[122] D. Wolpert, “Stacked Generalization (Stacking),” Neural Networks, vol. 5, pp. 241–259, 1992.

[123] H. He, W. Zhang, and S. Zhang, “A novel ensemble method for credit scoring: Adaption of different imbalance ratios,” Expert Syst. Appl., vol. 98, pp. 105–117, 2018, doi: https://doi.org/10.1016/j.eswa.2018.01.012.

[124] S. Layeghian Javan, M. M. Sepehri, M. Layeghian Javan, and T. Khatibi, “An intelligent warning model for early prediction of cardiac arrest in sepsis patients,” Comput. Methods Programs Biomed., vol. 178, pp. 47–58, 2019, doi: 10.1016/j.cmpb.2019.06.010.

[125] L. I. Kuncheva and E. Alpaydin, Combining Pattern Classifiers: Methods and Algorithms, vol. 18, no. 3. 2007.

[126] M. D. Muhlbaier, A. Topalis, and R. Polikar, “Learn++ .NC: Combining ensemble of classifiers with dynamically weighted consult-and-vote for efficient incremental learning of new classes,” IEEE Trans. Neural Networks, vol. 20, no. 1, pp. 152–168, 2009, doi: 10.1109/TNN.2008.2008326.

[127] Garcia-Pedrajas, “Constructing Ensembles of Classifiers by Means of Weighted Instance Selection,” vol. 20, no. 2, pp. 258–277, 2009.

[128] A. Rahman and B. Verma, “Novel layered clustering-based approach for generating ensemble of classifiers,” IEEE Trans. Neural Networks, vol. 22, no. 5, pp. 781–792, 2011, doi: 10.1109/TNN.2011.2118765.

[129] D. Opitz and R. Maclin, “Popular Ensemble Methods: An Empirical Study,” J. Artif. Intell. Res., vol. 11, no. April, pp. 169–198, 1999, doi: 10.1613/jair.614.

[130] T. G. Dietterich, “Ensemble methods in machine learning,” Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics), vol. 1857 LNCS, pp. 1–15, 2000, doi: 10.1007/3-540-45014-9_1.

[131] R. Polikar, “Ensemble based systems in decision making,” IEEE Circuits Syst. Mag., vol. 6, no. 3, pp. 21–44, 2006, doi: 10.1109/MCAS.2006.1688199.

[132] L. Rokach, “Ensemble-based classifiers,” Artif. Intell. Rev., vol. 33, no. 1–2, pp. 1–39, 2010, doi: 10.1007/s10462-009-9124-7.

[133] L. I. Kuncheva and L. C. Jain, “Designing classifier fusion systems by genetic algorithms,” IEEE Trans. Evol. Comput., vol. 4, no. 4, pp. 327–336, 2000, doi: 10.1109/4235.887233.

[134] M. L. Raymer, W. F. Punch, E. D. Goodman, L. A. Kuhn, and A. K. Jain, “Dimensionality reduction using genetic algorithms,” IEEE Trans. Evol. Comput., vol. 4, no. 2, pp. 164–171, 2000, doi: 10.1109/4235.850656.

[135] M. Chabbouh, S. Bechikh, C. C. Hung, and L. Ben Said, “Multi-objective evolution of oblique decision trees for imbalanced data binary classification,” Swarm Evol. Comput., vol. 49, no. May 2018, pp. 1–22, 2019, doi: 10.1016/j.swevo.2019.05.005.

[136] A. Bequé and S. Lessmann, “Extreme learning machines for credit scoring: An empirical evaluation,” Expert Syst. Appl., vol. 86, pp. 42–53, 2017, doi: 10.1016/j.eswa.2017.05.050.

[137] M. Abdar, W. Książek, U. R. Acharya, R. S. Tan, V. Makarenkov, and P. Pławiak, “A new machine learning technique for an accurate diagnosis of coronary artery disease,” Comput. Methods Programs Biomed., vol. 179, no. August, 2019, doi: 10.1016/j.cmpb.2019.104992.

[138] Y. Liu, Y. Wang, X. Ren, H. Zhou, and X. Diao, “A Classification Method Based on Feature Selection for Imbalanced Data,” IEEE Access, vol. 7, pp. 81794–81807, 2019, doi: 10.1109/ACCESS.2019.2923846.

[139] W. C. Lin, Y. H. Lu, and C. F. Tsai, “Feature selection in single and ensemble learning-based bankruptcy prediction models,” Expert Syst., vol. 36, no. 1, pp. 1–8, 2019, doi: 10.1111/exsy.12335.

[140] H. Faris et al., “Improving financial bankruptcy prediction in a highly imbalanced class distribution using oversampling and ensemble learning: a case from the Spanish market,” Prog. Artif. Intell., vol. 9, no. 1, pp. 31–53, 2020, doi: 10.1007/s13748-019-00197-9.

[141] T. Pisula, “An Ensemble Classifier-Based Scoring Model for Predicting Bankruptcy of Polish Companies in the Podkarpackie Voivodeship,” J. Risk Financ. Manag., vol. 13, no. 2, p. 37, 2020, doi: 10.3390/jrfm13020037.

[142] I. Hermadi, Y. Nurhadryani, I. Ranggadara, and R. Amin, “A Review of Contribution and Challenge in Predictive Machine Learning Model at Financial Industry,” J. Phys. Conf. Ser., vol. 1477, no. 3, pp. 5–9, 2020, doi: 10.1088/1742-6596/1477/3/032021.

[143] Y. C. Chang, K. H. Chang, and G. J. Wu, “Application of eXtreme gradient boosting trees in the construction of credit risk assessment models for financial institutions,” Appl. Soft Comput. J., vol. 73, pp. 914–920, 2018, doi: 10.1016/j.asoc.2018.09.029.

[144] W. Lei, R. Zhang, Y. Yang, R. Wang, and W. S. Zheng, “Class-Center Involved Triplet Loss for Skin Disease Classification on Imbalanced Data,” Proc. - Int. Symp. Biomed. Imaging, vol. 2020-April, pp. 16–20, 2020, doi: 10.1109/ISBI45749.2020.9098718.

[145] J. F. Ramirez Rochac, N. Zhang, L. Thompson, and T. Oladunni, “A Data Augmentation-Assisted Deep Learning Model for High Dimensional and Highly Imbalanced Hyperspectral Imaging Data,” 9th Int. Conf. Inf. Sci. Technol. ICIST 2019, pp. 362–367, 2019, doi: 10.1109/ICIST.2019.8836913.

[146] B. Jonathan, P. H. Putra, and Y. Ruldeviyani, “Observation Imbalanced Data Text to Predict Users Selling Products on Female Daily with SMOTE, Tomek, and SMOTE-Tomek,” Proc. - 2020 IEEE Int. Conf. Ind. 4.0, Artif. Intell. Commun. Technol. IAICT 2020, pp. 81–85, 2020, doi: 10.1109/IAICT50021.2020.9172033.

[147] M. Z. Alom et al., “A state-of-the-art survey on deep learning theory and architectures,” Electron., vol. 8, no. 3, pp. 1–67, 2019, doi: 10.3390/electronics8030292.

[148] M. Azizjon, A. Jumabek, and W. Kim, “1D CNN based network intrusion detection with normalization on imbalanced data,” 2020 Int. Conf. Artif. Intell. Inf. Commun. ICAIIC 2020, pp. 218–224, 2020, doi: 10.1109/ICAIIC48513.2020.9064976.

[149] T. Hosaka, “Bankruptcy prediction using imaged financial ratios and convolutional neural networks,” Expert Syst. Appl., vol. 117, pp. 287–299, 2019, doi: 10.1016/j.eswa.2018.09.039.

[150] F. Mai, S. Tian, C. Lee, and L. Ma, “Deep learning models for bankruptcy prediction using textual disclosures,” Eur. J. Oper. Res., vol. 274, no. 2, pp. 743–758, 2019, doi: 10.1016/j.ejor.2018.10.024.

[151] C. Clement, “Machine Learning in Bankruptcy Prediction,” J. Public Adm. Financ. Law, no. 17, p. 20, 2020.

[152] H. Dong, B. Zhu, and J. Zhang, “A Cost-sensitive Active Learning for Imbalance Data with Uncertainty and Diversity Combination,” ACM Int. Conf. Proceeding Ser., pp. 218–224, 2020, doi: 10.1145/3383972.3384002.

[153] N. W. S. Wardhani, M. Y. Rochayani, A. Iriany, A. D. Sulistyono, and P. Lestantyo, “Cross-validation Metrics for Evaluating Classification Performance on Imbalanced Data,” 2019 Int. Conf. Comput. Control. Informatics its Appl. Emerg. Trends Big Data Artif. Intell. IC3INA 2019, pp. 14–18, 2019, doi: 10.1109/IC3INA48034.2019.8949568.

[154] L. B. Laureano, A. M. Sison, and R. P. Medina, “Handling imbalanced data through affinity propagation and SMOTE,” ACM Int. Conf. Proceeding Ser., pp. 22–26, 2019, doi: 10.1145/3366650.3366665.

[155] B. Krawczyk, M. Koziarski, and M. Wozniak, “Radial-based oversampling for multiclass imbalanced data classification,” IEEE Trans. Neural Networks Learn. Syst., vol. 31, no. 8, pp. 2818–2831, 2020, doi: 10.1109/TNNLS.2019.2913673.

[156] T. M. Khoshgoftaar, K. Gao, and J. Van Hulse, “A novel feature selection technique for highly imbalanced data,” 2010 IEEE Int. Conf. Inf. Reuse Integr. IRI 2010, pp. 80–85, 2010, doi: 10.1109/IRI.2010.5558961.

[157] T. M. Khoshgoftaar, K. Gao, and A. Napolitano, “Exploring an iterative feature selection technique for highly imbalanced data sets,” Proc. 2012 IEEE 13th Int. Conf. Inf. Reuse Integr. IRI 2012, pp. 101–108, 2012, doi: 10.1109/IRI.2012.6302997.

[158] C. L. Calvert and T. M. Khoshgoftaar, “Threshold based optimization of performance metrics with severely imbalanced big security data,” Proc. - Int. Conf. Tools with Artif. Intell. ICTAI, vol. 2019-Novem, pp. 1328–1334, 2019, doi: 10.1109/ICTAI.2019.00184.

[159] S. Lu, F. Gao, C. Piao, and Y. Ma, “Dynamic Weighted Cross Entropy for Semantic Segmentation with Extremely Imbalanced Data,” Proc. - 2019 Int. Conf. Artif. Intell. Adv. Manuf. AIAM 2019, pp. 230–233, 2019, doi: 10.1109/AIAM48774.2019.00053.

[160] A. Fernández, F. J. Berlanga, M. J. Del Jesus, and F. Herrera, “Genetic cooperative-competitive fuzzy rule based learning method using genetic programming for highly imbalanced data-sets,” 2009 Int. Fuzzy Syst. Assoc. World Congr. 2009 Eur. Soc. Fuzzy Log. Technol. Conf. IFSA-EUSFLAT 2009 - Proc., pp. 42–47, 2009.

[161] M. Uriz, D. Paternain, H. Bustince, and M. Galar, “A first approach towards the usage of classifiers’ performance to create fuzzy measures for ensembles of classifiers: A case study on highly imbalanced datasets,” IEEE Int. Conf. Fuzzy Syst., vol. 2018-July, pp. 1–8, 2018, doi: 10.1109/FUZZ-IEEE.2018.8491440.

[162] M. Uriz, D. Paternain, H. Bustince, and M. Galar, “An empirical study on supervised and unsupervised fuzzy measure construction methods in highly imbalanced classification,” IEEE Int. Conf. Fuzzy Syst., vol. 2020-July, pp. 1–8, 2020, doi: 10.1109/FUZZ48607.2020.9177789.

[163] E. Bringer, A. Israeli, Y. Shoham, A. Ratner, and C. Ré, “Osprey: Weak Supervision of Imbalanced Extraction Problems without Code,” Proc. ACM SIGMOD Int. Conf. Manag. Data, 2019, doi: 10.1145/3329486.3329492.

[164] W. Ding, D. Y. Huang, Z. Chen, X. Yu, and W. Lin, “Facial action recognition using very deep networks for highly imbalanced class distribution,” Proc. - 9th Asia-Pacific Signal Inf. Process. Assoc. Annu. Summit Conf. APSIPA ASC 2017, vol. 2018-Febru, no. December, pp. 1368–1372, 2018, doi: 10.1109/APSIPA.2017.8282246.

[165] E. Kim and S. Bang, “Weighted L 1 -Norm Support Vector Machine for the,” vol. 28, pp. 9–21, 2015.

[166] A. Fernández, M. J. Del Jesus, and F. Herrera, “Improving the performance of fuzzy rule based classification systems for highly imbalanced data-sets using an evolutionary adaptive inference system,” Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics), vol. 5517 LNCS, no. PART 1, pp. 294–301, 2009, doi: 10.1007/978-3-642-02478-8_37.

[167] P. Villar, A. FernÁndez, R. A. Carrasco, and F. Herrera, “Feature selection and granularity learning in genetic fuzzy rule-based classification systems for highly imbalanced data-sets,” Int. J. Uncertainty, Fuzziness Knowlege-Based Syst., vol. 20, no. 3, pp. 369–397, 2012, doi: 10.1142/S0218488512500195.

[168] B. Liu, M. Zhang, W. Ma, X. Li, Y. Liu, and S. Ma, “A two-step information accumulation strategy for learning from highly imbalanced data,” Int. Conf. Inf. Knowl. Manag. Proc., vol. Part F1318, pp. 1289–1298, 2017, doi: 10.1145/3132847.3132940.

[169] P. Villar, A. Fernández, and F. Herrera, “A genetic algorithm for feature selection and granularity learning in fuzzy rule-based classification systems for highly imbalanced data-sets,” Commun. Comput. Inf. Sci., vol. 80 PART 1, pp. 741–750, 2010, doi: 10.1007/978-3-642-14055-6_78.

[170] Y. Tang, Y. Q. Zhang, and N. V. Chawla, “SVMs modeling for highly imbalanced classification,” IEEE Trans. Syst. Man, Cybern. Part B Cybern., vol. 39, no. 1, pp. 281–288, 2009, doi: 10.1109/TSMCB.2008.2002909.

[171] P. Villar, A. Fernández, and F. Herrera, “A genetic learning of the fuzzy rule-based classification system granularity for highly imbalanced data-sets,” IEEE Int. Conf. Fuzzy Syst., pp. 1689–1694, 2009, doi: 10.1109/FUZZY.2009.5277304.

[172] S. Sardari, M. Eftekhari, and F. Afsari, “Hesitant fuzzy decision tree approach for highly imbalanced data classification,” Appl. Soft Comput. J., vol. 61, pp. 727–741, 2017, doi: 10.1016/j.asoc.2017.08.052.

[173] H. Ahumada, G. L. Grinblat, L. C. Uzal, P. M. Granitto, and A. Ceccatto, “REPMAC: A new hybrid approach to highly imbalanced classification problems,” Proc. - 8th Int. Conf. Hybrid Intell. Syst. HIS 2008, pp. 386–391, 2008, doi: 10.1109/HIS.2008.142.

[174] A. Anand, G. Pugalenthi, G. B. Fogel, and P. N. Suganthan, “An approach for classification of highly imbalanced data using weighting and undersampling,” Amino Acids, vol. 39, no. 5, pp. 1385–1391, 2010, doi: 10.1007/s00726-010-0595-2.

[175] Q. Li and Y. Xie, “A behavior-cluster based imbalanced classification method for credit card fraud detection,” ACM Int. Conf. Proceeding Ser., pp. 134–139, 2019, doi: 10.1145/3352411.3352433.

[176] B. Gyawali, T. Solorio, M. Montes-Y-Gómez, B. Wardman, and G. Warner, “Evaluating a semisupervised approach to phishing URL identification in a realistic scenario,” ACM Int. Conf. Proceeding Ser., pp. 176–183, 2011, doi: 10.1145/2030376.2030397.

[177] S. R. Hashemi, S. S. M. Salehi, D. Erdogmus, S. P. Prabhu, S. K. Warfield, and A. Gholipour, “Asymmetric Loss Functions and Deep Densely-Connected Networks for Highly-Imbalanced Medical Image Segmentation: Application to Multiple Sclerosis Lesion Detection,” IEEE Access, vol. 7, pp. 1721–1735, 2019, doi: 10.1109/ACCESS.2018.2886371.

[178] Y. Lu, J. H. Zhou, and C. Guan, “Minimizing Hybrid Dice Loss for Highly Imbalanced 3D Neuroimage Segmentation,” Proc. Annu. Int. Conf. IEEE Eng. Med. Biol. Soc. EMBS, vol. 2020-July, pp. 1059–1062, 2020, doi: 10.1109/EMBC44109.2020.9176663.

[179] P. Jain, A. Agarwal, and R. Behara, “An approach to supervised classification of highly imbalanced and high dimensionality COPD readmission data on HPCC,” SysCon 2019 - 13th Annu. IEEE Int. Syst. Conf. Proc., pp. 5–9, 2019, doi: 10.1109/SYSCON.2019.8836797.

[180] C. F. Yeh, A. Heidel, H. Y. Lee, and L. S. Lee, “Recognition of highly imbalanced code-mixed bilingual speech with frame-level language detection based on blurred posteriorgram,” ICASSP, IEEE Int. Conf. Acoust. Speech Signal Process. - Proc., pp. 4873–4876, 2012, doi: 10.1109/ICASSP.2012.6289011.

[181] C. F. Yeh and L. S. Lee, “An improved framework for recognizing highly imbalanced bilingual code-switched lectures with cross-language acoustic modeling and frame-level language identification,” IEEE Trans. Audio, Speech Lang. Process., vol. 23, no. 7, pp. 1144–1159, 2015, doi: 10.1109/TASLP.2015.2425214.

[182] Y. Tang, S. Krasser, P. Judge, and Y. Q. Zhang, “Fast and effective spam sender detection with granular SVM on highly imbalanced mail server behavior data,” 2006 Int. Conf. Collab. Comput. Networking, Appl. Work. Collab., 2006, doi: 10.1109/COLCOM.2006.361856.

[183] I. Almomani et al., “Android Ransomware Detection Based on a Hybrid Evolutionary Approach in the Context of Highly Imbalanced Data,” IEEE Access, vol. 9, pp. 57674–57691, 2021, doi: 10.1109/ACCESS.2021.3071450.

[184] M. Abouelenien, Xiaohui Yuan, P. Duraisamy, and Xiaojing Yuan, “Improving classification performance for the minority class in highly imbalanced dataset using boosting,” 2012 Third Int. Conf. Comput. Commun. Netw. Technol., pp. 1–6, 2013, doi: 10.1109/icccnt.2012.6477850.

[185] M. Bach, A. Werner, J. Żywiec, and W. Pluskiewicz, “The study of under- and over-sampling methods’ utility in analysis of highly imbalanced data on osteoporosis,” Inf. Sci. (Ny)., vol. 384, pp. 174–190, 2017, doi: 10.1016/j.ins.2016.09.038.

[186] J. C. Lamirel, “Dealing with highly imbalanced textual data gathered into similar classes,” Proc. Int. Jt. Conf. Neural Networks, pp. 1–7, 2013, doi: 10.1109/IJCNN.2013.6707044.

[187] J. Guan, J. Liu, J. Sun, P. Feng, T. Shuai, and W. Wang, “META METRIC LEARNING FOR HIGHLY IMBALANCED AERIAL SCENE CLASSIFICATION College of Computer Science and Technology , Harbin Engineering University , Harbin , 150001 , China State Key Laboratory of Space-Ground Integrated Information Technology , Beijing , ,” ICASSP 2020 - 2020 IEEE Int. Conf. Acoust. Speech Signal Process., pp. 4042–4046, 2020.

[188] F. Wu, X. Y. Jing, S. Shan, W. Zuo, and J. Y. Yang, “Multiset feature learning for highly imbalanced data classification,” 31st AAAI Conf.

Artif. Intell. AAAI 2017, vol. PP, no. c, pp. 1583–1589, 2017, doi: 10.1109/tpami.2019.2929166.

[189] L. V. B. Beltrán, M. Coustaty, N. Journet, J. C. Caicedo, and A. Doucet, "Multi-attribute learning with highly imbalanced data," Proc. - Int. Conf. Pattern Recognit., pp. 9219–9226, 2020, doi: 10.1109/ICPR48806.2021.9412634.

[190] A. N. Jahromi, H. Karimpour, J. Sakhnini, and A. Dehghantanha, "A deep unsupervised representation learning approach for effective cyber-physical attack detection and identification on highly imbalanced data," CASCON 2019 Proc. - Conf. Cent. Adv. Stud. Collab. Res. - Proc. 29th Annu. Int. Conf. Comput. Sci. Softw. Eng., pp. 14–23, 2020.

[191] S. Li and W. Deng, "Real world expression recognition: A highly imbalanced detection problem," 2016 Int. Conf. Biometrics, ICB 2016, pp. 1–6, 2016, doi: 10.1109/ICB.2016.7550074.

[192] A. C. Neocleous, K. H. Nicolaides, and C. N. Schizas, "Intelligent Noninvasive Diagnosis of Aneuploidy: Raw Values and Highly Imbalanced Dataset," IEEE J. Biomed. Heal. Informatics, vol. 21, no. 5, pp. 1271–1279, 2017, doi: 10.1109/JBHI.2016.2608859.

[193] Y. B. Hagos et al., "Cell abundance aware deep learning for cell detection on highly imbalanced pathological data," Proc. - Int. Symp. Biomed. Imaging, vol. 2021-April, pp. 1438–1442, 2021, doi: 10.1109/ISBI48211.2021.9433994.

[194] J. C. Lamirel, "Dealing with highly imbalanced textual data gathered into similar classes," Proc. Int. Jt. Conf. Neural Networks, 2013, doi: 10.1109/IJCNN.2013.6707044.

[195] S. Kübler, C. Liu, and Z. A. Sayyed, To use or not to use: Feature selection for sentiment analysis of highly imbalanced data, vol. 24, no. 1. 2018.

[196] N. Anantrasirichai, D. Bull, and D. Bull, "DEFECTNET : MULTI-CLASS FAULT DETECTION ON HIGHLY-IMBALANCED DATASETS N . Anantrasirichai and David Bull," 2019 IEEE Int. Conf. Image Process., pp. 2481–2485, 2019.

[197] P. Jain, A. Agarwal, and R. Behara, "An approach to supervised classification of highly imbalanced and high dimensionality COPD readmission data on HPCC," SysCon 2019 - 13th Annu. IEEE Int. Syst. Conf. Proc., pp. 1–7, 2019, doi: 10.1109/SYSCON.2019.8836797.

[198] K. Fujiwara, M. Shigeno, and U. Sumita, "A New Approach for Developing Segmentation Algorithms for Strongly Imbalanced Data," IEEE Access, vol. 7, pp. 82970–82977, 2019, doi: 10.1109/ACCESS.2019.2923524.

[199] Y. Lu, Y. M. Cheung, and Y. Y. Tang, "Bayes Imbalance Impact Index: A Measure of Class Imbalanced Data Set for Classification Problem," IEEE Trans. Neural Networks Learn. Syst., vol. 31, no. 9, pp. 3525–3539, 2020, doi: 10.1109/TNNLS.2019.2944962.

[200] E. R. Q. Fernandes, A. C. P. L. F. De Carvalho, and X. Yao, "Ensemble of classifiers based on multiobjective genetic sampling for imbalanced data," IEEE Trans. Knowl. Data Eng., vol. 32, no. 6, pp. 1104–1115, 2020, doi: 10.1109/TKDE.2019.2898861.

[201] K. Yang et al., "Hybrid Classifier Ensemble for Imbalanced Data," IEEE Trans. Neural Networks Learn. Syst., vol. 31, no. 4, pp. 1387–1400, 2020, doi: 10.1109/TNNLS.2019.2920246.

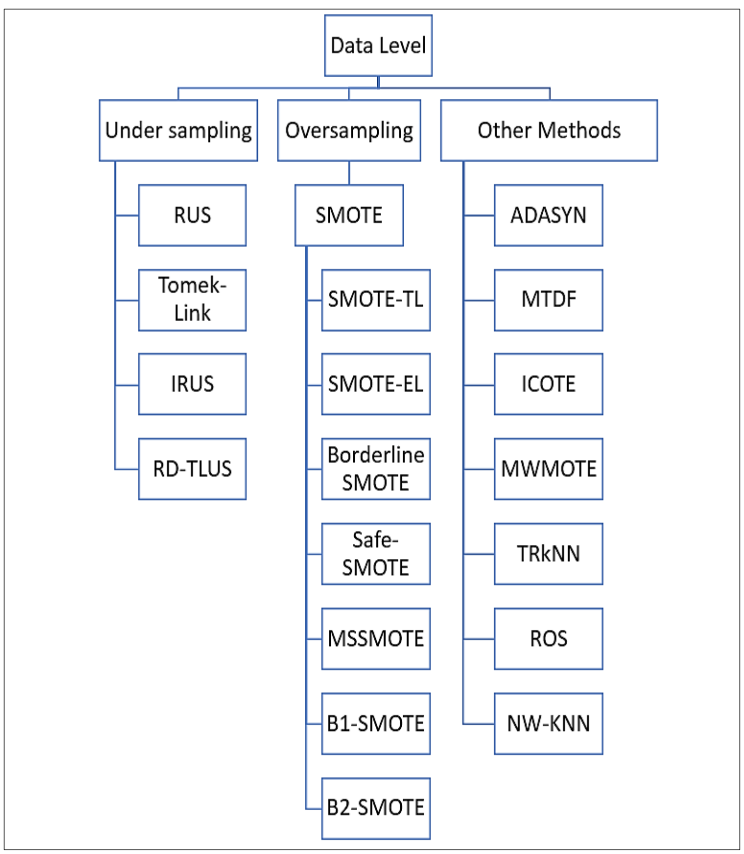
[202] G. Wang, J. Ma, L. Huang, and K. Xu, "Two credit scoring models based on dual strategy ensemble trees," Knowledge-Based Syst., vol. 26, pp. 61–68, 2012, doi: 10.1016/j.knosys.2011.06.020.

[203] E. Rendón, R. Alejo, C. Castorena, F. J. Isidro-Ortega, and E. E. Granda-Gutiérrez, "Data sampling methods to dealwith the big data multi-class imbalance problem," Appl. Sci., vol. 10, no. 4, 2020, doi: 10.3390/app10041276.

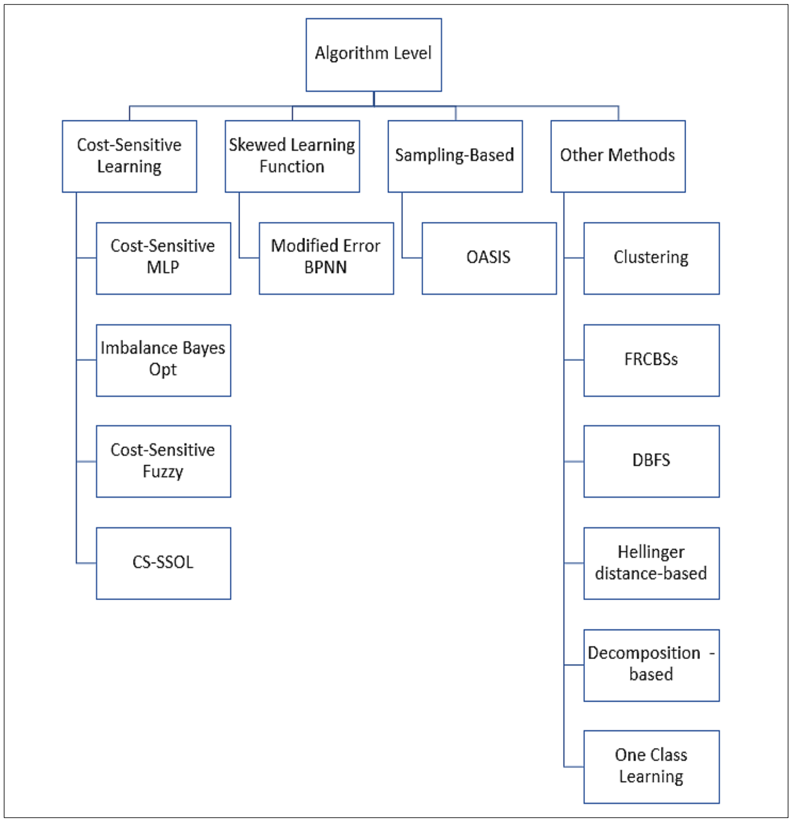
[204] C. L. Liu and P. Y. Hsieh, "Model-Based Synthetic Sampling for Imbalanced Data," IEEE Trans. Knowl. Data Eng., vol. 32, no. 8, pp. 1543–1556, 2020, doi: 10.1109/TKDE.2019.2905559.

[205] G. Shi, C. Feng, W. Xu, L. Liao, and H. Huang, "Penalized multiple distribution selection method for imbalanced data classification," Knowledge-Based Syst., vol. 196, p. 105833, 2020, doi: 10.1016/j.knosys.2020.105833.

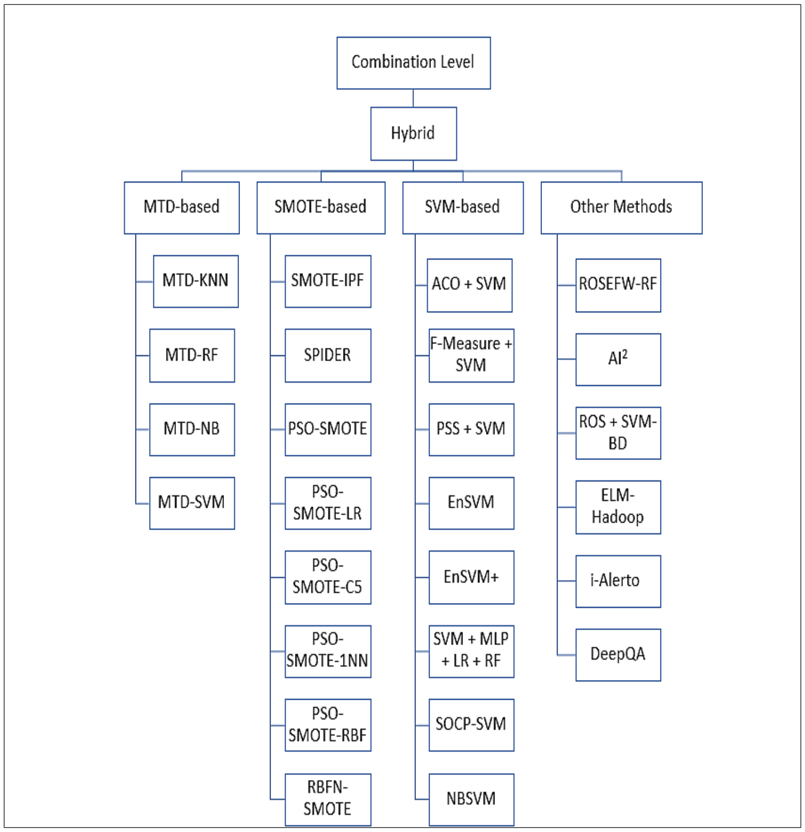
Appendix. I. METHODS IN DATA LEVEL STRATEGY



Appendix. II. METHODS IN ALGORITHM LEVEL STRATEGY



Appendix. III. METHODS IN COMBINATION LEVEL STRATEGY (HYBRID)



Appendix. IV. METHODS IN COMBINATION LEVEL STRATEGY (ENSEMBLE)