Penerbit UTHM

JOURNAL OF APPLIED SCIENCES TECHNOLOGY AND COMPUTING
e-ISSN: 3036-0250

JASTEC

Vol. 1 No. 2 (2024) p. 20-27
https://publisher.uthm.edu.my/ojs/index.php/jastec

# Handling Imbalanced Datasets in Machine Learning: Challenges, Approaches, and Best Practices

**Rusma Anieza Ruslan[1], Nureize Arbaiy[1]**

[1] *Faculty of Computer Science and Information Technology,*
*Universiti Tun Hussein Onn Malaysia, Parit Raja, Batu Pahat, 86400, MALAYSIA*

*Corresponding Author: hi230027@student.uthm.edu.my

DOI: https://doi.org/10.30880/jastec.2024.01.02.003

**Article Info**

Received: 2 August 2024
Accepted: 1 October 2024
Available online: 12 November 2024

**Keywords**

Imbalanced dataset, machine learning, resampling, ensemble method, cost sensitive laerning

**Abstract**

Determining the performance of a machine learning model is usually about the model's ability to make accurate predictions, which is assessed using an accuracy measure. However, other characteristics such as the quality and balance of the data must also be examined. Models may tend to make certain predictions that provide a high percentage of accurate predictions but have poor overall performance. There are balanced and imbalanced data situations in the dataset. An imbalanced dataset is a dataset that contains a minority class with a limited sample compared to the majority class. This makes it more likely that the model will favor the majority class, resulting in biased predictions and poor performance for the minority class. Therefore, it is important to remove the imbalance between the classes so that the model can make more accurate predictions. Several methods to solve this problem can be found in the literature, including the resampling method. Therefore, in this study, a comprehensive overview of techniques for dealing with imbalanced datasets in machine learning is given. This technique includes resampling techniques, ensemble methods and cost-sensitive learning. The study concludes that all techniques can be effective strategies for dealing with imbalanced class set problems and improving the performance of classification models in different domains.

## 1. Introduction

Machine learning (ML) is a subfield of artificial intelligence (AI) that focuses on the development of models and algorithms that can recognize patterns in data to make predictions and decisions [1, 3] without being explicitly programmed. There are different types of ML, including supervised, unsupervised, semi-supervised and reinforcement learning [2]. ML algorithms learn from data, and datasets play a crucial role in the development and performance of ML models. Datasets consist of rows, which represent observations, and columns, which represent features or characteristics. Datasets can be numeric, categorical, univariate, multivariate or time series and are essential for training, testing and evaluating ML models.

Datasets are a fundamental element of machine learning, as the quality and characteristics of the dataset significantly affect the performance and generalizability of ML models. Datasets can be labeled, i.e. each instance is associated with a corresponding label or target value, or unlabeled, i.e. the dataset has no predefined labels. Labeled datasets are used in supervised learning, while unlabeled datasets are used in unsupervised learning. Datasets can also be structured, unstructured, text-based or image-based and require different processing and analysis techniques. Each dataset has different characteristics in terms of size, complexity and balance.

The balance of datasets is divided into two categories: balanced datasets and imbalanced datasets. In a balanced dataset, each class is equally represented whether positive or negative, while an imbalanced dataset has an uneven distribution with underrepresented classes. The balance of the dataset balance is critical to the performance of each model in machine learning. If a minority class is significantly underrepresented compared to a majority class, this is significant [4]. A particular challenge when working with datasets is the problem of imbalanced datasets, where the distribution of classes is skewed and one class is being significantly underrepresented compared to the other. This can lead to biased and inaccurate predictions, especially in critical applications such as fraud detection [8, 9] and medical diagnosis [10, 11, 12], where accurate prediction of the minority class is crucial [5]. Several techniques have been developed to address this problem. These include resampling methods, such as oversampling the minority class or undersampling the majority class [6], as well as class weighting and ensemble methods.

This paper is organized as follows: Section 1 provides background information on imbalanced datasets in machine learning. Section 2 gives an overview of the techniques used to deal with class imbalances Section 3 discusses the results of experiments using the researcher's methods. The last section contains the conclusion.

## 2. Literature Review

A dataset is considered imbalanced if the number of input samples varies for each output or target class. Simply put, an imbalanced dataset is a skewed and uneven distribution of classes, which can pose various challenges for machine learning [4]. For example, if the dataset contains a smaller number of samples of the minority class than the majority class, then the machine learning model will tend to favour the majority class. As a result, the performance of the minority class is weak and the predictions are inaccurate or biassed [7]. This happens because the model prefers to learn patterns and make predictions for the majority class because it is better represented than the minority. For some real-world applications, such as fraud detection [8, 9] and medical diagnosis [10, 11, 12], accurate prediction of the minority class is important. This is because there are usually fewer fraudulent transactions than non-fraudulent ones. Therefore, these rare fraudulent activities need to be identified more accurately to avoid financial losses. Just like a medical diagnosis, doctors need to accurately recognise rare diseases to ensure the patient's well-being and receive timely treatment. If the imbalance in records is not addressed, it can lead to adverse effects such as losses from fraudulent transactions and misdiagnosis of life-threatening diseases.

An imbalanced dataset refers to a group of datasets with unequal class samples [13]. It can also be interpreted to mean that the number of minority and majority classes is unequal [14]. The majority class is the most abundant class, while the minority class is the least abundant class. Class imbalances in the dataset can affect the performance of the algorithm, leading to biased and inaccurate predictions. Therefore, there are several methods to deal with data imbalances, including resampling [15].

**Table 1** *Existing methods in handling imbalance class*

| Methods | Title | Contribution |
|---|---|---|
| Resampling [16] | Handling Class Imbalance In Credit Card Fraud Using Resampling Methods | The classification models performed better than the initial imbalanced dataset, were more realistic, and showed statistically significant improvements |
| Ensemble Methos [17] | Deep Neural Network Ensemble for the Intelligent Fault Diagnosis of Machines Under Imbalanced Data | An ensemble convolutional neural network (EnCNN) approach, which combines multiple base classifiers trained on balanced subsets of an imbalanced dataset using under-sampling, with weighted voting for the final prediction |
| Cost-Sensitive Learning [18] | Using Cost-Sensitive Learning and Feature Selection Algorithms to Improve the Performance of Imbalanced Classification | A Cost-sensitive Feature selection General Vector Machine (CFGVM) algorithm that combines the General Vector Machine and Binary Ant Lion Optimizer algorithms |

Table 1 shows three approaches for dealing with imbalanced datasets: Resampling, ensemble methods, and cost-sensitive learning. Each approach to dealing with imbalanced datasets has its strengths and contributes to improving the performance of classification models on imbalanced datasets. Furthermore, this approach demonstrates the importance of addressing this challenge in different applications.

## 2.1 Resampling Method

Resampling is a technique that can be used to correct data imbalances in machine learning by adjusting the number of representatives in the majority and minority classes [19]. Resampling is divided into two main approaches, as shown in Figure 1: Undersampling reduces the representation of the majority class and oversampling duplicates the representation of the minority class.

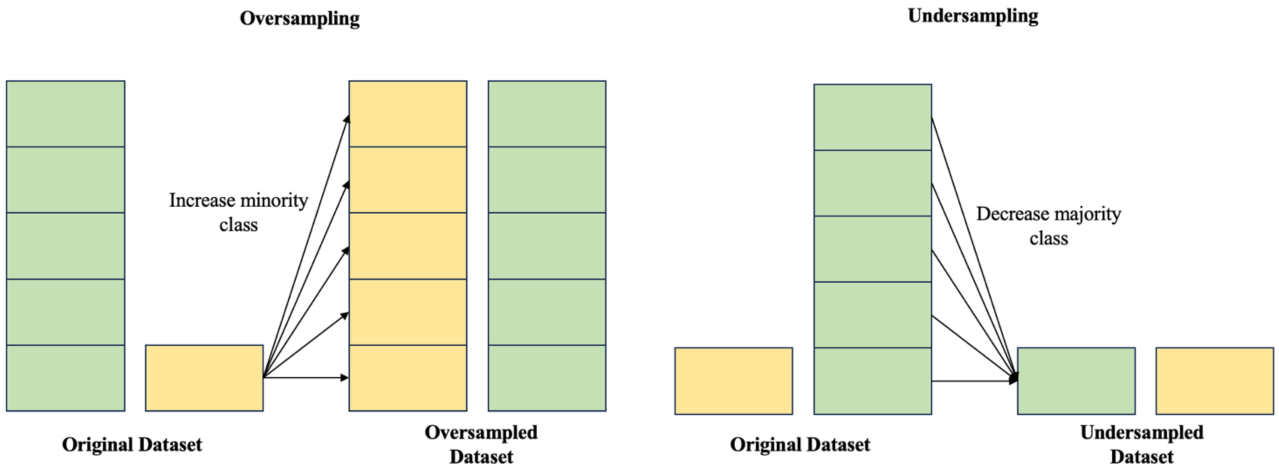


**Figure 1** *Difference between oversampling and undersampling*

Oversampling involves adding data from minority classes. However, oversampling can lead to overfitting as it only replicates instances from minority classes without adding new information. Oversampling is usually used with training data. Other approaches that have the same function as oversampling include SMOTE, ADASYN and data augmentation. SMOTE is a technique that generates new synthetic samples for minority classes [20]. The artificial samples generated in SMOTE are consistent across the feature space. However, this can lead to the feature space being crowded with synthetic SMOTE samples. ADASYN, on the other hand, focuses primarily on minority samples, which are challenging to classify accurately. It uses a weighted distribution to select minority classes that differ according to their difficulty in learning. In other words, ADASYN generates synthetic samples of minority classes that are close to the decision environment and thus more challenging.

Undersampling reduces the data in the majority class. However, undersampling can result in important information being lost as instances are removed from the majority class. This method can be applied directly to the training dataset to adapt the machine learning model. Near-miss undersampling is another method that serves the same purpose as undersampling. When using near-miss, the examples are selected according to how far the majority class example is from the minority class example. Near Miss is divided into three versions, which are determined by the Euclidean distance [21]:

i. NearMiss-1: The example from the majority class that is closest on average to the three examples from the minority class, has the smallest distance.

ii. NearMiss-2: The example from the majority class is closest on average to the three examples from the minority.

iii. NearMiss-3: The instance from the majority class is closest to each instance from the minority class.

**Table 2** *Application of resampling technique*

| Title | Contribution | Findings |
|---|---|---|
| Handling Class Imbalance in Credit Card Fraud using Resampling Methods [16] | Addressing the significant problem of class imbalance in credit card fraud detection, which has an impact on the effectiveness of current classification techniques | The classification models performed better than the initial imbalanced dataset, were more realistic, and showed statistically significant improvements. |
| A method for resampling imbalanced datasets in binary classification tasks for real-world problems [22] | A unique approach to resampling that blends undersampling and oversampling methods. | Comparing other resampling methods and not using any resampling, there was a significant improvement in classifier performance on imbalanced datasets, particularly in the detection of rare events or minority classes. |
| Selecting the Suitable Resampling Strategy for Imbalanced Data Classification Regarding Dataset Properties. An Approach Based on Association Models [23] | Creation of models that enable the automatic determination of the most effective resampling technique for any given dataset by its unique properties | The efficiency of oversampling and undersampling methods in resolving imbalanced data classification is based upon various factors, including the imbalance ratio, dataset dimensions and size, class overlap, and borderline examples. |
| Resampling imbalanced data for network intrusion detection datasets [24] | Study and understand how resampling affects ANN classifier performance in the context of NIDS | The efficiency of resampling in resolving the imbalanced cybersecurity data issue and enhancing the identification of minority-class data (attacks) |
| Comparing Different Resampling Methods in Predicting Students' Performance Using Machine Learning Techniques [25] | Investigates the distinction between binary and multiclass classification, the effect of feature structures, and the effectiveness of several machine learning classifiers using different resampling strategies and validation techniques | The Random Forest classifier performs better than other models when improved using the SVM-SMOTE method. |

Table 2 shows the flexibility of resampling techniques in solving imbalanced class problems in various domains, including credit card fraud detection, network intrusion detection and student performance prediction. The study emphasizes the importance of understanding the characteristics of the dataset and exploring different resampling strategies to achieve optimal classifier performance.

## 2.2 Ensemble Method

Ensemble techniques are another option for problems with imbalanced classes. This technique combines multiple base models into a larger model with more robust performance [26]. Indirectly, this technique can improve classification accuracy on imbalanced datasets. Several popular ensemble learning methods can be used to improve the machine learning process: bagging and boosting [27].

Freund and Schapire first introduced the boosting method in 1996 [28]. In this algorithm, several weak learners are successively adapted in an extremely adaptive manner [27]. Each model in the sequence is adapted, and observations in the dataset that were poorly handled by the previous models in the sequences are given more weight. Similar to bagging, boosting can be applied to regression and classification problems. There are three different types of boosting algorithms: Extreme Gradient Boosting (XGB), also called XGBoost [29], Stochastic Gradient Boosting (SGB) [30], and Adaptive Boosting (AdaBoost) [28]. The function of the boosting algorithms is shown as follows (1):

$$f(x) = \sum_t a_t h_t(x) \tag{1}$$

where several weak learners $h_t(x)$ are combined to create a strong learner $f(x)$. To do this, a model is first developed using the training dataset and then a second model is developed to correct the errors in the first model $a_t$.

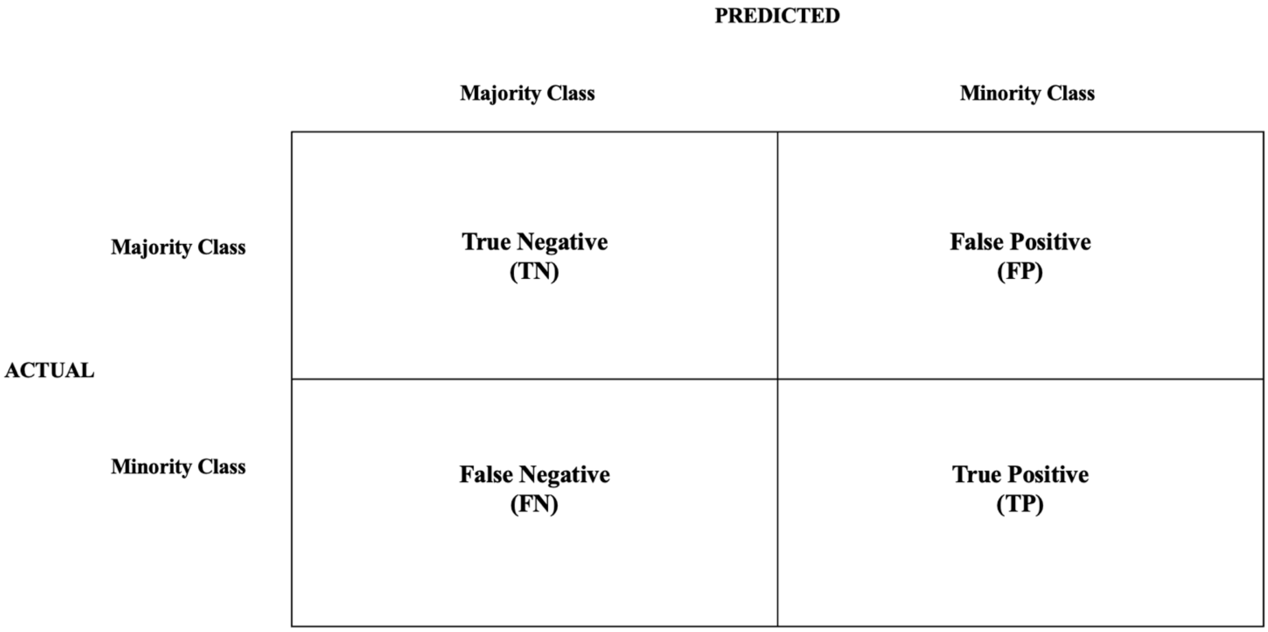
Bootstrap aggregation is another name for the bagging method. It is an algorithm that is completely data specific [31]. Bagging is the combination of bootstrapping and aggregation. By modifying the stochastic distribution of the training dataset, where slight changes in the training set lead to significant deviations in the model predictions, bagging aims to generate more diverse predictive models [32]. The function of bagging algorithms is illustrated as follows (2):

$$f(x) = \frac{1}{B}\sum_{B=1}^{B} f_{b(x)} \tag{2}$$

Where the bootstrapping sets $\frac{1}{B}$ are generated by weak learners $f_{b(x)}$.

## 2.3 Cost-Sensitive Learning Method

To maintain the balance between the classes, the cost-sensitive learning method splits the costs between the majority and minority classes. Higher costs are assigned to the minority class and lower costs are assigned to the majority class [32]. The confusion matrix in Figure 2 describes cost-sensitive learning.



| | | PREDICTED | |
|---|---|---|---|
| | | Majority Class | Minority Class |
| ACTUAL | Majority Class | True Negative<br>(TN) | False Positive<br>(FP) |
| | Minority Class | False Negative<br>(FN) | True Positive<br>(TP) |

**Figure 2** *Confusion matrix*

It is known that confusion matrices to cost matrices. The cost matrix provides the costs associated with the four outcomes, referred to as CTP, CFP, CFN and CTN, shown in Figure 1. CTP and CTN are set to 0 because, as is often the case in cost-sensitive learning, no cost is assigned to a correct classification [33]. Usually CFN > CFP is used because the positive class is misclassified to the negative class. For example, in a fraud detection model, classifying fraudulent transactions as legitimate (FN) may result in losses, while labeling legitimate transactions as fraudulent (FP) may inconvenience customers. The best metric to evaluate classification performance when the cost of misclassification is known is the total cost. The formula for the total cost is as follows (3):

$$Total\ Cost = (FN \times CFN) + (FP \times CFP) \tag{3}$$

## 3. Result and Discussion

The result of the study and the analysis based on the predictive model constructed are intended to explain the performance of the model and the results of each method discussed in Section 2.

To distinguish between fraudulent and non-fraudulent credit card transactions, Hordri et al., [16] investigated the classification performance of different models, including Random Forest (RF), Naive Bayes (NB), Logistic Regression (LR) and Multilayer Perceptron (MLP), and used different resampling techniques, including

Random Undersampling (RUS), Random Oversampling (ROS) and Synthetic Minority Oversampling Technique (SMOTE). The results show that ROS performs better than SMOTE when it comes to improving classification results. RF outperformed the other three resampling methods and showed robust performance among the classification techniques (NB, LR and MLP). Although SMOTE has been shown to be generally effective in the literature, the authors point out that some of the synthetic data it produces may have limitations, as some data include both minority and majority classes. The use of advanced resampling methods such as "improve-SMOTE", which may be more successful in improving classification performance on imbalanced credit card fraud datasets, is something the authors recommend for future research. This could entail additional investigation into the accuracy and distribution of the synthetic data generated by these methods to ensure ideal classification results.

Jia et al., [17] present an Ensemble Convolutional Neural Network (EnCNN) to solve the problem of imbalanced data in intelligent machine fault diagnosis. The main results are that the EnCNN achieves higher and more stable diagnosis accuracy compared to single CNNs in the presence of imbalanced data and can automatically select the hyperparameters of the deep network. The main advantage of EnCNN is that only one hyperparameter (the number of base classifiers) need to be set, and increasing this parameter generally improves the diversity and generalization of the ensemble. However, EnCNN cannot handle scenarios in which new, unrecognized faults occur after the training phase. The authors have recognized this limitation and plan to address it in future work. Possible future research directions include exploring techniques such as few-shot learning, overt sentence detection or incremental learning to improve EnCNN's capabilities in handling new, unseen errors. In addition investigating the impact of different base classifier architectures, ensemble techniques, and dataset equalization strategies could further improve EnCNN's performance.

To solve the problem of imbalanced data classification that frequently occurs in a variety of real-world applications, Feng et al., [18] present a novel Cost-sensitive Feature selection General Vector Machine (CFGVM) algorithm that combines the strengths of the General Vector Machine (GVM) and the Binary Ant Lion Optimizer (BALO) algorithm. The GVM algorithm has a good generalization capability, but it not very efficient with imbalanced datasets. The BALO algorithm, on the other hand, has a high usability and a fast convergence rate, making it suitable for feature selection and cost weight optimization. The CFGVM algorithm uses BALO to determine the optimal cost weights and select the most significant features, which helps to improve the classification performance of the minority class samples. The results show that the CFGVM algorithm significantly outperforms similar and state-of-the-art algorithms in tackling imbalanced classification problems. The combination of cost-sensitive learning and feature selection based on BALO and GVM addresses these challenges. The CFGVM algorithm has its limitations, especially for larger datasets with high-dimensional features, where more iterations may be required to achieve convergence. According to the authors, the CFGVM algorithm is also constructive when it misclassifies the minority class, which can have serious consequences, e.g. in medical applications. An ensemble algorithm based on GVM and BALO, a hybrid algorithm that combines the advantages of undersampling and feature selection, an extension of the CFGVM algorithm to handle concept drift in online learning scenarios, and a new algorithm based on CFGVM to handle high-dimensional imbalanced dataset with small samples are just some of the future research directions suggested by the paper.

## 4. Conclusion

The problem of class imbalance in machine learning models is a major challenge that can affect the performance of classification tasks. Resampling techniques such as oversampling with methods like SMOTE and ADASYN and undersampling with techniques like Near-Miss can effectively address this problem. These approaches aim to equalize the distribution of classes in the training data, which can lead to improved model accuracy. Ensemble methods, such as bagging and boosting (XGBoost, SGB, AdaBoost) can also indirectly improve the performance of models on imbalanced datasets. By combining multiple weak learners, these ensemble methods can effectively capture the complexity of the data and improve overall classification accuracy. For critical applications, cost-sensitive learning can be a valuable approach. This technique assigns a higher cost to the minority class, effectively prioritizing and accurately classifying the less represented class. This can be particularly important in areas where the minority class is of greater importance, such as medical diagnostics or fraud detection.

Dealing with imbalanced datasets is crucial as this can lead to biased and inaccurate predictions, especially in applications where accurate prediction of the minority class is crucial, such as fraud detection and medical diagnosis. If this problem not solved, it can lead to significant financial losses and adverse health effects. Potential areas for future research in imbalanced classification include exploring advanced resampling techniques, investigating the impact of different base classifier architectures and ensemble techniques, extending existing algorithms to handle concept drift and high-dimensional imbalanced datasets, and developing hybrid algorithms that combine the benefits of different applications..

## Acknowledgement

This research was supported by Universiti Tun Hussein Onn Malaysia (UTHM) through Tier 1 (Vot Q507) and GPPS (Vot Q662).

## Conflict of Interest

The authors declare no conflict of interest regarding the paper's publication.

## Author Contribution

*The author confirms sole responsibility for the following: study conception and design, data collection, analysis and interpretation of results, and manuscript preparation.*

## References

[1] Mahesh, B. (2020). Machine learning algorithms-a review. International Journal of Science and Research (IJSR). [Internet], 9(1), 381-386.

[2] Ozgur, A. (2004). Supervised and unsupervised machine learning techniques for text document categorization. Unpublished Master's Thesis, İstanbul: Boğaziçi University.

[3] Cho, S., Vasarhelyi, M. A., Sun, T., & Zhang, C. (2020). Learning from machine learning in accounting and assurance. Journal of Emerging Technologies in Accounting, 17(1), 1-10.

[4] Ramyachitra, D., & Manikandan, P. (2014). Imbalanced dataset classification and solutions: a review. International Journal of Computing and Business Research (IJCBR), 5(4), 1-29.

[5] Sarker, I. H. (2021). Machine learning: Algorithms, real-world applications and research directions. SN computer science, 2(3), 160.

[6] Kotsiantis, S., Kanellopoulos, D., & Pintelas, P. (2006). Handling imbalanced datasets: A review. GESTS international transactions on computer science and engineering, 30(1), 25-36.

[7] Paulus, J. K., & Kent, D. M. (2020). Predictably unequal: understanding and addressing concerns that algorithmic clinical prediction may increase health disparities. NPJ digital medicine, 3(1), 99.

[8] Makki, S., Assaghir, Z., Taher, Y., Haque, R., Hacid, M. S., & Zeineddine, H. (2019). An experimental study with imbalanced classification approaches for credit card fraud detection. IEEE Access, 7, 93010-93022.

[9] Zhang, Y. F., Lu, H. L., Lin, H. F., Qiao, X. C., & Zheng, H. (2022). The optimized anomaly detection models based on an approach of dealing with imbalanced dataset for credit card fraud detection. Mobile Information Systems, 2022(1), 8027903.

[10] Hung, L. C., Hu, Y. H., Tsai, C. F., & Huang, M. W. (2022). A dynamic time warping approach for handling class imbalanced medical datasets with missing values: A case study of protein localization site prediction. Expert Systems with Applications, 192, 116437.

[11] Fotouhi, S., Asadi, S., & Kattan, M. W. (2019). A comprehensive data level analysis for cancer diagnosis on imbalanced data. Journal of biomedical informatics, 90, 103089.

[12] Behrad, F., & Abadeh, M. S. (2022). An overview of deep learning methods for multimodal medical data mining. Expert Systems with Applications, 200, 117006.

[13] Haixiang, G., Yijing, L., Shang, J., Mingyun, G., Yuanyue, H., & Bing, G. (2017). Learning from class-imbalanced data: Review of methods and applications. Expert systems with applications, 73, 220-239.

[14] Felix, E. A., & Lee, S. P. (2019). Systematic literature review of preprocessing techniques for imbalanced data. IET Software, 13(6), 479-496.

[15] Charte, F., Rivera, A. J., del Jesus, M. J., & Herrera, F. (2015). Addressing imbalance in multilabel classification: Measures and random resampling algorithms. Neurocomputing, 163, 3-16.

[16] Hordri, N. F., Yuhaniz, S. S., Azmi, N. F. M., & Shamsuddin, S. M. (2018). Handling class imbalancein credit card fraud using resampling methods. Int. J. Adv. Comput. Sci. Appl, 9(11), 390-396.

[17] Jia, F., Li, S., Zuo, H., & Shen, J. 2020). Deep neural network ensemble for the intelligent fault diagnosis of machines under imbalanced data. IEEE Access, 8, 120974-120982.

[18] Feng, F., Li, K. C., Shen, J., Zhou, Q., & Yang, X. (2020). Using cost-sensitive learning and featureselection algorithms to improve the performance of imbalanced classification. IEEE Access, 8, 69979-69996.

[19] Sasada, T., Liu, Z., Baba, T., Hatano, K., & Kimura, Y. (2020). A resampling method for imbalanceddatasets considering noise and overlap. Procedia Computer Science, 176, 420-429.

[20] Van den Goorbergh, Ruben; van Smeden, Maarten; Timmerman, Dirk; Van Calster, Ben. (2022). "The harm of class imbalance corrections for risk prediction models: illustration and simulation using logistic regression". Journal of the American Medical Informatics Association. 29 (9): 1525–1534.

[21] Mqadi, N. M., Naicker, N., & Adeliyi, T. (2021). Solving misclassification of the credit card imbalance problem using near miss. Mathematical Problems in Engineering, 2021, 1-16.

[22] Cateni, S., Colla, V., & Vannucci, M. (2014). A method for resampling imbalanced datasets in binary classification tasks for real-world problems. Neurocomputing, 135, 32-41.

[23] Kraiem, M. S., Sánchez-Hernández, F., & Moreno-García, M. N. (2021). Selecting the suitable resampling strategy for imbalanced data classification regarding dataset properties. An approach based on association models. Applied Sciences, 11(18), 8546.

[24] Bagui, S., & Li, K. (2021). Resampling imbalanced data for network intrusion detection datasets. Journal of Big Data, 8(1), 1-41

[25] Ghorbani, R., & Ghousi, R. (2020). Comparing different resampling methods in predicting students' performance using machine learning techniques. IEEE Access, 8, 67899-67911

[26] Jian, C., Gao, J., & Ao, Y. (2016). A new sampling method for classifying imbalanced data based onsupport vector machine ensemble. Neurocomputing, 193, 115-122.

[27] Mohammed, A., & Kora, R. (2023). A comprehensive review on ensemble deep learning: Opportunities and challenges. Journal of King Saud University-Computer and Information Sciences, 35(2), 757-774.

[28] Freund, Y., & Schapire, R. E. (1996, July). Experiments with a new boosting algorithm. In icml (Vol. 96, pp. 148-156).

[29] Friedman, J., Hastie, T., & Tibshirani, R. (2000). Additive logistic regression: a statistical view of boosting (with discussion and a rejoinder by the authors). The annals of statistics, 28(2), 337-407.

[30] Friedman, J. H. (2001). Greedy function approximation: a gradient boosting machine. Annals of statistics, 1189-1232.

[31] Breiman, L. (1996). Bagging predictors. Machine learning, 24, 123-140.

[32] Ali, A., Shamsuddin, S. M., & Ralescu, A. L. (2013). Classification with class imbalance problem. Int. J. Advance Soft Compu. Appl, 5(3), 176-204.

[33] Weiss, G. M., McCarthy, K., & Zabar, B. (2007). Cost-sensitive learning vs. sampling: Which is best for handling unbalanced classes with unequal error costs?. Dmin, 7(35-41), 24.