Mediterranean Smart Cities Conference (MSCC 2024)
2-4 May, 2024, Martil, Morocco

An Improved Hybrid Sampling Method for Classifying imbalanced Data to Predict Student Performance

Mohamed Bellaj
Abdelmalek Essadi University
Hight Normal School
Tetouan, Morocco
mohamed.bellaj@etu.uae.ac.ma

Ahmed Ben dahmane
Abdelmalek Essadi University
Hight Normal School
Tetouan, Morocco
abendahman@uae.ac.ma

Abstract—**Data imbalance in Machine Learning is defined as an unbalanced distribution of classes within a dataset. This issue is most commonly encountered in classification jobs where the distribution of classes or labels in a given dataset is not uniform. The simplest way to handle this problem is to use resampling, which involves adding records to the minority class or removing them from the majority class. In this research, we experimented with two frequently used resampling techniques: oversampling (ROS) and undersampling (RUS). We also proposed a hybrid sampling algorithm that combines ROS and RUS. To investigate all strategies, we picked a public educational dataset and used three types of machine learning algorithms: Random Forest (RF), Logistic regression (LR), and AdaBoost. One of the important conclusions of this paper is that oversampling outperforms undersampling for different classifiers and achieves greater scores in many assessment criteria. The experimental results suggest that a hybrid strategy combining ROS and RF outperforms the other benchmark techniques. This technique has a high potential for enhancing the students' performance prediction models.**

Keywords—machine learning; imbalanced Data; resampling; ROS; RUS

I. INTRODUCTION

In machine learning and statistics, classification is the process of using labeled datasets to train a system to recognize new, unseen datasets and determine which class they belong to. The amount of data has recently increased significantly, but there is a dearth of high-quality labeled data. Various classical machine learning techniques were predicated on the target classes being distributed uniformly. In machine learning and statistics, classification is the process of training a system to recognize new, previously unknown datasets and determine which class they belong to. The volume of data has recently expanded dramatically, but there is a scarcity of high-quality labeled data. Several classical machine learning approaches relied on the target classes being distributed equally. This assumption, however, is false in a lot of applications since the vast majority of instances are labeled with one class and very few with the other. As a result, the models eliminate the minority class while favoring the majority class more. Unbalanced datasets have a negative impact on model performance. The phrase for this problem is class imbalance. As a result, while great accuracy may be achieved in this scenario, other assessment measures like as precision, recall, F1-score [1], and ROC score are not much improved.

In general, there are three approaches to dealing with uneven data: data level, algorithm level, and ensemble techniques [2]. Resampling the data is the data-level approach's method for reducing class imbalance. The two main resampling procedures are undersampling the majority class and oversampling the minority class. The most basic type of oversampling is known as random oversampling (ROS). Rac [3] employed the adaptive synthetic (ADASYN) strategy to predict student performance, as well as the synthetic minority oversampling technique (SMOTE), SVMSMOTE, and borderline-SMOTE. Using a variety of classifiers, they observed that the Borderline-SMOTE approach generated the best prediction results. Random undersampling (RUS) is the most commonly utilized undersampling approach. Certain researchers have used hybrid sampling, which includes both oversampling and undersampling methodologies [4]. These tactics contribute to ensuring that classifiers are not biased toward any one class by providing a balanced dataset. To resolve unbalanced data, [5] used a hybrid sampling technique that included RUS followed by ROS and a balance loss cost function. They discovered that oversampling and then undersampling outperformed undersampling and then oversampling. Furthermore, they discovered that the proposed strategy improved performance by 7%. The ROS-RUS method has the virtue of being simple, lacking heuristics, and implying nothing about the data. In a separate study, the under, over, and SMOTE sampling approaches were used to address the unbalanced issue on ten datasets. They discovered that the hybrid sampling strategy outperformed the alternatives.

The paper's remaining sections are organized as follows: This section describes the ways for overcoming an unbalanced dataset. In the second section, we discuss similar studies. And we present the methodology in section3. Section 4 features a discussion after the presentation of the experiment analysis and appraisal of an unbalanced dataset. This research is completed in Section 5.

II. Related Works

Several approaches to imbalanced data classification have been explored over the years. This section gives a brief introduction of sampling techniques, categorizing them as under-sampling, over-sampling, or hybrid sampling depending on the algorithms utilized.

We're interested in the following resampling methods: ROS for oversampling, RUS for undersampling, SMOTE-NC for nominal and continuous, and RUS for hybrid sampling. These techniques were chosen based on the findings of [6] research, which demonstrated that ROS, RUS, and the SMOTE hybrid perform well with educational data. According to [7], ROS and SMOTE outperformed ensemble classifiers when coping with multi-class imbalances in student log data. In [8], a hybrid approach combining SMOTE and one-sided selection undersampling was demonstrated to efficiently eliminate class imbalance in EDM. In [9], SMOTE, ROS, and RUS were discovered to increase classification models' effectiveness in forecasting students' final grades when dealing with class imbalance.

III. Metodology

A. *Datasets Collection*

Using the experience API, a method for tracking student actions (xAPI), the data used in this study comes from the Kalboard 360 learning management system, which was implemented in junior high schools and includes academic data. Developed with the intention of improving learning, Kalboard 360 is a multi-agent junior high school. The experience API allows providers to define students, activities, and objects that comprise the learning process. A unique approach that allows simultaneous access to instructional information via internet-connected devices.

There are 480 instances in the collection, and each one has 16 attributes. Three categories are used to group these traits: (1) demographic characteristics, like nationality and gender; (2) academic background characteristics, like education, class, and grade level; and (3) social characteristics, like involvement in class discussions, availability of scientific learning resources, responsiveness to parent surveys, and parental satisfaction with the school. The dataset's class distribution looks like this: There are 127 instances in the low class, 211 in the medium class, and 142 in the high class.

TABLE I. THE DATASET ATTRIBUTES

Feature Category	**Attribute**	**Description**
Demographical	gender	The gender of the student {female or male}
	Nationality	Student nationality
	Relation	Student's contact parent such as {father or mum}
Academic Background	StageID	Stage student belongs such as {Lower level , Middle level , and high level }
	GradeID	Grade student belongs such as {G-01, G-02, G-03, G-04, G-05, G-06, G-07, G-08, G-09, G-10, G-11, G-12}
	SectionID	Section student belongs such as {A, B, C}.
	Topic	Course topic such as {Math, English, IT, Arabic, Science, Quran}
	StudentAbsence Days	Student absence days {Above-7, Under-7}
	Semester	School year semester as {First or second}
Behavioral	RaisedHands	How many times the student raises his/her hand on classroom {numeric: 0- 100}
	VisitedResources	How many times

		the student visits a course content {numeric: 0-100}
	Announcements View	How many times the student checks the new announcements{numeric:0- 100}
	Discussion	How many times the student participated on discussion groups {numeric:0-100}
Parent Participation in learning process	ParentSchool Satisfaction	This feature obtains the Degree of parent satisfaction from school as follow {Good, Bad}
	ParentAnswering Survey	Parent is answering the surveys that provided from school or not {Yes, No}.
Performance Level	Class	Total score {Low-Level: 0 to 69, Middle-Level: 70 to 89, High: interval 90 to 100}

B. *Tools and Instruments*

This study used Jupyter Notebook and Python to conduct the experiment. The Imblearn Python module was used to resample the imbalance class, and the Scikit-learn Python library was used to apply the machine learning technique.

C. *Preprocessing*

A sequence of preprocessing steps is carried out using the Scikitlearn preprocessing tool for Python: first, preliminary feature filtering was performed to remove any unnecessary data from the list; second, the category data is transformed into numerical values to check that they are acceptable for modeling. Preprocessing is an important step in data mining that enhances data quality and guarantees the modeling process is more efficient [11].

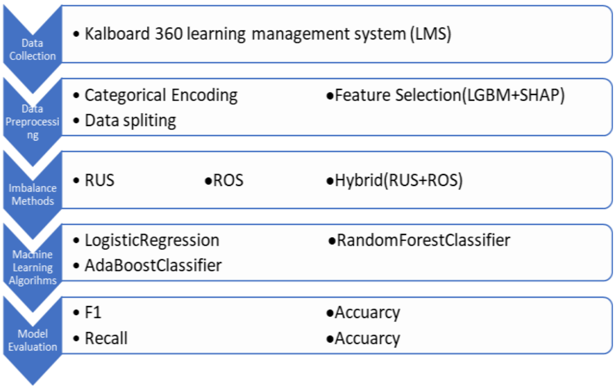


Figure 1: Preprocessing phases

D. *Data encoding*

It is common in the field of machine learning to encounter datasets with multiple labels in one or more columns. These labels can be character or numerical values. Nevertheless, the original format of such data cannot be directly incorporated into a machine learning model; instead, label encoding—a technique that converts labels into numeric notation—is used to make data comprehensible to models[12]. For supervised learning techniques, this preprocessing stage is critical.

It is typical to come across datasets in the field of machine learning that have several labels in one or more columns. These labels could be represented by letters or numbers. Such data, though, cannot be put straight into the original framework of a machine learning model. One popular method for providing models with understandable data is label encoding [12]. The process of translating labels into numeric notation for use in a machine learning model is known as label encoding. This preprocessing stage is required for methods involving supervised learning.

Each value in a category column is typically substituted in label encoding by a number between 0 and N-1. One utility class that makes label normalization easier is the Label Encoder.

E. *Feature Selection*

- LGBM

Feature significance in LightGBM (Light Gradient Boosting Machine) is a technique for figuring out which features (variables) in your dataset influence the model's predictions the most [13]. Two different feature importance approaches are used by LightGBM:

Gain (or Split Importance): In trees, this method determines the relative contribution of each feature to the model by calculating its overall gain. The amount that a feature increases accuracy on the branches it impacts is indicated by its gain. A trait is more significant for making predictions if its gain value is higher. Put simply, it quantifies the extent to which a feature enhances the model's predictions, mostly by focusing on the number of better splits it produces.

Split (or Frequency Importance): This method evaluates a feature's relevance by counting how often it occurs in a model. An attribute is deemed more significant if it is frequently used to distinguish between sets of data. With this method, a feature's usage in data partitioning across all trees is counted; a feature with a larger count is considered more important.

- SHAP

Any machine learning model's output is described using (SHapley Additive exPlanations) values. It makes use of a game theory technique that evaluates each player's input into the final outcome. Every feature in machine learning is given a significance value that indicates how much of an impact it has on the model's output.

The feature significance and a summary of the SHAP dependent charts are shown in the SHAP summary plot [14]. As mentioned in (7), the graphic stacks attributes vertically after sorting them according to significance. For every feature Xi, the row plot depicts a description of the SHAP dependency plot. The horizontally displayed SHAP value φ(x (j) I for each patient is represented by a dot. The value of each feature is represented by a dot, which ranges from low (blue) to high (red). Missing values are indicated by black dots. Plotting blue dots on the upper side and red points on the lower side indicates that the hazard grows as the value rises. The SHAP study can be summarized using a SHAP summary plot, which shows the significance of feature values, and an abstract of the SHAP dependency plot.

F. Resampling Technique

The procedure of balancing the uneven classes in order to address the issue of unbalanced data is known as resampling. Three categories of sampling techniques exist [15]:

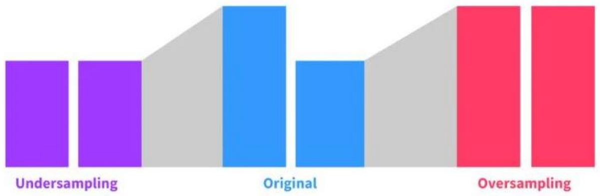


Figure 2:Undersampling and oversampling data explanation

Using methods for handling unbalanced datasets, the dataset is first balanced. Currently, the individual applications of random undersampling (RUS), random oversampling (ROS), and the hybrid technique (RUS+ROS) are being made. The balanced dataset is then subjected to machine learning techniques to provide the findings. Lastly, the results of machine learning techniques on the imbalanced dataset are contrasted with those of machine learning techniques on the balanced dataset using the imbalanced dataset management strategies.

- Oversampling Techniques

In 2002, a reported study by [17] developed an oversampling approach called SMOTE that can build a minority synthetic class. It became more popular and the most commonly used method in unbalanced classification situations. However, oversampling strategies that replicate or create additional instances might result in overfitting. This strategy will be copied or used to create new minority class instances [16]. Examples of this approach include the Synthetic Minority Oversampling Technique (SMOTE), Random Oversampling (ROS), and Adaptive Synthetic Sampling (ADASYN).

ROS: A straightforward sampling technique for growing the number of the minority group is random oversampling. This technique [18] chooses data points from the minor class at random and duplicates them exactly. Because of this, there were more minority samples in order to maintain equilibrium between the two classes.

- Undersampling Techniques

Data instances from the majority class are removed by employing this technique [18]. Among the most widely used undersampling techniques are Random Undersampling (RUS), Edited Nearest Neighbors (ENN), and Tomek Links (TL). This method works well with big datasets. This approach works well, but it has a drawback in that less data means less information. It is a well-liked and practical technique for data scientists despite its limitations, such as the loss of potentially important data.Data instances from the majority class are removed by employing this technique [18]. Among the most widely used undersampling techniques are Random Undersampling (RUS), Edited Nearest Neighbors (ENN), and Tomek Links (TL). This method works well with big datasets. This approach works well, but it has a drawback in that less data means less information. It is a well-liked and practical technique for data scientists despite its limitations, such as the loss of potentially important data.

Rus: In order to balance the distribution of classes for the learning process, the most basic and straightforward method for resampling an unbalanced dataset is random undersampling, in which samples from the majority class are randomly eliminated from the class.

- Hybrid Techniques

This method combines ensemble approaches with oversampling, or oversampling with undersampling. Examples of this strategy are SMOTE-ENN, SMOTE-TL, SMOTEBoost, and RUSBoost.

G. Classification using Machine Learning Classifiers

This work employs five different approaches to ensemble learning. The Decision Tree technique serves as the foundation algorithm for ensemble learning classifiers by default. The following is a brief description:

- Random Forest

The variance of the model is decreased by pooling many DT in the RF ensemble[19]. The process of averaging forecasts in an RF helps to generate predictions for unknown samples.

$$I = \frac{1}{N}\sum_{N=1}^{N} \int(x) \quad (1)$$

The RF method gathers predictions from multiple DTs, aggregates them, and determines the best course of action based on the analysis of the data. Additionally, RF is capable of handling datasets with missing values and functions through the use of the bagging technique and an ensemble learning paradigm.

- Logistic Regression

A popular linear classification algorithm, LR can simulate the probability of a binary outcome, either 1 or 0, depending on one or more predictor variables [20]. When applied to a linear combination of the predictor variables, the logistic regression equation takes the form of the logistic function, also referred to as the sigmoid function. The following is the standard expression for the formulation for LR:

$$P(Y=1|X)= g(\mathbf{x}) = \operatorname{sign}\left(\sum_{m=1}^{M} \alpha_m f_m(\mathbf{x})\right) \quad (2)$$

- AdaBoost

is an algorithm for boosting that gets around the drawbacks of boosting. Focus was placed on the hard-to-classify area or pattern. Each subgroup receives the same distribution of weights. As a result, the properly categorized instance weight will decrease and the misclassified instance weight will increase. Ultimately, the voting process was used to turn weaker learning groups into stronger learning groups [21].

$$\frac{1}{1+e^{-(\beta 0+\beta 1X1+\beta 2X2+\cdots+\beta pXp)}} \quad (3)$$

- Performance Measure

The measure technique is assessed using a range of metrics, such as accuracy, F-measure, and ROC, in terms of assessment metrics. The accuracy statistic uses the three performance classifications (low, medium, and high) to show the total number of properly detected instances. The F-measure is an evaluation metric that combines and optimizes the advantages of recall and precision. Recall, sometimes referred to as sensitivity, is the percentage of retrieved and relevant examples divided by the total number of relevant instances, whereas precision, also known as positive predictive value, is the proportion of recovered instances that are relevant. The classifier's ability to discern between positive and negative instances is examined using the ROC area (or curve), which is then used to establish a threshold for doing so [22].

Accuracy=(TP+TN)/(TP+FP+FN+TN) (4)

Precision=TP/(TP+FP) (5)

Recall =TP/(TP+FN) (6)

F1-score=(2×Precision×Recall)/(Precision+Recall) (7)

IV. RESULT AND DISCUSSION

Four experimental phases comprised this study: (1) feature selection; (2) testing with many machine learning (ML) algorithms; (3) testing with random oversampling and undersampling (ROS and RUS); and (4) testing with a hybrid technique (oversampling-undersampling). Initially, each classifier's performance was assessed by the use of the ML approach (Random Forest, Logistic Regression, and Adaboost). Secondly, the ROS, RUS, and Hybride approaches were used for testing.

A. *Feature selection*

Using only pertinent data and eliminating noise from the data is a technique called feature selection that helps you reduce the number of input variables for your model.

To reduce the overall number of features and compare training times, we first employed the LGBM classifier to ascertain the degree of correlation between the variables in order to establish the importance ranking.
After selecting features, we focus on the eight most important factors during the training and testing stage.

B. *Resampling Results*

Every Figure displays the outcomes of every resampling method. Data that includes both the majority and minority classes was used, without any resampling.

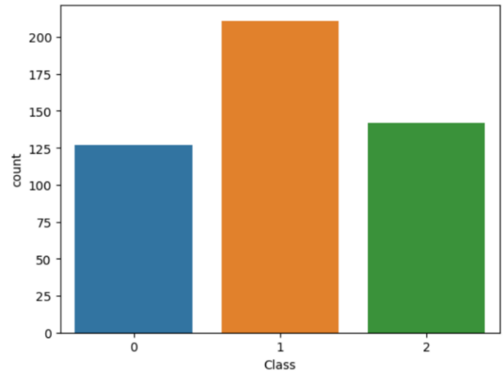


Figure 3: Distribution of the data without resampling

"0" denotes students with a low level (127 students), "1" denotes students with a midle level (211 students), and "2" denotes students with a high level (142 students) in the somewhat unbalanced dataset.

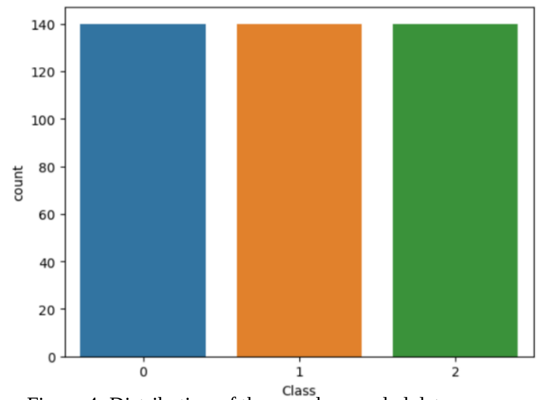
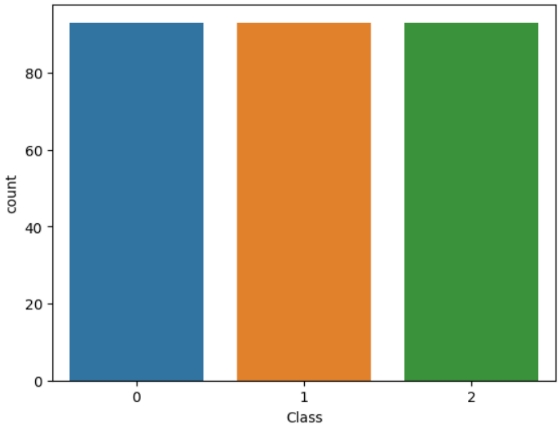


Figure 4: Distribution of the overdersampled data

The sample size of students with low performance level increased to 140 in figure 4, and the sample size of students with low performance level became 140. In other words, ROS raised the size of the minority class by



randomly resampling to match the size of the majority class.

On the other hand, RUS resized the majority class at random to make it equal to the size of the minority class. In other words, the sample size for all student performance levels was lowered to 90.

C. Evaluating Results Using maching learning algorithms without resampling methods

Without resampling the original data, this subsection displays the performance results of the two ensemble models and the five traditional machine learning models.

TABLE II. THE PERFORMANCE OF THE USED MODELS

Model		F1	Precision	Accuracy	Roc-Auc	Recall
LR	train	0.79	0.79	0.79	0.92	0.79
	test	0.76	0.76	0.76	0.91	0.76
RF	train	0.77	0.77	0.77	0.91	0.77
	test	0.73	0.73	0.73	0.89	0.73
Ada Boost	train	0.72	0.72	0.72	0.85	0.72
	test	0.78	0.78	0.78	0.84	0.78

The two ensemble models, LR and RF, perform better than the other common machine learning models in the training stage in terms of accuracy, precision, f-measure, AUC, and recall, according to Table II's data. One clear benefit of LR over other approaches of machine learning. The testing procedure then introduces adaboost. This indicates that while Adaboost performs best in the other measurements, LR has the highest ROC AUC performance.

D. Comparison of performance metrics for all machine learning after resampling

Three machine learning algorithms—ROS, RUS, and the hybrid ROS-RUS sampling algorithm—are evaluated in this study. Additionally, the method without resampling—that is, one that lacks a clear method for handling imbalance—was employed as a performance benchmark. Table III's results show that Adaboost with Original data performed better than any other metric, with the exception of ROC AUC, which was followed by LR.

With the exception of Adaboost (ROC-AUC = 0.83), RUS underperformed for the moderately imbalanced dataset, a s evidenced by its lowest overall performance metric score s when applying LR (accuracy = 0.69, precision = 0.71, re call = 0.69, and F1-score = 0.71), compared to the highest performance metric when applying RF (accuracy = 0.75, p recision = 0.76, recall = 0.75, ROC-AUC = 0.93, and F1-s core = 0.76).

TABLE III. THE PERFORMANCE OF USING ROS, RUS AND ROS+RUS

Model		F1	Precision	Accuracy	Roc-Auc	Recall
LR	test	0.76	0.76	0.76	0.91	0.76
	ROS	0.75	0.75	0.74	0.91	0.74
	RUS	0.71	0.71	0.69	0.90	0.69
	ROS+ RUS	0.72	0.72	0.71	0.91	0.71
RF	test	0.73	0.73	0.73	0.89	0.73
	ROS	**0.78**	**0.78**	**0.78**	**0.93**	**0.78**
	RUS	**0.76**	**0.76**	**0.75**	**0.93**	**0.75**
	ROS+ RUS	**0.74**	**0.74**	**0.73**	**0.91**	**0.71**
Ada Boost	test	0.78	0.78	0.78	0.84	0.78
	ROS	0.69	0.69	0.67	0.83	0.67
	RUS	0.73	0.74	0.72	0.83	0.72
	ROS+ RUS	0.73	0.73	0.72	0.83	0.72

Applying the Adabost classifier produced the lowest results (accuracy = 0.67, precision = 0.69, recall = 0.78, ROC-AUC = 0.83, and F1-score = 0.67), but ROS fared the best when using RF (accuracy = 0.78, precision = 0.78, recall = 0.78, ROC-AUC = 0.93, and F1-score = 0.69).

The lowest metric results achieved by LR (accuracy = 0.71, precision = 0.72, recall = 0.71, and F1-score = 0.72) after utilizing the hybrid resampling method ROS+RUS, with the exception of Adaboost (ROC-AUC = 0.83). Results demonstrate the greatest gains when RF is used (accuracy = 0.73, precision = 0.74, ROC-AUC = 0.91, and F1-score = 0.74), with the exception of Adaboost (recall = 0.75).

The results of the study showed that classification performance was significantly impacted by oversampling, undersampling, and hybrid sampling. RF performed better than all other classifiers when ROS was used for oversampling. But using hybrid approaches (ROS+RUS) and RUS enhances RF performance much further. All resampling techniques, however, produce a negative result with no improvement in performance when combined with the two additional classifiers, Adaboost and LR.

V. CONCLUSION AND FUTURE WORKS

The usefulness of using data sampling techniques to develop prediction models for unbalanced education quality data was illustrated in this paper. These methods mostly involve transforming an unbalanced dataset into a balanced dataset using preprocessing techniques like RUS, ROS, and ROS-RUS (hybrid sampling), which are then used to three machine learning classification models.

To analyze and assess the issue, we used the unbalanced dataset. Using the dataset, we experimented with oversampling, undersampling, and hybrid resampling strategies. We also assessed classifiers using a variety of criteria. Across all assessment metrics, ROS achieves higher scores than RUS and ROS+RUS for a large number of classifiers. After utilizing various resampling techniques, including ROS, RF proved to be the most effective model for the issue under discussion in this paper.

In order to compare the two, we intend to use several deep learning approaches in conjunction with resampling techniques in future work.

References

[1] Y. Liu, L. Zhu, L. Ding, H. Sui, and W. Shang, "A hybrid sampling method for highly imbalanced and overlapped data classification with complex distribution," Inf. Sci. (Ny)., p. 120117, 2024.

[2] M. A. Arefeen, S. T. Nimi, and M. S. Rahman, "Neural network-based undersampling techniques," IEEE Trans. Syst. Man, Cybern. Syst., vol. 52, no. 2, pp. 1111–1120, 2020.

[3] 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.

[4] X. Yi, Y. Xu, Q. Hu, S. Krishnamoorthy, W. Li, and Z. Tang, "ASN-SMOTE: a synthetic minority oversampling method with adaptive qualified synthesizer selection," Complex Intell. Syst., vol. 8, no. 3, pp. 2247–2272, 2022.

[5] A. Newaz, S. Hassan, and F. S. Haq, "An empirical analysis of the efficacy of different sampling techniques for imbalanced classification," arXiv Prepr. arXiv2208.11852, 2022.

[6] N. Mqadi, N. Naicker, and T. Adeliyi, "A SMOTe based oversampling data-point approach to solving the credit card data imbalance problem in financial fraud detection," Int. J. Comput. Digit. Syst., vol. 10, no. 1, pp. 277–286, 2021.

[7] X.-Y. Gao, A. Amin Ali, H. Shaban Hassan, and E. M. Anwar, "Improving the accuracy for analyzing heart diseases prediction based on the ensemble method," Complexity, vol. 2021, pp. 1–10, 2021.

[8] K. S. Selim and S. S. Rezk, "On predicting school dropouts in Egypt: a machine learning approach," Educ. Inf. Technol., pp. 1–32, 2023.

[9] S. D. A. Bujang et al., "Imbalanced Classification Methods for Student Grade Prediction: A Systematic Literature Review," IEEE Access, 2022.

[10] E. A. Amrieh, T. Hamtini, and I. Aljarah, "Preprocessing and analyzing educational data set using X-API for improving student's performance," in 2015 IEEE Jordan conference on applied electrical engineering and computing technologies (AEECT), 2015, pp. 1–5.

[11] A. A. Jasim, L. R. Hazim, and W. D. Abdullah, "Characteristics of data mining by classification educational dataset to improve student's evaluation," J. Eng. Sci. Technol., vol. 16, no. 4, pp. 2825–2844, 2021.

[12] M. Schuld, R. Sweke, and J. J. Meyer, "Effect of data encoding on the expressive power of variational quantum-machine-learning models," Phys. Rev. A, vol. 103, no. 3, p. 32430, 2021.

[13] M. Massaoudi, S. S. Refaat, I. Chihi, M. Trabelsi, F. S. Oueslati, and H. Abu-Rub, "A novel stacked generalization ensemble-based hybrid LGBM-XGB-MLP model for Short-Term Load Forecasting," Energy, vol. 214, p. 118874, 2021.

[14] Y. Meng, N. Yang, Z. Qian, and G. Zhang, "What makes an online review more helpful: an interpretation framework using XGBoost and SHAP values," J. Theor. Appl. Electron. Commer. Res., vol. 16, no. 3, pp. 466–490, 2020.

[15] M. Khushi et al., "A comparative performance analysis of data resampling methods on imbalance medical data," IEEE Access, vol. 9, pp. 109960–109975, 2021.

[16] D. Gonzalez-Cuautle et al., "Synthetic minority oversampling technique for optimizing classification tasks in botnet and intrusion-detection-system datasets," Appl. Sci., vol. 10, no. 3, p. 794, 2020.

[17] P. Soltanzadeh and M. Hashemzadeh, "RCSMOTE: Range-Controlled synthetic minority over-sampling technique for handling the class imbalance problem," Inf. Sci. (Ny)., vol. 542, pp. 92–111, 2021.

[18] A. Sayli and S. BAŞARIR, "Sampling Techniques and Application in Machine Learning in order to Analyse Crime Dataset," Avrupa Bilim ve Teknol. Derg., no. 38, pp. 296–310, 2022.

[19] M. Sheykhmousa, M. Mahdianpari, H. Ghanbari, F. Mohammadimanesh, P. Ghamisi, and S. Homayouni, "Support vector machine versus random forest for remote sensing image classification: A meta-analysis and systematic review," IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens., vol. 13, pp. 6308–6325, 2020.

[20] P. Schober and T. R. Vetter, "Logistic regression in medical research," Anesth. Analg., vol. 132, no. 2, p. 365, 2021.

[21] A. Shahraki, M. Abbasi, and Ø. Haugen, "Boosting algorithms for network intrusion detection: A comparative evaluation of Real AdaBoost, Gentle AdaBoost and Modest AdaBoost," Eng. Appl. Artif. Intell., vol. 94, p. 103770, 2020.

[22] R. Soleymani, E. Granger, and G. Fumera, "F-measure curves: A tool to visualize classifier performance under imbalance," Pattern Recognit., vol. 100, p. 107146, 2020.