ภาคผนวก ก

ผลงานที่ตีพิมพ์ในระหว่างการศึกษา

ผลงานที่ตีพิมพ์ในระหว่างการศึกษา

Panjainam, P., & Kanjanawattana, S. (2024). A Comparison of the Hybrid Resampling Techniques for Imbalanced Medical Data. ICRSA 2024: 2024 7th International Conference on Robot Systems and Applications, Bangkok, Thailand. 12 – 14 September 2024, 5 Pages, https://doi.org/10.1145/3702468.3702477

A Comparison of the Hybrid Resampling Techniques for Imbalanced Medical Data

Paonrat Panjainam
School of Computer Engineering Institute of Engineering
Suranaree University of Technology
Nakhon Ratchasima, Thailand
m6500979@g.sut.ac.th

Sarunya Kanjanawattana
School of Computer Engineering Institute of Engineering
Suranaree University of Technology
Nakhon Ratchasima, Thailand
sarunya.k@sut.ac.th

Abstract

Extremely severe preeclampsia is a disorder that emerges during pregnancy and is defined by the development of high blood pressure; both the mother and the fetus may perish as a result. However, the number of preeclampsia patients is substantially fewer in terms of statistics compared to that in a typical pregnancy. Thus, the uneven data offers a barrier to the building of highly effective machine learning models. This study attempted to overcome the problem of data imbalance in the preeclampsia dataset. Four unique undersampling algorithms were implemented: Random Undersampling (RUS), Tomek Link (Tomek), Edited Nearest Neighbors (ENN), and Repeated Edited Nearest Neighbors (RENN). Similarly, four other oversampling procedures were employed: Adaptive Synthetic Sampling (ADASYN), Borderline-SMOTE, Synthetic Minority Oversampling Technique (SMOTE), and Random Oversampling (ROS). In order to investigate the balanced data, Decision Tree, Naive Bayes, Random Forest, and XGBoost were deployed. According to the experimental findings, the XGBoost model, which employed the hybrid resampling technique Tomek and ROS, demonstrated the greatest degree of efficacy, as indicated by an AUC analysis.

CCS Concepts

• **Computing methodologies** → **Machine learning**; *Machine learning approaches*; *Bio-inspired approaches*; Generative and developmental approaches.

Keywords

Data imbalance, Data resampling, Undersampling, Preeclampsia data, Oversampling

ACM Reference Format:
Paonrat Panjainam and Sarunya Kanjanawattana. 2024. A Comparison of the Hybrid Resampling Techniques for Imbalanced Medical Data. In *2024 7th International Conference on Robot Systems and Applications (ICRSA) (ICRSA 2024), September 12–14, 2024, Bangkok, Thailand.* ACM, New York, NY, USA, 5 pages. https://doi.org/10.1145/3702468.3702477

ICRSA 2024, September 12–14, 2024, Bangkok, Thailand

ACM ISBN 979-8-4007-1703-1/24/09
https://doi.org/10.1145/3702468.3702477

1 Introduction

Currently, preeclampsia is a specific disease occurring during pregnancy that involves the development of high blood pressure during pregnancy, which is very severe, resulting in death for the pregnant woman and the fetus. The initial symptoms of this condition in pregnant women are protein leakage in the urine and elevated blood pressure. Additionally, close monitoring and attention should be given to this confluence of conditions. Epilepsy, which is directly associated with hypertension during pregnancy, may occur in severe cases of symptoms. A preeclampsia disorder affects both the mother and the fetus in terms of health. Additionally, mortality may be a possibility. It tends to occur after the 20th week of pregnancy, even in the absence of medical documentation indicating hypertension which, the systolic and diastolic blood pressures constantly exceed 140 mmHg and 90 mmHg, respectively. In addition, the presence of protein in the urine is a critical symptom. Hypertension is considered a notable indicator of preeclampsia. Globally, preeclampsia claims the lives of approximately 76,000 women and 500,000 infants across the globe. Eight to ten percent of pregnancies are impacted[25].

There are several approaches to mitigating the severity of preeclampsia, including the division of high-risk groups that necessitate constant monitoring to prevent the condition during pregnancy. The following symptoms have been identified as representing the high-risk category: Pregnant women who experience symptoms of epigastric tightness, blurred vision, pain, swelling, high blood pressure, protein in the urine, 40-year-old-over pregnant women, having medical history of preeclampsia, or have congenital disorders including to obesity, hypertension, renal dysfunction, diabetes, immunodeficiency, or abnormal weight gain.

To investigate the current state of research concerning the risk of preeclampsia, we examined artificial intelligence methods utilized to predict the occurrence of preeclampsia in expectant women, as well as prevention guidelines that account for potential preeclampsia initiating factors [5]. As anticipated, artificial intelligence technology has taken on a progressively significant role in addressing numerous challenges within the context of medical science by helping in decision-making and effective treatment planning [3].

The effectiveness of artificial intelligence is dependent on the quality of the data. In addressing the issue of unbalanced data, the unbalanced distribution of information poses an intriguing challenge because if the quantity and quality of the data are sufficient to discern patterns within the data, then a machine learning model may exhibit respectable performance[6]. Conversely, the model may have lacked sufficient intelligence to learn and classify data for only a small portion of the collection [15]. Typically, significant

imbalances occur in data in the actual world. Imbalanced datasets frequently exhibit a relatively small population size for the class of interest in comparison to the other classes, leading to substantial inconsistencies in the data and disparities in proportions between classes, especially for medical and health data [7].

In this study, our objective was to rectify the problem of imbalanced data pertaining to the risk of preeclampsia through the implementation of hybrid, undersampling, and oversampling techniques. Additionally, an assessment was conducted on the efficacy of specific machine learning models, including Extreme Gradient Boosting (XGBoost), Naive Bayes, Decision Tree, and Random Forest. As our data characteristic, there were three distinct groups in the datasets: normal, risk, and preeclampsia. To consider how data unbalancing, the data ratio between the majority data sample represented by the normal class and the minority data sample represented by the risk class was measured roughly 76. Moreover, a data ratio between the majority data samples (the risk group) and a minority of data (Preeclampsia group) is roughly 14. The calculated ratios were explicitly defined as imbalanced data [13].

2 Literature Reviews

2.1 Data Pre-processing

Data Pre-processing addressing the issue of unbalanced data during data distribution presents an intriguing challenge [6]. Too small portion of data directly impacts to the quality of machine learning models because the models learn the pattern from the majority of data [15]. To effectively implement the sampling method, the unbalanced data should be transformed into numerical data prior to addressing the problem [13].

Our unbalanced dataset consisted of three groups: 9,341 records for the normal group, 123 records for the high-risk group, and 9 records for the preeclampsia group. To guarantee an enhancement in the quality of our dataset, it was therefore essential to implement the potential data unbalancing technique. Our approach involved boosting the proportion of the small data set, such as the preeclampsia group, while decreasing the size of the large data set, such as the normal group [16]. Subsequently, the quantity of data in each class should be approximately equivalent to that of the high-risk group. Eventually, each cohort comprised a substantial quantity of data.

As an indicator of data imbalance, the data imbalance ratio compares the proportions of the majority and minority data groups [13]. The ratio is ranged between zero to one. The nearly one ratio means data more balanced and vice versa [17]. As previously stated, medical science datasets were more susceptible to the problem of data imbalance than those of academic fields. A review of several prior studies led us to the conclusion that they had addressed the issue at hand, such as the diagnosis of lung cancer [13], genetic material identification [10], medical adverse effects [22], and so forth.

The oversampling methods were as follows: Random Oversampling (ROS), Adaptive Synthetic Sampling (ADASYN), Synthetic Minority Over-sampling Technique (SMOTE), and Borderline-SMOTE.

(1) ROS is an early data analysis technique that involves the random addition of minority class samples to the data [1].
(2) ADASYN divides minority class data samples according to learning difficulty and employs a weighted distribution. A smaller subset of the data in a group serves to generate a majority sample, which is a more challenging set to grasp [9].
(3) SMOTE uses KNN to pre-synthesize estimates on the minority data [4].
(4) Borderline-SMOTE applies SMOTE to data on the borderline of the data samples frequently misclassified as demographic clusters [8].

The undersampling methods are as follows: Random Undersampling (RUS), Tomek Link (Tomek), Edited Nearest Neighbors (ENN), and Repeated Edited Nearest Neighbors (RENN).

(1) RUS is the earliest technique since its inception for undersampling. By doing so, the random sample is diminished to encompass the majority group [21].
(2) Tomek evaluates two data samples as potentially related if they are adjacent and of distinct data types. Error-containing (Error data) or noisy (Noise data) data, as well as samples from the majority of data categories, are eliminated [24].
(3) ENN was applied to every sample that was evaluated by KNN on the remaining sample data. Erroneously classified segments of sample data are discarded. Additionally, the corrected data set comprises the remaining samples [26].
(4) RENN is the iterative application of ENN until the elimination of the untrainable data set no longer has an impact [26].

Hybrid resampling methods, which combine and pair undersampling of majority data and oversampling of minority data, were utilized to increase the population of minority groups while decreasing the population of majority data.

2.2 Classifiers

The study applied four machine learning classification methods, namely Decision Tree, Naive Bayes, Random Forest, and XGBoost, to develop models for predicting unbalanced medical data. The performance of the models was evaluated using the Area Under the Curve (AUC) measure, as defined by [11, 19].

The Decision Tree algorithm is a powerful tool for tackling classification problems. Krawczyk et al. [14] applied Decision Trees to detect data with imbalanced characteristics, a prominent trait in actual datasets. In addition, Sanz et al. [23] deployed Decision Trees on 11 datasets to successfully solve the issue of imbalanced data across several categories, yielding promising experimental outcomes [20].

Naive Bayes, as studied by Yap et al. [28], did a comparative investigation of machine learning algorithms to determine the most effective technique for managing imbalanced datasets. In their study, Bhandari et al. [2] applied the Naive Bayes algorithm to evaluate datasets with a high number of dimensions and imbalanced data. They demonstrated the usefulness of Naive Bayes in reliably recognizing different populations within these datasets. Moreover, Naive Bayes is widely seen in the literature while dealing with the classification of medical research data [11].

The Random Forest technique, introduced by Wu et al. [27], has proven to be a flexible and successful solution for text classification

tasks that deal with imbalanced data. Khushi et al. [13] confirmed its higher performance in a lung cancer diagnostic trial compared to competing approaches. In their research work, Khabsa et al. [12] suggested adopting Random Forest as a good strategy for categorizing imbalanced data.

XGBoost, as noted by Papacharalampous et al. [19], demonstrated great performance as compared to other tree-based classification algorithms when applied to rainfall datasets. Ogunleye et al. [18] suggested the application of XGBoost for the diagnosis of chronic renal sickness, offering additional proof of its usefulness in dealing with diverse types of datasets.

3 Methodology

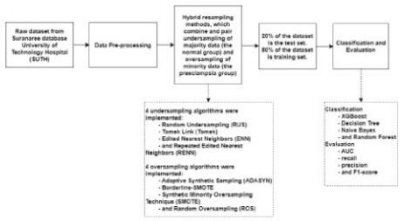

Figure 1: The framework of methodology

Figure 1 presents the methods and steps of the methodology. The purpose of this research was to find the appropriate resampling method to deal with imbalanced data sets related to preeclampsia in pregnancy. These datasets from the database of Suranaree University of Technology Hospital (SUTH) have three classes: the normal group, the risk group, and the preeclampsia group. As indicated by the data ratio, the distribution of the data in this set was unbalanced. Hence, the risk of preeclampsia, as assessed by the four chosen classifiers, can be predicted through the implementation of data imbalance methods.

3.1 Data Pre-processing

Data preparation required various actions, including converting the data into a suitable format for analysis, arranging the data for inquiry, modifying data attributes, and addressing missing data. The goal of the tests was to create models and determine the most effective approach for efficiently addressing data imbalances. Significant data items were acquired from patient histories, such as pregnancy age, number of pregnancies, and symptoms including headache, impaired vision, swollen tongue, and general edema. The information was separated into three independent groups: normal, at-risk, and preeclampsia. Notably, to determine the danger group, we carefully distinguished those normal pregnancy groups that incorporate damaging material. A case of a mother enduring hypertension without being diagnosed with preeclampsia is one example. Her disease necessitates more regular monitoring by doctors compared to a normal pregnancy, which supports her identification as a high-risk person.

The at-risk group comprised patients who possessed significant attributes, including raised blood pressure and protein in the urine, headache, swollen tongue, blurred vision, pre-existing pregnancies characterized by such symptoms, patients aged 40 and above, a prior diagnosis of preeclampsia, chronic illnesses including hypertension, kidney disease, diabetes, autoimmune disorders, obesity, abnormal weight gain in comparison to pregnancy weight gain standards, and fatigue. The patient became immediately classed into the risk group in the event that any symptoms showing a sign of the normal pregnancy group were detected.

Then, the data transformation was conducted. We changed categorical data to numerical data to prepare the process of data balance. This study utilized four separate undersampling techniques: RUS, Tomek, ENN, RENN, as well as four distinct oversampling techniques: Borderline-SMOTE, ROS, ADASYN, SMOTE.

3.2 Classification and Evaluation

To assess the efficacy of our resampling approaches, we choose four classification techniques: Decision Tree, Naive Bayes, Random Forest, and XGBoost. We deployed each model using its default parameter settings to preserve a uniform basis for comparison. The study employed these algorithms to validate which method of data unbalancing handling was suitable for imbalanced medical data.

This investigation implements the Area Under the Curve (AUC) as a metric to assess the performance of the classifier. A relationship between the True Positive Rate (TPR) and False Positive Rate, which ranges from 0 to 1, is expressed by the area under the curve. Effectiveness of the model increases as the value approaches 1. AUC is applicable to model performance evaluation in unbalanced data sets [11, 19]. Furthermore, the effectiveness of the models is evaluated through the utilization of recall, precision, and F1-score metrics.

4 Results

Naïve Bayes, the tomek_adasyn had the best performance. These are shown in Table 1, which produced the following superior values for AUC, F1-score, Recall, and Precision: 0.7769, 0.6970, 0.7325, and 0.7564, respectively.

Decision Tree, the renn_smote had the best performance in AUC and Precision, enn_adasyn. The renn_adasyn had the best performance in F1-score and Recall. These are shown in Table 2, which resulted in the following more effective values for AUC, F1-score, Recall, and Precision: 0.9749, 0.9688, 0.9722, and 0.9697, respectively.

Random Forest, the tomek_smote had the most significant performance in AUC, F1-score, and Recall. The tomek_adasyn had the best result in Precision. These are shown in Table 3, which produced the following superior values for AUC, F1-score, Recall, and Precision: 0.9929, 0.9730, 0.9785, and 0.9706, respectively.

XGBoost, the tomek_ros had the best performance. These are shown in Table 4, which produced the following superior values for AUC, F1-score, Recall, and Precision: 1.0000, 0.9864, 0.9892, and 0.9841, respectively.

Table 1: Results for Naïve Bayes on test set

Model	Technique	AUC	F1-score	Recall	Precision
Naive Bayes	enn_adasyn	0.7399	0.6692	0.6952	0.7183
	enn_borderline_smote	0.7442	0.6587	0.6616	0.6615
	enn_ros	0.7442	0.6587	0.6616	0.6615
	enn_smote	0.7442	0.6587	0.6616	0.6615
	renn_adasyn	0.7399	0.6692	0.6952	0.7183
	renn_borderline_smote	0.7442	0.6587	0.6616	0.6615
	renn_ros	0.7442	0.6587	0.6616	0.6615
	renn_smote	0.7442	0.6587	0.6616	0.6615
	ras_adasyn	0.4907	0.4641	0.4783	0.4803
	ras_borderline_smote	0.5439	0.2922	0.3484	0.2530
	ras_ros	0.5715	0.3906	0.3977	0.3982
	ras_smote	0.4914	0.3930	0.3880	0.4049
	tomek_adasyn	**0.7769**	**0.6970**	**0.7325**	**0.7564**
	tomek_borderline_smote	0.7082	0.6727	0.6839	0.6834
	tomek_ros	0.7363	0.5211	0.6667	0.4641
	tomek_smote	0.7423	0.6686	0.7075	0.7156

Table 2: Results for Decision Tree on test set

Model	Technique	AUC	F1-score	Recall	Precision
Decision Tree	enn_adasyn	0.9747	0.9688	**0.9722**	0.9683
	enn_borderline_smote	0.9522	0.9380	0.9420	0.9444
	enn_ros	0.9633	0.9535	0.9565	0.9565
	enn_smote	0.9633	0.9535	0.9565	0.9565
	renn_adasyn	0.9747	0.9688	**0.9722**	0.9683
	renn_borderline_smote	0.9633	0.9535	0.9565	0.9565
	renn_ros	0.9417	0.9223	0.9275	0.9333
	renn_smote	**0.9749**	**0.9690**	0.9710	**0.9697**
	ras_adasyn	0.8801	0.8565	0.8522	0.8629
	ras_borderline_smote	0.8171	0.7823	0.7851	0.7816
	ras_ros	0.8769	0.8473	0.8555	0.8488
	ras_smote	0.8464	0.8135	0.8185	0.8157
	tomek_adasyn	0.9597	0.9507	0.9507	0.9507
	tomek_borderline_smote	0.9208	0.9048	0.9070	0.9032
	tomek_ros	0.9689	0.9598	0.9677	0.9565
	tomek_smote	0.9335	0.9190	0.9237	0.9164

Table 3: Results for Random Forest on test set

Model	Technique	AUC	F1-score	Recall	Precision
Random Forest	enn_adasyn	0.9876	0.9383	0.9371	0.9400
	enn_borderline_smote	0.9806	0.9556	0.9565	0.9593
	enn_ros	0.9795	0.9556	0.9565	0.9593
	enn_smote	0.9818	0.9556	0.9565	0.9593
	renn_adasyn	0.9900	0.9546	0.9547	0.9562
	renn_borderline_smote	0.9793	0.9556	0.9565	0.9593
	renn_ros	0.9780	0.9431	0.9458	0.9431
	renn_smote	0.9805	0.9431	0.9458	0.9431
	ras_adasyn	0.9609	0.9120	0.9088	0.9171
	ras_borderline_smote	0.9534	0.9005	0.9001	0.9046
	ras_ros	0.9636	0.9157	0.9168	0.8902
	ras_smote	0.9683	0.9157	0.9168	0.9219
	tomek_adasyn	0.9892	0.9624	0.9583	**0.9706**
	tomek_borderline_smote	0.9889	0.9592	0.9618	0.9571
	tomek_ros	0.9859	0.9603	0.9640	0.9586
	tomek_smote	**0.9929**	**0.9730**	**0.9785**	0.9697

Table 4: Results for XGBoost on test set

Model	Technique	AUC	F1-score	Recall	Precision
XGBoost	enn_adasyn	0.9779	0.9531	0.9547	0.9522
	enn_borderline_smote	0.9888	0.9535	0.9565	0.9565
	enn_ros	0.9888	0.9535	0.9565	0.9565
	enn_smote	0.9888	0.9535	0.9565	0.9565
	renn_adasyn	0.9779	0.9531	0.9547	0.9522
	renn_borderline_smote	0.9888	0.9535	0.9565	0.9565
	renn_ros	0.9888	0.9535	0.9565	0.9565
	renn_smote	0.9888	0.9535	0.9565	0.9565
	ras_adasyn	0.9530	0.8995	0.9017	0.8999
	ras_borderline_smote	0.9358	0.9038	0.8995	0.9144
	ras_ros	0.9683	0.8868	0.8915	0.9219
	ras_smote	0.9410	0.8934	0.8947	0.8954
	tomek_adasyn	0.9953	0.9751	0.9722	0.9798
	tomek_borderline_smote	0.9806	0.9720	0.9667	0.9798
	tomek_ros	**1.0000**	**0.9864**	**0.9892**	**0.9841**
	tomek_smote	0.9934	0.9720	0.9667	0.9798

5 Discussion

Based on the actual data developed by employing the Naïve Bayes classifier, it was concluded that the tomek_adasyn model offered higher values for AUC, F1-score, Recall, and Precision: 0.7769, 0.6970, 0.7325, and 0.7564, respectively. In the instance of the decision tree, the renn_adasyn and enn_adasyn processes yielded the largest feasible F1-score and Recall values, which were 0.9688 and 0.9722, respectively. The optimum values for both AUC and Precision have been determined using the renn_smote data preparation method, resulting in values of 0.9697 and 0.9749, respectively.

When processing data in the tomek_smote format, the best results for Random Forest AUC, F1-score, and Recall were 0.9929, 0.9730, and 0.9785, respectively. Using the tomek_adasyn format, the highest Precision value of 0.9706 was reached. The most suitable values for AUC, F1-score, Recall, and Precision for XGBoost were obtained using the tomek_ros data preparation approach, with values of 1.0000, 0.9864, 0.9892, and 0.9841, respectively.

From this data, it can be claimed that the Naïve Bayes classifier offers pretty average results when compared with other models. Naïve Bayes performs well on data with independent features, which is rare in medical datasets where qualities are highly correlated. Conversely, it performs well with Decision Tree, Random Forest, and XGBoost models.

Decision Trees typically overfit owing to the over-memorization of training data, rendering them unable to accurately forecast incoming input, especially when dealing with complex datasets or uncontrolled tree depth. Random Forest, based on a bagging ensemble technique, reduces overfitting and increases model robustness by training each tree on unique subsets of features and a random sample of data. XGBoost, a boosting ensemble technique, exceeds both Decision Trees and Random Forest thanks to its superior optimization approaches, regularization, and handling of challenging datasets. Iteratively improving model predictions, XGBoost generates decision trees sequentially, with each tree focused on repairing the faults committed by the preceding ones. This extensive development makes it extremely accurate and robust.

6 Conclusion

This study tries to uncover the correct technique to handle the unequal clinical dataset related with preeclampsia. The study used four independent undersampling strategies (RUS, Tomek, ENN, RENN) and four distinct oversampling approaches (ADASYN, SMOTE, Borderline-SMOTE). An assessment of the efficacy of four prediction models was conducted: XGBoost, Decision Tree, Naive Bayes, and Random Forest. Each model was assessed using AUC, F1-score, Recall, and Precision. The best AUC, F1-score, Recall, and Precision

values for the dataset attained with tomek_ros and the XGBoost model were 1.0000, 0.9864, 0.9892, and 0.9841, respectively.

However, some of the limitations of the study include; the use of one dataset and therefore the study can only be applied narrowly. As for the future studies, the authors suggest including different dataset clarifying performance with cross-validation testing and using demographics or medical situations different from the current study. It is necessary to derive new treatments on the algorithm level like cost-sensitive learning or ensemble learning to evade biases and constraints of data level approaches. Thus, comparative analysis of the data level and algorithm level might allow in the future to make a concerted system in the form of a combined approach. Thus, future studies must broaden the variety of methods and data sets, which will enhance the robustness and practicality of resampling approaches in medical data analysis.

Acknowledgments

This work was supported by (i) Suranaree University of Technology (SUT), (ii) Thailand Science Research and Innovation (TSRI), (iii) National Science, and Research and Innovation Fund (NSRF) (NRIIS Project Number 179264). This work has been improved grammatically and organizationally through the use of QuillBot, ChatGPT, and Grammarly. The author expresses gratitude for the help of these tools, which contribute to improving the overall quality of this work.

References

[1] G. E. Batista, R. C. Prati, and M. C. Monard. 2004. A study of the behavior of several methods for balancing machine learning training data. *ACM SIGKDD Explorations Newsletter* (2004), 20–29. https://doi.org/10.1145/1007730.1007735

[2] Surya Mukesh Bhandari and Kunal Patel. 2015. A review on using clustering and classification techniques to predict student failure with high dimensional and imbalanced data. (2015).

[3] Carlos Briceño-Pérez, Liliana Briceño-Sanabria, and Paulino Vigil-De Gracia. 2009. Prediction and Prevention of Preeclampsia. *Hypertension in Pregnancy* (2009), 138–155. https://doi.org/10.1080/10641950802022384

[4] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer. 2002. SMOTE: Synthetic minority over-sampling technique. *Journal of Artificial Intelligence Research* (2002), 321–357. https://doi.org/10.1613/jair.953

[5] A. C. De Kat, J. Hirst, M. Woodward, S. Kennedy, and S. A. Peters. 2019. Prediction models for preeclampsia: A systematic review. *Pregnancy Hypertension* (2019), 48–66. https://doi.org/10.1016/j.preghy.2019.03.005

[6] S. Fotouhi, S. Asadi, and M. W. Kattan. 2019. A comprehensive data level analysis for cancer diagnosis on imbalanced data. *Journal of Biomedical Informatics* (2019). https://doi.org/10.1016/j.jbi.2018.12.003 Advance online publication.

[7] X. Guo, Y. Yin, C. Dong, G. Yang, and G. Zhou. 2008. On the Class Imbalance Problem. In *IEEE 2008 Fourth International Conference on Natural Computation*. 192–201. https://doi.org/10.1109/ICNC.2008.871

[8] H. Han, W. Y. Wang, and B. H. Mao. 2005. Borderline-SMOTE: A New Over-Sampling Method in Imbalanced Data Sets Learning. In *Advances in Intelligent Computing. ICIC 2005*, D. S. Huang, X. P. Zhang, and G. B. Huang (Eds.). 3644. https://doi.org/10.1007/11538059_91

[9] H. He, Y. Bai, E. A. Garcia, and S. Li. 2008. ADASYN: Adaptive synthetic sampling approach for imbalanced learning. In *2008 IEEE International Joint Conference on Neural Networks (IEEE World Congress on Computational Intelligence)*. 1322–1328. https://doi.org/10.1109/IJCNN.2008.4633969

[10] N. Herndon and D. Caragea. 2016. A Study of Domain Adaptation Classifiers Derived From Logistic Regression for the Task of Splice Site Prediction. *IEEE Transactions on NanoBioscience* (2016), 75–83. https://doi.org/10.1109/TNB.2016.2522400

[11] Harpreet Kaur, Himanshu S. Pannu, and Amandeep K. Malhi. 2019. A Systematic Review on Imbalanced Data Challenges in Machine Learning: Applications and Solutions. *Comput. Surveys* (2019). https://doi.org/10.1145/3343440

[12] Madian Khabsa, Ahmed Elmagarmid, Ihab Ilyas, Hossam Hammady, and Mourad Ouzzani. 2016. Learning to identify relevant studies for systematic reviews using random forest and external information. *Machine Learning* (2016), 465–482. https://doi.org/10.1007/s10994-015-5535-7

[13] M. Khushi, K. Shaukat, T. M. Alam, I. A. Hameed, S. Uddin, S. Luo, X. Yang, and M. C. Reyes. 2021. A Comparative Performance Analysis of Data Resampling Methods on Imbalance Medical Data. *IEEE Access* (2021). https://doi.org/10.1109/access.2021.3102399

[14] Bartosz Krawczyk, Michał Woźniak, and Gerald Schaefer. 2014. Cost-sensitive decision tree ensembles for effective imbalanced classification. *Applied Soft Computing* (2014), 554–562. https://doi.org/10.1016/j.asoc.2013.08.014

[15] Q. Li and Y. Mao. 2014. A review of boosting methods for imbalanced data classification. *Pattern Analysis and Applications* (2014), 679–693. https://doi.org/10.1007/s10044-014-0392-8

[16] R. Longadge and S. Dongre. 2013. Class imbalance problem in data mining review. (2013). https://doi.org/10.48550/arXiv.1305.1707

[17] N. Noorhalim, A. Ali, and S. M. Shamsuddin. 2019. Handling Imbalanced Ratio for Class Imbalance Problem Using SMOTE. In *Proceedings of the Third International Conference on Computing, Mathematics and Statistics (iCMS2017)*. 19–43. https://doi.org/10.1007/978-981-13-7279-7

[18] Ayokunle Ogunleye and Qian-Gen Wang. 2020. XGBoost Model for Chronic Kidney Disease Diagnosis. *IEEE/ACM Transactions on Computational Biology and Bioinformatics* (2020), 2131–2140. https://doi.org/10.1109/TCBB.2019.2911071

[19] Georgios Papacharalampous, Hristos Tyralis, Anastasios Doulamis, and Nikolaos Doulamis. 2021. Comparison of tree-based ensemble algorithms for merging satellite and earth-observed precipitation data at the daily time scale. *Hydrology* (2021). https://doi.org/10.3390/hydrology10020050

[20] Yoonsik Park and Joydeep Ghosh. 2014. Ensembles of (α)-Trees for Imbalanced Classification Problems. *IEEE Transactions on Knowledge and Data Engineering* (2014), 131–143. https://doi.org/10.1109/TKDE.2012.255

[21] J. Prusa, T. M. Khoshgoftaar, D. J. Dittman, and A. Napolitano. 2015. Using Random Undersampling to Alleviate Class Imbalance on Tweet Sentiment Data. In *2015 IEEE International Conference on Information Reuse and Integration*. 197–202. https://doi.org/10.1109/IRI.2015.39

[22] S. Santiso, A. Casillas, and A. Pérez. 2019. The class imbalance problem detecting adverse drug reactions in electronic health records. *Health Informatics Journal* (2019), 1768–1778. https://doi.org/10.1177/1460458218799

[23] Javier Alfonso Sanz, Daniel Bernardo, Francisco Herrera, Humberto Bustince, and Hani Hagras. 2015. A Compact Evolutionary Interval-Valued Fuzzy Rule-Based Classification System for the Modeling and Prediction of Real-World Financial Applications With Imbalanced Data. *IEEE Transactions on Fuzzy Systems* (2015), 973–990. https://doi.org/10.1109/TFUZZ.2014.2336263

[24] I. Tomek, A. Z. Broder, A. M. Bruckstein, and J. Koplowitz. 1985. Two modifications of CNN. *On the performance of edited nearest neighbor rules in high dimensions* (1985), 136–139. https://doi.org/10.1109/TSMC.1985.6313401

[25] P. K. Vata, N. Chauhan, A. Nallathambi, and F. E. R. Hussein. 2015. Assessment of prevalence of preeclampsia from Dilla region of Ethiopia. (2015). https://doi.org/10.1186/s13104-015-1821-5

[26] D. L. Wilson. 1972. Asymptotic Properties of Nearest Neighbor Rules Using Edited Data. *IEEE Transactions on Systems, Man, and Cybernetics* (1972), 408–421. https://doi.org/10.1109/TSMC.1972.4309137

[27] Qinglong Wu, Yuhang Ye, Huan Zhang, Michael K. Ng, and Su-En Ho. 2014. ForesTexter: An efficient random forest algorithm for imbalanced text categorization. *Knowledge-Based Systems* (2014), 105–116. https://doi.org/10.1016/j.knosys.2014.06.004

[28] Bee Wah Yap, Khairul Azmi Abu Rani, Hani Al-Faris Abd Rahman, Simon Fong, Zuraini Khairudin, and Nor Aniza Abdullah. 2014. An application of oversampling, undersampling, bagging and boosting in handling imbalanced datasets. In *Proceedings of the 1st International Conference on Advanced Data and Information Engineering*. 13–22. https://doi.org/10.1007/978-981-4585-18-7_2