A novel data balancing technique via resampling majority and minority classes toward effective classification

Mahmudul Hasan, Md. Fazle Rabbi, Md. Nahid Sultan, Adiba Mahjabin Nitu, Md. Palash Uddin
Department of Computer Science and Engineering, Hajee Mohammad Danesh Science and Technology University, Dinajpur, Bangladesh

Article Info

Article history:

Received Mar 26, 2023
Revised Jul 15, 2023
Accepted Jul 30, 2023

Keywords:

Divide and conquer
Heart disease prediction
Imbalance data handling
Machine learning
Medical informatics

ABSTRACT

Classification is a predictive modelling task in machine learning (ML), where the class label is determined for a specific example of predefined features. In determining handwriting characters, identifying spam, detecting disease, identifying signals, and so on, classification requires training data with many features and label instances. In medical informatics, high precision and recall are mandatory issues besides the high accuracy of the ML classifiers. Most of the real-life datasets have imbalanced characteristics that hamper the overall performance of the classifiers. Existing data balancing techniques perform the whole dataset at a time that sometimes causes overfitting and underfitting. We propose a data balancing technique that follows the divide and conquer procedure to cluster the dataset into several segments, and both oversampling and undersampling operation is performed on each cluster. Finally, the cluster joined together and built a balanced dataset. We chose the sample data of two heart disease datasets: Hungarian and Long Beach. Logistic regression and random forest classifier are the representatives of ML algorithms. We compare our proposed techniques with existing SMOTE, NearMiss, and SMOTETomek data balancing techniques. Both algorithms perform better on the proposed technique-balanced dataset. This technique can be the optimal solution for the imbalanced data handling strategy.

Corresponding Author:

Mahmudul Hasan
Department of Computer Science and Engineering
Hajee Mohammad Danesh Science and Technology University
Dinajpur-5200, Bangladesh
Email: mahmudulmoon123@gmail.com

1. INTRODUCTION

In practical use, classification models frequently face the imbalanced dataset problem, where the number of instances from the majority class is substantially more than those from the minority class, preventing the model from learning well from the minority class [1]. When the minority group's contributions to a dataset are increasingly crucial, such as disease diagnosis, churn, or fraud identification, this becomes a significant issue. Both oversampling the minority group and undersampling the majority group are standard methods for addressing this imbalanced dataset issue. The problem is that each of these methods has its weaknesses. The principle behind the oversampling vanilla method is to replicate a subset of the minority class at random; as a result, this approach does not generate any novel insights [2]. Undersampling involves deleting some random samples from the majority class, which results in losing some information in the original data. When the dataset is highly imbalanced, oversampling creates massive synthetic data for the minority class that reduce the variance of the class, causing oversampling and increasing the bias during classification [3]. Oversampling sometimes creates model overfitting, and undersampling causes the loss of information and reduces the performance of the classifiers [4]. Existing hybrid oversampling and undersampling methods try to fix the issues, but it fails for the data distributions in some domains, namely healthcare informatics, biostatistics, and bioinformatics. The instances of individual classes are close to each other and sometimes overlap. It misguides the machine learning (ML) classifiers during the time of classification and creates ambiguity during the learning stage of the ML models. Dataset balancing is one of the powerful preprocessing techniques in ML. Many researchers use this concept in different domains. Among many of the work in this research, Batista *et al.* [5] perform a comprehensive experimental evaluation comparing ten techniques dealing with the class imbalance problem on thirteen University of California (UCI) datasets. They found through their experiments that class differences do not consistently reduce the efficiency of learning systems. To detect the code smells, some researchers use ML and found this procedure offers a minimum performance due to the high imbalance characteristics of the dataset. They use synthetic minority oversampling technique (SMOTE) in preprocessing stage and conclude that data balancing does not dramatically improve the performance of the models [6]. The extended work of the same researchers [7] uses five different data-balancing techniques and shows their impact on code smell detection in object-oriented systems. The results demonstrate that skipping the balancing stage does not significantly impact accuracy. In another study Lemaıtre *et al.* [8] present the imbalanced-learn application programming interface (API), a Python toolbox to handle the imbalance datasets in ML. They discuss several existing data-balancing techniques and compare the models in binary and multiclass data balancing; additionally, they also present the techniques of the methods, either oversampling or undersampling. In heart disease prediction [9], this study uses a hybrid approach combining SMOTE with edited nearest neighbor (ENN) to balance the dataset. They use a balancing technique on electrocardiogram (ECG) data, train ML models, and show the result of balanced and imbalanced datasets separately. The hybrid SMOTE-ENN significantly increases the classifiers' performance, proving the importance of data balancing in healthcare. To find a suitable data balancing technique on heart disease cleveland dataset classification, Sahid *et al.* [10] use SMOTE, NearMiss, and synthetic minority oversampling technique Tomek links (SMOTETomek) as data balancing techniques and use ML and ensemble ML models to check the priority of the balancing techniques. Stroke prediction using ML methods among the older Chinese in a high imbalance dataset, authors use SMOTE in preprocessing phase and get a significant result on the performance of the classifiers. It helps the classifiers to show a stable result and improves the accuracy of the classifiers at a reasonable rate [11]. To apply deep learning to medical data, Zhang *et al.* [12] uses data balancing techniques on ECG data. They propose a data balancing technique, an agent-based model (ABM), that adopts the Gaussian Naive Bayes algorithm to estimate the object sample and use the entropy as a query function to evaluate the result. Sensitive domains like healthcare require high precision and recall for each class [13]. The high biases of the balanced data show unstable classification reports for different classes. Sometimes, the accuracy is satisfactory, but the precision and recall show huge fluctuations among the classes that could be better for the classifier's performance. To mitigate these issues in this study,

- We proposed a divide-and-conquer-based data balancing technique that controls the classifier's performance's stability, high accuracy and prevents overfitting and underfitting.
- To check the performance of the proposed data balancing technique, we evaluate all possible combinations of all balancing techniques and classifiers.
- We treat the noisy, missing values using random forest regression to turn the dataset to be more ML-trainable.

2. PROPOSED METHOD

2.1. Method overview

In this study, we take two healthcare informatics data as the sample of imbalanced data. Our proposed methodology includes data preprocessing, and we apply the data balancing techniques to the dataset individually and fit the balanced dataset to the ML algorithm logistic regression (LR) and random forest classifier (RFC). The top-down view of the proposed method is in Figure 1. We evaluate the performance of each combination using accuracy, precision, recall, and f1 score and finally show the combinations' receiver operating characteristic (ROC). To check the stability of the dataset in the different folds, we use stratified K-fold cross-validation on imbalance data and K-fold cross-validation on the balanced datasets. We compare the result of each combination with the proposed data balancing techniques combinations and all the results present in the result section.

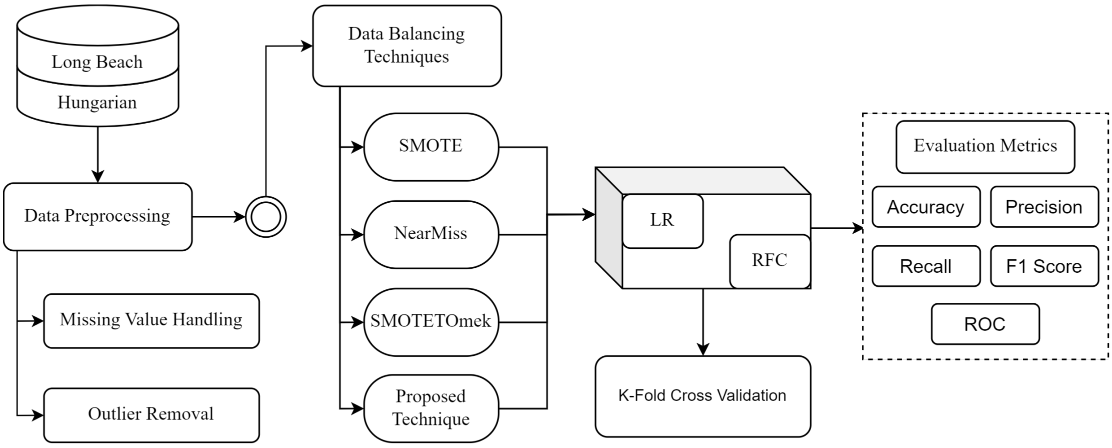

Figure 1. Top-down approach of proposed method

2.2. Datasets

We use two heart disease datasets, namely Hungarian and Long Beach Va Terrada [14]. Both datasets contain the same feature, and both are binary-labelled datasets. All the datasets contain 13 standard features (age, sex, ChestPainType, RestingBP, cholesterol, FastingBS, RestingECG, MaxHR, ExerciseAngina, oldpeak, and ST_Slope) and a target feature (heart disease). Hungarian contains 294 instances (106 in class 0 and 188 in class 1), and Long Beach Va contains 200 instances (51 in class 0 and 149 in class 1).

2.3. Preprocessing techniques

The two datasets are noisy and have many values that need to be added. We apply random forest regression [15] to find the value in missing places and fill in the data. The missing position is considered a dependent variable, and other features are independent variables; then, the regression output is put in the missing value section. This process performs on the columns that have a large number of missing values. For the columns that contain a few missing values, we handle it to fill the data by the arithmetic mean of the column.

We also check the outliers of the datasets and remove the outlier using turkey fences [16]. The dataset is split into three quartiles: Q1, Q2, and Q3. The first quartile, or Q1, is the value within the data set comprising 25% values below it. The third quartile, or Q3, is the value that accounts for 25% of the values above it. Outliers are also valued below or above the lower or upper limits, as (1) and (2):

$$lower\_limit = Q_1 - 1.5(Q_3 - Q_1) \tag{1}$$

$$lower\_limit = Q_1 + 1.5(Q_3 - Q_1) \tag{2}$$

Outliers that fall below the lower limit are replaced with a lower limit, and outliers above the upper limit are replaced with an upper limit.

2.4. Baseline data balancing techniques

The data are balanced using SMOTE, NearMiss, and combined over and undersampling techniques SMOTETomek. The effects of each technique on our suggested method are tabulated in the results section.

SMOTE: to address imbalanced data, the SMOTE stands out as a widely utilized approach [17]. This technique involves creating synthetic instances for the underrepresented class, enhancing the dataset without sacrificing information from the original records, thus contributing additional data points.

NearMiss: one common approach taken by NearMiss to rectify the problem of skewed data was to employ an undersampling machine learning method. It eliminates random samples from the majority group, which can lead to data loss. Hence, an underfitting model problem may result from a specific scenario.

SMOTETomek: to deal with unbalanced datasets, SMOTETOMEK employs a hybrid ML strategy [18]. It is a hybrid method that employs both undersampling and oversampling. As a result, the performance measures used for classifying data either move up or down depending on the dataset's underlying statistical properties.

2.5. Proposed data balancing technique

The above data-balancing techniques use either oversampling or undersampling to balance the dataset, where the total dataset considers as a cluster and performs the balancing operation. In our proposed data balancing techniques, we divide the dataset into several clusters based on the data characteristics determined by K-means clustering [19]. We balance each cluster separately, then merge the individual cluster and create the final balance dataset. In each cluster, we find the majority and minority class first, then apply the resampling techniques. In existing approaches, the majority and minority class are fixed, but in our proposed techniques, the majority and minority class changes based on the data sample of individual clusters. In each cluster, we choose random data from the minority class and calculate the distance between the random data and its k nearest neighbors. Then, we multiply the distance by a random number between 0 and 1 and add the new data point as a synthetic sample for the minority class. This step is continued until the minority class meets the desired proportion. Then, we choose another random data from the majority class and check the nearest neighbors of the random data point. If the neighbor data are from a minority class, we remove the random data point. Selecting the observation x and y needs to fulfil the following properties:

- The nearest neighbors of observation x is y
- The nearest neighbors of observation y is x
- Both x and y belong to a different class.

It means x and y belong to the majority and minority classes, and we select the two as a pair. Consider $d(x_i, x_j)$ denotes the Euclidean distance between the data point x_i and x_j, where x_i denotes the minority class sample and x_j denotes the majority class sample. If there is no sample x_k satisfies the following condition:

- $d(x_i, x_k) < d(x_i, x_j)$
- $d(x_j, x_k) < d(x_i, x_j)$

then the pair of $d(x_i, x_j)$ is the selected pair.

This technique can be used to identify and eliminate data samples from the majority class with the smallest Euclidean distance to the data from the minority class (i.e. the data from the majority class that is closest to the data from the minority class, thus making it ambiguous to differentiate).

2.6. Machine learning algorithms

LR: LR serves as a statistical model that assesses the likelihood of an event transpiring based on an analysis of certain independent variables. This model is particularly geared towards tackling classification challenges [20]. Notably, LR entails binary outcomes: the event either materializes or does't come to pass [21].

RFC: RFC represents an expanded iteration of bagging, an ensemble technique, and is fashioned through the amalgamation of numerous decision trees [22]. This approach addresses overfitting by opting for a subset of potential features while creating decision trees, in contrast to decision trees considering the entirety of features. The process involves crafting distinct decision trees from the training dataset, subsequently amalgamating these trees' outcomes to yield the ultimate result. In classification tasks, RFC employs a voting mechanism wherein the class with the highest votes count is selected as the final prediction [23].

2.7. Performance measure techniques

The purpose of the classification report is to evaluate classifier performance. Accuracy is among the metrics that can be employed to gauge the efficacy of classification algorithms. The precision of the test determines the count of samples predicted to be positive that indeed turn out to be positive [24]. This metric proves valuable when the aim is to minimize false positives. Recall serves as an indicator of how effectively optimistic predictions capture positive samples [25]. The F1 score, a harmonic average of precision and recall, offers a comprehensive synthesis of both measurement approaches. It holds the potential to outperform accuracy in scenarios involving imbalanced binary classification datasets. Furthermore, the receiver operating characteristics (ROC) curve contrasts the false positive rate (FPR) against the true positive rate (TPR). This graphical representation aids in assessing classifier performance across different thresholds.

3. RESULT AND DISCUSSION

In this study, we propose a divide-and-conquer-based data balancing technique and compare the classifier's performance to find the superiority of this proposed technique. Firstly, we apply LR and RFC in the imbalance dataset and then use SMOTE. NearMiss, SMOTETomek and proposed balancing techniques one by one and check the performance of the models.

In Table 1, the state of the data presents before and after data balancing. We show the five different states of both datasets. It shows that the datasets are imbalanced, and after applying the balancing techniques, it is balanced; sometimes, the instances increase and sometimes decrease. We use the balanced dataset separately from the ML classifiers and check the results using the classification report.

Table 1. State of the data sample before and after data balancing

State	Long Beach		Hungarian	
Class	Class 0	Class 1	Class 0	Class 1
Original dataset	51	149	106	188
SMOTE	149	149	188	188
NearMiss	51	51	106	106
SMOTETomek	51	149	106	188
Proposed balancing technique	134	134	173	173

3.1. Performance of the classifiers on imbalance datasets

We apply LR and RFC on both datasets. Table 2 shows the result of Long Beach and Hungarian datasets resulting in an imbalanced state. The ML classifiers' performance is not good, and individual classes' precision, recall, and f1 score are unstable. In Table 2, the precision and recall are comparatively low for class 0 than class 1 for both classifiers. The performance in Table 2 has a significant gap, but it is less than in Table 2. Because the Long Beach data is more imbalanced than the Hungarian dataset, to solve this issue, we need to use data balancing techniques to balance the data to get a stable output and good accuracy.

Table 2. Performance of the classifier without balancing on Long Beach and Hungarian dataset

Algorithms	Class	Long Beach dataset				Hungarian dataset			
		Precision	Recall	F1 Score	Accuracy	Precision	Recall	F1 Score	Accuracy
LR	0	0.62	0.27	0.37	0.78	0.75	0.59	0.66	0.77
	1	0.79	0.94	0.86		0.78	0.88	0.83	
RF	0	0.63	0.40	0.49	0.79	0.74	0.61	0.67	0.77
	1	0.82	0.92	0.87		0.79	0.87	0.83	

3.2. Performance of the classifiers after balancing in Long Beach dataset

Firstly, we use the balancing techniques on the Long Beach dataset and results are tabulated in Tables 3 and 4. The table indicates that the proposed data balancing technique outperforms other balancing techniques in both classifiers. In the imbalance Long Beach dataset, the accuracy of LR is 78% and RF is 79%, but the other metrics are unstable for both classes. After applying the data balancing techniques, we improved the stability of the other performance measurement techniques, but accuracy fell in SMOTE, NearMiss, and SMOTETomek. Nevertheless, the proposed data-balancing technique shows a different scenario; it improves the classification accuracy and stabilizes the other performance measurement techniques for both classes. It is proof of the superiority of the proposed technique. The performance of each combination is in Figure 2 by a ROC. We also apply the 10-fold cross-validation to imbalance and balanced datasets and get a slight standard deviation in average accuracy. Each fold shows good performance, and most cases are stable.

3.3. Performance of the classifiers after balancing in Hungarian dataset

We apply the same methodology in the Hungarian dataset to prove the superiority of the proposed data balancing technique. Like the previous dataset, our proposed technique performs better than other data balancing techniques. In the imbalance phase, LR shows 77%, and RF shows 77% accuracy also. Our proposed methods help the classifiers, and the accuracy goes LR from 77% to 89% and RF from 77% to 91%. Comparing the Table 2 with Table 5 shows that the result is stable in the balanced dataset compared to an imbalanced dataset. Precision and recall are stable in SMOTE balancing, but our proposed techniques show more balanced results and better accuracy of the classifiers. We visualize the ROC of all possible combinations in Figure 3, which shows the superiority of the proposed data balancing technique in Table 6, we show the 10-fold cross-validation score of the algorithms and get a small standard deviation in every case. The result clearly indicates that we can choose RF as a classifier to predict heart disease after applying the proposed data balancing technique.

Table 3. Performance of the classifiers after balancing on Long Beach dataset

Balancing technique	Algorithms	Class	Precision	Recall	F1 score	Accuracy
SMOTE	LR	0	0.71	0.73	0.72	0.72
		1	0.72	0.70	0.71	
	RFC	0	0.86	0.83	0.85	0.85
		1	0.84	0.87	0.85	
NearMiss	LR	0	0.56	0.56	0.56	0.62
		1	0.67	0.67	0.67	
	RFC	0	0.56	0.56	0.56	0.62
		1	0.67	0.67	0.67	
SMOTETomek	LR	0	0.50	0.25	0.33	0.70
		1	0.74	0.89	0.81	
	RFC	0	0.75	0.50	0.60	0.80
		1	0.81	0.93	0.87	
Proposed technique	LR	0	0.76	0.76	0.76	0.78
		1	0.79	0.79	0.79	
	RFC	0	0.92	0.88	0.90	0.91
		1	0.90	0.93	0.92	

Table 4. Stratified K fold and K fold cross validation score for Long Beach dataset

State	Algorithms	F1	F2	F3	F4	F5	F6	F7	F8	F9	F10	Average ± std
Original data	LR	0.80	0.75	0.80	0.75	0.70	0.75	0.70	0.80	0.75	0.55	0.74±0.07
	RFC	0.85	0.75	0.80	0.85	0.70	0.70	0.75	0.75	0.85	0.60	0.76±0.07
SMOTE	LR	0.73	0.70	0.70	0.67	0.73	0.77	0.70	0.90	0.93	0.69	0.75±0.09
	RFC	0.70	0.67	0.67	0.83	0.90	0.87	0.90	1.0	0.96	0.89	0.84±0.12
NearMiss	LR	0.82	0.36	0.90	0.80	0.40	0.84	0.80	0.70	0.90	0.60	0.71±0.18
	RFC	0.64	0.36	0.70	0.80	0.50	1.00	0.60	0.60	0.60	0.70	0.65±0.16
SMOTETomek	LR	0.80	0.75	0.80	0.75	0.70	0.75	0.70	0.80	0.75	0.55	0.74±0.07
	RFC	0.80	0.70	0.85	0.90	0.65	0.70	0.75	0.80	0.80	0.75	0.77±0.07
Porposed technique	LR	0.74	0.78	0.81	0.70	0.81	0.78	0.78	0.74	0.92	0.65	0.77±0.07
	RFC	0.78	0.93	0.95	0.81	0.85	0.89	0.89	0.74	0.96	0.77	0.86±0.08

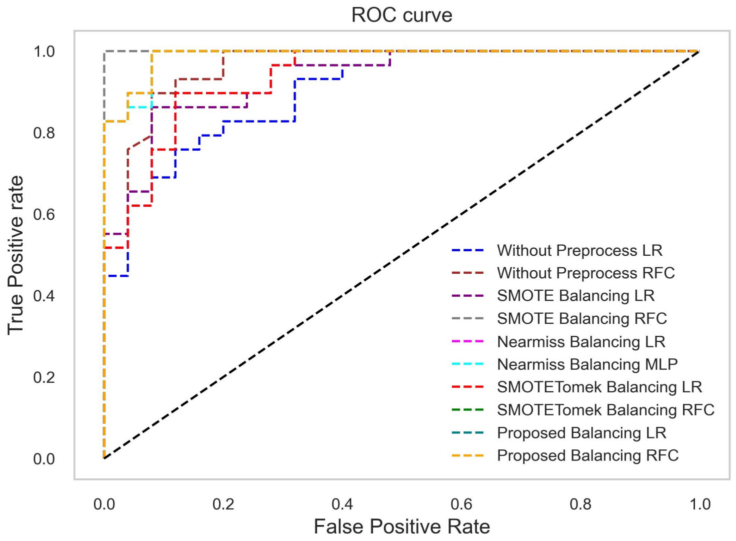

Figure 2. ROC curve of all possible combinations in Long Beach dataset

Table 5. Performance of the classifiers after balancing on Hungarian dataset

Balancing technique	Algorithms	Class	Precision	Recall	F1 Score	Accuracy
SMOTE	LR	0	1.00	0.76	0.86	0.87
		1	0.77	1.00	0.87	
	RFC	0	0.94	0.79	0.86	0.86
		1	0.78	0.94	0.85	
NearMiss	LR	0	0.67	0.78	0.72	0.67
		1	0.69	0.55	0.61	
	RFC	0	0.69	0.78	0.73	0.70
		1	0.71	0.60	0.65	
SMOTE	LR	0	0.84	0.75	0.79	0.81
		1	0.79	0.87	0.83	
Tomek	RFC	0	0.85	0.79	0.81	0.83
		1	0.82	0.87	0.84	
Proposed Technique	LR	0	0.88	0.88	0.88	0.89
		1	0.89	0.89	0.89	
	RFC	0	0.91	0.91	0.91	0.91
		1	0.92	0.92	0.92	

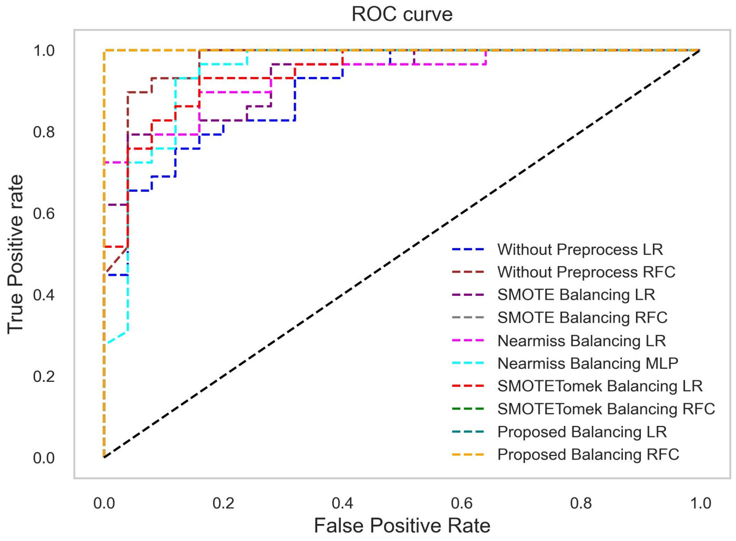

Figure 3. ROC curve of all possible combinations in Hungarian dataset

Table 6. Stratified K fold and K fold cross validation score for Hungarian dataset

State	Algorithms	F1	F2	F3	F4	F5	F6	F7	F8	F9	F10	Final
Original data	LR	0.83	0.83	0.87	0.97	0.76	0.86	0.83	0.79	0.83	0.72	0.83±0.06
	RFC	0.77	0.73	0.87	0.833	0.76	0.83	0.86	0.72	0.86	0.76	0.80±0.05
SMOTE	LR	0.79	0.89	0.89	0.74	0.79	0.82	0.92	0.84	0.84	0.73	0.83±0.06
	RFC	0.84	0.87	0.84	0.76	0.84	0.89	0.95	0.86	0.92	0.76	0.85±0.06
NearMiss	LR	0.77	0.73	0.71	0.86	0.76	0.76	0.86	0.76	1.00	0.81	0.80±0.08
	RFC	0.64	0.77	0.71	0.86	0.76	0.67	0.86	0.67	0.90	0.71	0.76±0.09
SMOTETomek	LR	0.83	0.83	0.87	0.97	0.76	0.86	0.83	0.79	0.83	0.72	0.83±0.06
	RFC	0.77	0.77	0.90	0.83	0.72	0.83	0.86	0.72	0.86	0.72	0.80±0.06
Porposed technique	LR	0.75	0.86	0.89	0.86	0.83	0.86	0.97	0.89	0.82	0.72	0.86±0.07
	RFC	0.89	0.91	0.97	0.89	0.83	0.80	0.94	0.92	0.85	0.82	0.88±0.05

4. CONCLUSION AND FUTURE WORK

This study proposes a new data-balancing technique to predict heart disease from two well-known datasets. We use LR and RF ML algorithms and first check the performance of the classifiers in the imbalanced dataset. The performance of the classifiers could be better level and precision, and recall could be more stable in both classes. Then, we apply three existing data-balancing techniques and check the performance of the classifiers. The overall performance sometimes falls, but precision and recall are more stable than the imbalanced dataset. In 10-fold cross-validation, the performance of the classifiers is stable, and accuracy fluctuates in a minimum range. The overall result shows that the proposed data balancing techniques outperform the three data balancing techniques, and RF shows better accuracy than LR. This study focuses on the binary classification problem, and the proposed data balancing technique is suitable for binary classification. The further study focuses on multi-label classification, and the different domains will consider as the study area.

REFERENCES

[1] A. S. Tarawneh, A. B. Hassanat, G. A. Altarawneh, and A. Almuhaimeed, "Stop oversampling for class imbalance learning: A review," in *IEEE Access*, vol. 10, pp. 47643-47660, 2022, doi: 10.1109/ACCESS.2022.3169512.

[2] G. Douzas and F. Bacao, "Effective data generation for imbalanced learning using conditional generative adversarial networks," *Expert Systems with applications*, vol. 91, pp. 464–471, 2018, doi: 10.1016/j.eswa.2017.09.030.

[3] V. P. K. Turlapati and M. R. Prusty, "Outlier-SMOTE: A refined oversampling technique for improved detection of COVID-19," *Intelligence-based medicine*, vol. 3, p. 100023, 2020, doi: 10.1016/j.ibmed.2020.100023.

[4] S. Park and H. Park, "Combined oversampling and undersampling method based on slow-start algorithm for imbalanced network traffic," *Computing*, vol. 103, no. 3, pp. 401–424, 2021, doi: 10.1007/s00607-020-00854-1.

[5] G. E. Batista, R. C. Prati, and M. C. Monard, "A study of the behavior of several methods for balancing machine learning training data," *ACM SIGKDD explorations newsletter*, vol. 6, no. 1, pp. 20–29, 2004, doi: 10.1145/1007730.1007735.

[6] F. Pecorelli, D. Di Nucci, C. De Roover, and A. De Lucia, "On the role of data balancing for machine learning-based code smell detection," in *Proceedings of the 3rd ACM SIGSOFT international workshop on machine learning techniques for software quality evaluation*, 2019, pp. 19–24, doi: 10.1145/3340482.3342744.

[7] F. Pecorelli, D. Di Nucci, C. De Roover, and D. Lucia, "A large empirical assessment of the role of data balancing in machine-learning-based code smell detection," *Journal of Systems and Software*, vol. 169, p. 110693, 2020, doi: 10.1016/j.jss.2020.110693.

[8] G. Lemaître, F. Nogueira, and C. K. Aridas, "Imbalanced-learn: A python toolbox to tackle the curse of imbalanced datasets in machine learning," *The Journal of Machine Learning Research*, vol. 18, no. 1, pp. 559–563, 2017.

[9] U. Nagavelli, D. Samanta, and P. Chakraborty, "Machine learning technology-based heart disease detection models," *Journal of Healthcare Engineering*, vol. 2022, 2022, doi: 10.1155/2022/7351061.

[10] M. A. Sahid, M. Hasan, N. Akter, and M. M. R. Tareq, "Effect of Imbalance Data Handling Techniques to Improve the Accuracy of Heart Disease Prediction using Machine Learning and Deep Learning," *2022 IEEE Region 10 Symposium (TENSYMP)*, Mumbai, India, 2022, pp. 1–6, doi: 10.1109/TENSYMP54529.2022.9864473.

[11] Y. Wu and Y. Fang, "Stroke prediction with machine learning methods among older chinese," *International journal of environmental research and public health*, vol. 17, no. 6, p. 1828, 2020, doi: 10.3390/ijerph17061828.

[12] H. Zhang, H. Zhang, S. Pirbhulal, W. Wu, and V. H. C. D. Albuquerque, "Active balancing mechanism for imbalanced medical data in deep learning–based classification models," *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)*, vol. 16, no. 1s, pp. 1–15, 2020, doi: 10.1145/3357253.

[13] T. Ching, D. S. Himmelstein, B. K. Beaulieu-Jones, A. A. Kalinin, and T. Do, "Opportunities and obstacles for deep learning in biology and medicine," *Journal of The Royal Society Interface*, vol. 15, no. 141, p. 20170387, 2018, doi: 10.1098/rsif.2017.0387.

[14] O. Terrada, B. Cherradi, A. Raihani, and O. Bouattane "Atherosclerosis disease prediction using supervised machine learning techniques," in *2020 1st International Conference on Innovative Research in Applied Science, Engineering and Technology (IRASET)*,Meknes, Morocco, 2020, pp. 1-5, doi: 10.1109/IRASET48871.2020.9092082.

[15] I. H. Rahmana, A. R. Febriyani, I. Ranggadara, S. Suhendra, and I. S. Karima, "Comparative study of extraction features and regression algorithms for predicting drought rates," *TELKOMNIKA (Telecommunication Computing Electronics and Control)*, vol. 20, no. 3, pp. 638–646, 2022, doi: 10.12928/telkomnika.v20i3.23156.

[16] Q. Zhou, S. Li, X. Li, W. Wang, and Z. Wang, "Detection of outliers and establishment of targets in external quality assessment programs," textslClinica chimica acta, vol. 372, no. 1-2, pp. 94–97, 2006, doi: 10.1016/j.cca.2006.03.033.

[17] W. Gata and A. Bayhaqy, "Analysis sentiment about islamophobia when christchurch attack on social media," *Telkomnika (Telecommunication Computing Electronics and Control)*, vol. 18, no. 4, pp. 1819– 1827, 2020, doi: 10.12928/telkomnika.v18i4.14179.

[18] M. Alenezi, M. Akour, and O. Al Qasem, "Harnessing deep learning algorithms to predict software refac- toring," *Telkomnika (Telecommunication Computing Electronics and Control)*, vol. 18, no. 6, pp. 2977–2982, 2020, doi: 10.12928/telkomnika.v18i6.16743.

[19] M. M. Uddin, M. M. Rashid, M. Hasan, M. A. Hossain, and Y. Fang, "Investigating corporate environmental risk disclosure using machine learning algorithm," *Sustainability*, vol. 14, no. 16, p. 10316, 2022, doi: 10.3390/su141610316.

[20] R. Zhu, X. Hu, J. Hou, and X. Li, "Application of machine learning techniques for predicting the consequences of construction accidents in China," *Process Safety and Environmental Protection*, vol. 145, pp. 293–302, Jan. 2021, doi: 10.1016/j.psep.2020.08.006.

[21] R. Tamakloe, S. Das, E. N. Aidoo, and D. Park, "Factors affecting motorcycle crash casualty severity at signalized and non-signalized intersections in ghana: Insights from a data mining and binary logit regression approach," *Accident Analysis and Prevention*, vol. 165, p. 106517, 2022, doi: 10.1016/j.aap.2021.106517.

[22] M. Hasan, U. Das, R. K. Datta, and M. Z. Abedin, "Model development for predicting the crude oil price: Comparative evaluation of ensemble and machine learning methods," in *Novel Financial Applications of Machine Learning and Deep Learning: Algorithms, Product Modeling, and Applications*, Springer, 2023, pp. 167–179, doi: 10.1007/978–3–031–18 552–6-10.

[23] S. B. Atitallah, M. Driss, and I. Almomani, "A novel detection and multi-classification approach for iot-malware using random forest voting of fine-tuning convolutional neural networks," *Sensors*, vol. 22, no. 11, p. 4302, 2022, doi: 10.3390/s22114302.

[24] J. Jasmir, S. Nurmaini, R. F. Malik, and B. Tutuko, "Bigram feature extraction and conditional random fields model to improve text

classification clinical trial document," *TELKOMNIKA (Telecommunication Computing Electronics and Control)*, vol. 19, no. 3, pp. 886–892, 2021, doi: 10.12928/telkomnika.v19i3.18357.

[25] H. S. Sint and K. K. Oo, "Comparison of two methods on vector space model for trust in social commerce," *TELKOMNIKA (Telecommunication Computing Electronics and Control)*, vol. 19, no. 3, pp. 809–816, 2021, doi: 10.12928/telkomnika.v19i3.18150.

BIOGRAPHIES OF AUTHORS

Mahmudul Hasan (GSMIEEE) is currently pursuing his Ph.D. in Information Technology (IT) at Deakin University, Melbourne, Victoria, Australia. He completed his M.Sc. (Eng.) in CSE from the Department of Computer Science and Engineering (CSE) at Hajee Mohammad Danesh Science and Technology University, Dinajpur, Bangladesh, in 2023. He completed his BSc (Eng.) in CSE from the same university in 2021. He is a former lecturer at the Department of Computer Science and Engineering at the University of Creative Technology, Chittagong, Bangladesh. He is also one of the instructors of the first research BootCamp in Bangladesh. His research interest includes federated learning, blockchain, machine learning, deep learning, cyber security, health informatics, business intelligence, and computational sociology. He can be contacted at email: mahmudulmoon123@gmail.com.

Md. Fazle Rabbi is currently working as an Associate Professor at the Department of Computer Science and Engineering in Hajee Mohammad Danesh Science and Technology University, Dinajpur-5200. He received his B.Sc (Eng.) degree in Computer Science and Engineering from Hajee Mohammad Danesh Science and Technology University, Dinajpur, Bangladesh in 2008. He completed his M.Sc. degree in Computer Science and Engineering from Jahangirnagar University, Savar, Dhaka, Bangladesh in 2018. His main research interest is machine learning, bioinformatics, image processing, data structures, and algorithm. He has several scientific research publications in various aspects of Computer Science and Engineering. He can be contacted at email: rabbi@hstu.ac.bd.

Md. Nahid Sultan is currently working as an assistant professor at the Department of Computer Science and Engineering in Hajee Mohammad Danesh Science and Technology University, Dinajjpur-5200. He completed his M.Sc. and B.Sc. in Computer Science and Engineering from Islamic University, Bangladesh. His research interest is artificial intelligence, machine learning, deep learning, bioinformatics, and internet of things. He has several scientific research publications in various aspects of Computer Science and Engineering. He can be contacted at email: nahid.sultan@hstu.ac.bd.

Adiba Mahjabin Nitu is currently working as a professor at the Department of Computer Science and Engineering in Hajee Mohammad Danesh Science and Technology University, Dinajpur-5200. She received his M.Sc. degree in Computer Science and Engineering from the University of Northern British Columbia, Canada in 2015 and another M.Sc. degree from the University of Rajshahi, Bangladesh in 2004. She completed his B.Sc. degree in Computer Science and Engineering from the University of Rajshahi, Bangladesh in 2003. Her main research interest is artificial intelligence, simulation, modeling, and the internet of things (IoT). She has several scientific research publications in various aspects of Computer Science and Engineering. She can be contacted at email: nitu.hstu@gmail.com.

Md. Palash Uddin received a Ph.D. degree in Information Technology from Deakin University, Australia in 2023. He also received a B.Sc. degree in Computer Science and Engineering from Hajee Mohammad Danesh Science and Technology University (HSTU), Bangladesh, and an M.Sc. degree in Computer Science and Engineering from Rajshahi University of Engineering and Technology, Bangladesh. He is currently working as a Postdoctoral Research Fellow at the School of Information Technology, Deakin University, Australia. He is also an academic faculty member at HSTU, Bangladesh. His research interests include machine learning, federated learning, blockchain, and remote sensing image analysis. He can be contacted by email: palash_cse@hstu.ac.bd.