

## Imbalanced Flood Forecast Dataset Resampling Using SMOTE-Tomek Link

Ainaa Hanis Zuhairi

Malaysia Japan International Institute of Technology, University Technology Malaysia

Fitri Yakub

Malaysia Japan International Institute of Technology, University Technology Malaysia

Mas Omar

Malaysia Japan International Institute of Technology, University Technology Malaysia

Muhammad Sharifuddin

Malaysia Japan International Institute of Technology, University Technology Malaysia

Amrul Faruq

Faculty of Engineering, Universitas Muhammadiyah Malang

他

---

<https://doi.org/10.5109/7323359>

出版情報 : Proceedings of International Exchange and Innovation Conference on Engineering & Sciences (IEICES). 10, pp.845-850, 2024-10-17. International Exchange and Innovation Conference on Engineering & Sciences

バージョン :

権利関係 : Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International

## Imbalanced Flood Forecast Dataset Resampling Using SMOTE-Tomek Link

Ainaa Hanis Zuhairi<sup>1</sup>, Fitri Yakub<sup>1</sup>, Mas Omar<sup>1</sup>, Muhammad Sharifuddin<sup>1</sup>, Khamarrul Azahari Razak<sup>1</sup>, Amrul Faruq<sup>2</sup>  
Malaysia Japan International Institute of Technology, University Technology Malaysia, <sup>2</sup>Malang University  
Indonesia

Corresponding author email: [ainaahanis@graduate.utm.my](mailto:ainaahanis@graduate.utm.my)

**Abstract:** *Imbalanced data is common and presents significant challenge towards classification of data. In this research, we present a combination of two techniques used for handling class imbalance in datasets, SMOTE (Synthetic Minority Over-sampling Technique) and Tomek Links. Each strategy handles the class imbalance problem in a unique way, and their combination attempts to create a more balanced and cleaner dataset for training machine learning models to handle binary classification by addressing problematic or difficult-to-classify data. Machine learning classifiers used in this study are K-Nearest Neighbour (KNN), Support Vector Machine (SVM), Logistic Regression, Decision Tree (DT), Random Forest (RF), Gradient Boosting, Extreme Gradient Boosting (XGBoost), Light Gradient Boosting (LGBM), AdaBoost and Catboost. It has been discovered that the mean F1 score for resampled datasets provides more trustworthy results for forecasting floods.*

**Keywords:** Imbalanced Dataset, flood forecast, Resampling, SMOTE-Tomek.

### 1. INTRODUCTION

Floods are one of the world's deadliest natural disasters and they are almost certain to occur on a recurrent basis if preventive measures are not taken. When water overflows across generally dry terrain, it causes a flood, which may result casualties or fatalities as well as cause significant damage to infrastructures and personal assets in the disaster region [1]. Machine learning (ML) techniques are commonly used for forecasting [2]. Imbalanced data has several applications in real life, this includes handling high-speed rail fault diagnostics [3], fraud, Information Security and Data Mining. The difficulty arises when the number of instances of one class exceeds that of another creating an imbalance. Using an effective technique is crucial for tackling binary classification imbalance issue.

One typical method for addressing this imbalance dataset problem is to either oversample the minority class or undersample the dominant class. These techniques, however, have their own flaws. The vanilla oversampling approach duplicates some random instances from the minority class therefore this strategy adds no additional information to the data. On the contrary, the undersampling approach is used to eliminate certain random samples from the majority class while also eliminating some information from the original data [4]. In this study a technique is used to handle this difficulty, a technique has been proposed that is Synthetic Minority Oversampling Technique – Tomek links (SMOTE-Tomek Links).

### 2. LITERATURE REVIEW

A prevalent issue in classification tasks is data imbalance, which poses a barrier to the majority of traditional machine learning algorithms in terms of precisely predicting the target class [5]. In classification situations where the distribution of instances in the classes is skewed one way over the other, imbalanced data presents a challenge. As per Fig. 1, the majority class has far more samples, whereas a minority class has far less. The majority-minority class ratio might be 100:1 to 1000:1 or

more, indicating that majority class instances outnumber minority class instances [6].

Categorization should be able to be done by classifiers of Machine Learning (ML) without bias, unfortunately this may not be the case for imbalanced data. For desirable outcome, it is preferable to feed the Machine Learning algorithm with balanced data. It is also critical to investigate the various performance assessment measures for binary classification issues, to avoid being misled by better classification accuracy.

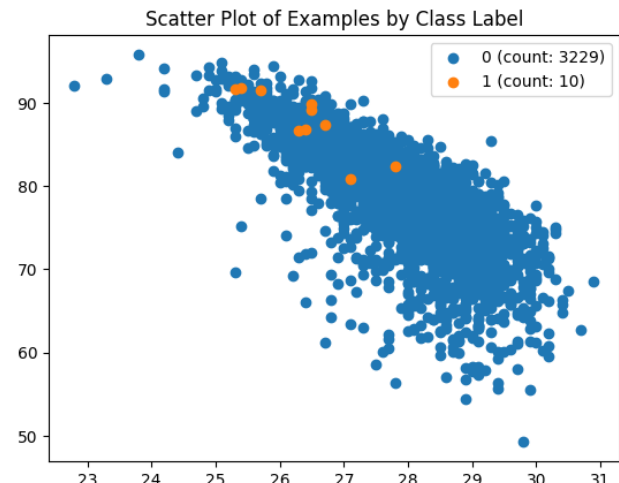


Fig. 1. Scatter Plot of Imbalanced flood data in Binary

#### 2.1 Data Level Approach

Resampling strategies can be broadly classified into three categories, hybrid methods, oversampling, and undersampling. Preprocessing techniques are used in the data-level approach to balance the unbalanced datasets on training data. Data-centric approaches also refer to the methods utilized in preprocessing stages to balance the unbalanced data in order to provide balanced training data and reduce the imbalance ratio between classes, these preprocessing techniques operate directly on the complete dataset [7].

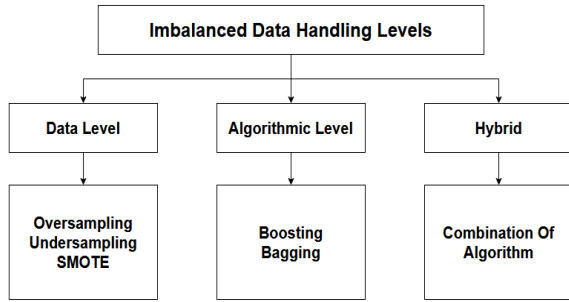


Fig. 2. Imbalanced Data Handling Level

## 2.2 Resampling

Issue with class Imbalance occurs when the number of samples for one class is much greater than other classes. The three characteristic of class imbalance are small disjunct, overlapping, and small sample size [8]. Resampling can help address the problems caused by unbalanced data. The process of adjusting the number of occurrences in the majority and minority classes to create balanced data is called resampling. Various resampling approaches, such as under sampling, oversampling, and hybrid methods, have been presented and are still in use today [9][10][11]. Using resampling approaches to balance the percentage of majority and minority samples in the training data is one of the most crucial strategies. In general, the data level approach uses two resampling techniques: under sampling and oversampling. The steps for resampling datasets are shown in Fig. 3.

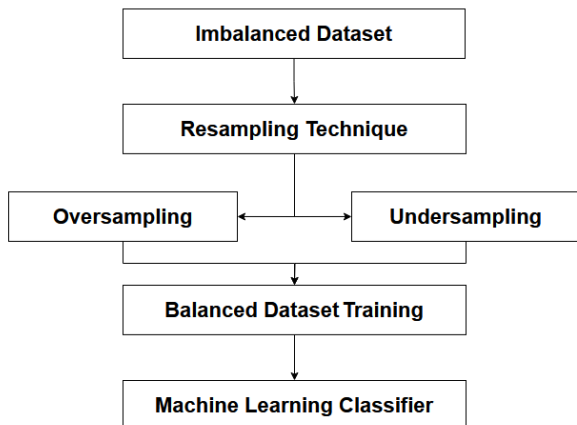


Fig. 3. Resampling technique flow for Imbalanced Data

## 2.3 Oversampling

Samples from the minority class are oversampled to balance them out with samples from the majority class. One of the most popular methods is simple random oversampling, which creates random samples from the majority class to correspond with the minority class. The primary concern regarding the oversampling strategy is that it does not supplement the dataset with fresh instances or information, potentially leading to overfitting of the classifiers [12].

## 2.4 Feature Selection

In data mining, feature selection has always been crucial. Typically, it entails a combination of search, calculation of the utility of the qualities, and assessment in relation to a certain learning scheme [13]. Because of two factors, feature selection is crucial for classification in high-dimensional datasets: first, some classification rules cannot be obtained if there are more features than samples; second, removing features with little variability

across samples can enhance classification performance [14]. Furthermore, by expediting the learning process and enhancing the model's capacity for generalization, feature selection can enhance the effectiveness of the classification algorithm even more [15]. Numerous feature selection techniques exist, they are broadly classified into three groups: filter techniques, wrapper techniques, and embedding techniques.

## 3. METHODOLOGY

This section covers the flood dataset description, the preparation methods for the data, and the learning techniques that were employed in the experiments.

### 3.1 Dataset Description

The information utilized in this study was gathered over a 10-year period by the Department of Irrigation and Drainage Malaysia and the Malaysian Metrological Department. The basic dataset consists of 3239 rows and 7 characteristics. This dataset is majorly skewed since the number of floodings classified as 1 in binary is 10, whereas 0 corresponds to non-flooding is 3229 (see Fig.1). This means that the minority class make up only 0.3% while the majority class make up to 99.7%. As a result, this dataset must be balanced before being fed into the Machine Learning classifier, as a balanced dataset is required for a classifier to make an accurate prediction.

### 3.2 SMOTE-Tomek link

Data preprocessing is an important stage in the machine learning pipeline since raw data may contain missing values and irrelevant variables. In this study, the data are encoded, since the optimal model is produced by the machine learning classifier using numerical values. The next stage is to use MinMaxScaler (MM) to scale all the numerical values. The MM approach is used to translate features into a specified range often 0, 1. Equation 1 represents how the values of a feature are scaled, where a, b is the range in which the data must be scaled [16].

$$x(\text{scaled}) = \frac{x - a}{b - a} \quad (1)$$

The dataset is divided into test and training sets using an 80-20 percent distribution afterwards. Synthetic Minority Oversampling Technique (SMOTE) intends to modify the classifier learning bias toward the minority class by producing an arbitrary number of artificial minority class data through interpolation. The fundamental concept is to locate K-nearest neighbors, defined as the K elements belonging to the minority class for each minority class sample  $x_i$ , and then randomly choose one of these neighbors. Using interpolation theory, we may produce a new sample  $x_{new}$  as per equation 2 [17].

$$x_{new} = x_i + (\hat{x}_i - x_i) \times \delta \quad (2)$$

Ivan Tomek introduced the idea of Tomek connections, a method for cleaning data. Tomek Links finds pairs of instances from the closest opposing classes [18]. A pair of neighbours separated by a minimal Euclidian distance is called a Tomek connection.  $(x_i, x_j)$  with  $x_i$  being a member of the minority group and  $x_j$  the majority class,  $d(x_i, x_j)$  signify the distance in Euclidian. In the event that no sample  $x_k$  exists fulfils the subsequent requirement:  $d(x_i, x_k) < d(x_i, x_j)$  or  $d(x_j, x_k) < d(x_i, x_j)$ , the pair of  $(x_i, x_j)$  is a Tomek link [17].

To address the issue of imbalanced data, in this article combination preprocessing strategy is used, SMOTE with Tomek connections. Tomek linkages were successfully used as a data cleaning strategy to eliminate samples produced by the SMOTE method that were close to the classification border. It is simple to determine the border between various classes by combining the Tomek connections technique.

### 3.3 Machine Learning (ML) Classifiers

One of the most often used methods for analysis is machine learning (ML), which enables computational models made up of several processing layers to learn representations of data with various levels of abstraction [19]. Large amounts of data may be processed fast and effectively by machine learning algorithms, allowing for the investigation of complex correlations and patterns [20]. This study uses multiple ML Classifiers, these are K-Nearest Neighbor (KNN), Support Vector Classifier (SVC), Logistic Regression, Decision Tree (DT), Random Forest (RF), Gradient Boosting (GB), Extreme Gradient Boosting (XGB), Light Gradient Boosting (LGB), AdaBoost and Catboost.

A non-parametric instance-based classifier is the KNN classifier [21]. It is a lazy learning technique that just saves all of the samples in the training set rather than learning from them [22]. The training step requires these saved values. The nearest neighbour estimate serves as the foundation for this approach. The distance metric, which is a similarity measure, is used to categorize the new cases. The most widely used measure is Euclidean distance. Finding the closest neighbour in a big training set takes a lot of time, which is a drawback of the KNN classifier.

SVM is a conventional machine learning model that is used in regression and classification. It is a component of the Supervised ML methodology. To classify the new data points, the SVM classifier separates the data by fitting an ideal line, or decision boundary, in the n-dimensional space. This decision boundary is known as the hyperplane [23][24]. Logistic Regression is a traditional machine learning classifier from the Supervised ML technique that is applied to classification problems. Using several independent attributes to predict the target attribute is the basic objective of logistic regression; the result can be either discrete or categorical [23][25].

The fundamental tenet of the decision tree, a traditional machine learning technique, is that comparable inputs lead to similar results. By evaluating the choices made for the various sample attributes and placing the samples in the next leaf node, decision-making using tree results aims to classify or regress the samples with the same attributes. The process of classifying data using a set of rules is called a decision tree. It offers a methodical methodology to determine which ideals will be attained in what circumstances. Decision trees come in two varieties: regression trees for continuous variables and classification trees for discrete variables [26].

Another technique for supervised classification is the Random Forest algorithm. It selects a random value and offers multiple options. It yields exact outcomes. The Random Forest system makes use of the Gini and

Entropy properties. This feature is utilized in a decision tree to select the optimal branch [27]. Gradient Boosting is a member of the Gradient Boosting Decision Trees class, which creates a set of dependent predictors by integrating several trees [28]. A scalable end-to-end tree boosting technique known as Extreme Gradient Boost is frequently used in data mining competitions [29]. LightGBM is a Gradient Boosting Decision Tree-based algorithm that divides the tree into leaves [30]. Unlike the traditional Gradient Boosting Decision Tree, the fraction of data instances for each feature is lowered dramatically during information gain estimate.

The Adaboost method is an iterative process that creates a powerful Bayesian classifier by combining many weak classifiers. Using the unweighted training sample data, the Adaboost builds a weak classifier to provide class labels. The weight assigned to a training data item that has been wrongly classified is called training [31]. Gradient boosting on decision trees is used in CatBoost, an efficient classification method that manages categorical features in data [32]. It uses statistical techniques to handle categorical data automatically, while other systems require the categorical data to be pre-sorted.

### 3.4 Evaluation Metrics

Evaluation metrics play a critical role in both evaluating the classification performance and refining the classifier modeling. As per the sources [3][33], the classification performance of unbalanced data may be evaluated using many assessment measures. Merely depending on accuracy in cases of extreme dataset skewness is misleading, as the model can attain high accuracy by learning from the majority class alone. Investigating the many performance evaluation measures is crucial to categorize the unbalanced data. The many performance assessment measures for classification problems are discussed in the context of unbalanced classification in this section.

The Confusion matrix [34], sometimes referred to as the Error matrix, is a helpful and straightforward statistic to utilize when working with classification issues. In a confusion matrix, which is a matrix table with rows and columns, there are four quadrants that each reflect the results of a single data point. For most binary classification tasks, metrics used in performance evaluation are computed using two or more quadrants of the confusion matrix.

The following is a description of each confusion matrix quadrant: True Positive (TP) quadrant represents the samples that are correctly predicted to be positive, meaning they are true. The samples that are supposed to be negative and are in fact true (predicted to decline and it is true) are shown in the True Negative (TN) quadrant. The False Positive (FP) quadrant, sometimes referred to as Type-1 error, shows samples that are projected to be positive but are false (i.e., predicted to continue but are false). The False Negative (FN) quadrant, sometimes referred to as the Type-2 Error, shows samples that are anticipated to be negative but are false (i.e., they are supposed to decrease).

The degree to which a classifier can correctly anticipate the classes is measured by its accuracy. It is also the most

popular assessment metric for classification tasks, however since a machine learning classifier is more likely to learn from the majority class, it should be avoided, especially when dealing with imbalanced datasets. This formula can be used to determine the accuracy [35]:

$$A = \frac{TP + TN}{TP + TN + FP + FN} \quad (3)$$

The precision measure shows the number of relevant samples that are included in the projected samples, so revealing the mistakes made in labelling a sample as correct when it is not. The ratio of TP to the total number of samples in the positive class is shown. The following is how it is expressed mathematically [36]:

$$P = \frac{TP}{TP + FP} \quad (4)$$

Recall quantifies how well the positive class was anticipated. The ratio of TP samples to all genuinely positive samples is known as the TP Rate, or sensitivity. The following is how it is expressed mathematically [36]:

$$R = \frac{TP}{TP + FN} \quad (5)$$

Essentially, the F1-Score is the harmonic mean of recall and precision. When recall or precision are both comparatively low, the F1-Score score is low. It is shown as follows [35]:

$$F1\ Score = \frac{2 \times P \times R}{P + R} \quad (6)$$

#### 4. RESULT AND DISCUSSION

Results are acquired against the performance evaluation metrics, namely Accuracy, Precision, Recall, F1-Score, listed in equations (3), (4), (5) and (6). Results of Before and after applying the SMOTE-Tomek link technique are displayed in Tables 1 and 2. As an initial comparison, we constructed a model without resampling, as indicated in Table 1 and Fig. 4, prior to resampling the data. Table 1's results demonstrate how terrible the model is and how often it predicts false positives as positives. Therefore, although having high accuracy and mean recall, decreased precision still results in a very low mean F1 score as the harmonic mean of recall and precision is the F1 score. A low F1 score means that precision is poor even with a strong recall as the percentage of genuine positives among all the model's positive predictions is known as precision.

Since the classifier is learning from the majority class, high accuracy occurs when the dataset's data distribution is disproportionate, this phenomenon is also known as the accuracy paradox. Resampling techniques are employed to solve this problem. Using nine classifiers, we evaluated the classification performance of the combined SMOTE-Tomek link algorithm to demonstrate the impact of the combined preprocessing technique, experiments are conducted.

The resampled data results are shown in Table.2 and Fig. 5. The resampling step utilising combined SMOTE and

Tomek connections approach assessment metrics had an improvement on the mean precision, mean recall, accuracy, and especially mean F1-score which can be clearly seen in Mean F1 score bar chart Fig.4 and Fig.5. This can be seen in all classifiers result after being resampled. The assessment measures that are clearly improved come from distinct classifiers. The method of merging SMOTE and Tomek linkages performs better across a range of classifiers, indicating that the combined method can be used to diverse environments including data and classifiers.

The findings demonstrated that evaluation measures utilizing both SMOTE and Tomek connections together are significantly better than evaluation metrics without any preprocessing, which is thought to be a good preprocessing technique in some recent literature [37].

Table 1. Imbalanced data result

Classifier	Mean Accuracy	Mean Precision	Mean Recall	Mean F1
KNN	0.9892	0.5840	0.8700	0.2500
SVM	0.9985	0.7807	0.8123	0.5833
LogReg	0.9985	0.7807	0.8123	0.5833
DT	0.9985	0.8434	0.8764	0.5000
RF	0.9985	0.7807	0.8123	0.4583
GB	0.9981	0.7703	0.8121	0.5625
XGB	0.9977	0.8019	0.9365	0.6875
LGB	0.9977	0.8019	0.9365	0.6875
AdaBoost	0.9965	0.7847	0.8316	0.5972
CatBoost	0.9977	0.8019	0.9365	0.6875

Table 2. Resampled data result

Classifier	Mean Accuracy	Mean Precision	Mean Recall	Mean F1
KNN	0.9977	0.6238	0.6250	0.2732
SVM	0.9985	0.8644	0.9369	0.7708
LogReg	0.9965	0.8021	0.9983	0.7083
DT	0.9981	0.7494	0.8121	0.5416
RF	0.9977	0.7704	0.8742	0.6041
GB	0.9981	0.7807	0.8121	0.5833
XGB	0.9981	0.8331	0.9367	0.7291
LGB	0.9975	0.8300	0.9315	0.7083
AdaBoost	0.9981	0.8017	0.8744	0.6458
CatBoost	0.9981	0.8331	0.9367	0.7291

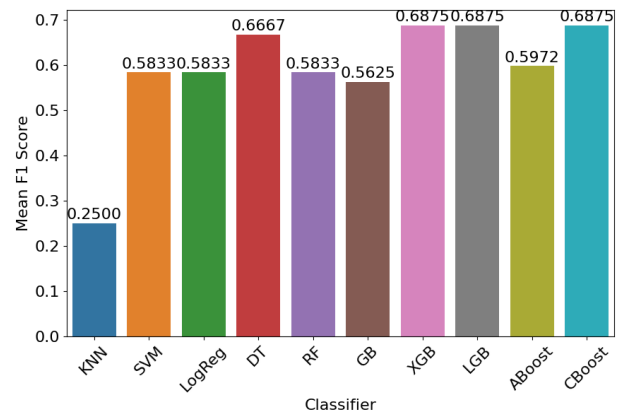


Fig. 4. Imbalanced Data mean F1 score Bar chart.

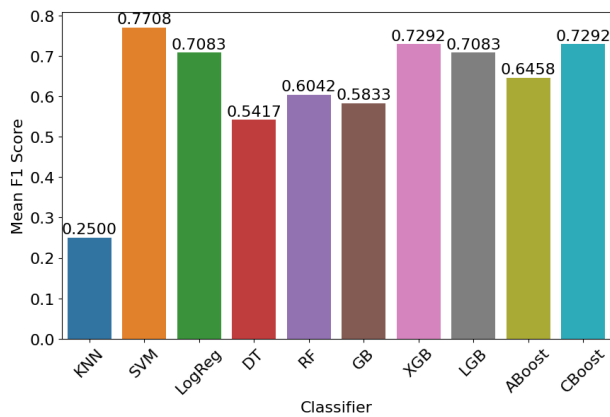


Fig. 5. Resampled Data mean F1 score Bar chart.

## 5. CONCLUSION

One of the problems with flood forecasting is high data imbalance, and it can be difficult to anticipate the dropout rate with machine learning systems. This paper discusses the challenges of applying machine learning algorithms to highly skewed data and the significance of researching performance evaluation metrics other than accuracy for binary classification. This is because a skewed dataset may make it difficult to determine the trained model's accuracy and the process of resampling aids in the creation of a balanced dataset, which improves classification performance.

## 6. REFERENCES

Reference to a journal publication:

- [1] T. Tingsanchali, Urban flood disaster management, *Procedia Eng.*, 32, (2012) 25–37.
- [2] Q. Di, Q. Jinbo, and C. Mingti, Application of Machine Learning in Flood Forecast: A Survey, *Proc. - 2022 Int. Conf. Virtual Reality, Human-Computer Interact. Artif. Intell. VRHCIAI.* (2022) 177–181.
- [3] X. Y. Liu, J. Wu, and Z. H. Zhou, Exploratory undersampling for class-imbalance learning, *IEEE Trans. Syst. Man, Cybern. Part B Cybern.* 39 (2009) 539–550.
- [4] S. W. Masood and S. A. Begum, Comparison of Resampling Techniques for Imbalanced Datasets in Student Dropout Prediction, *Proc. - 2022 IEEE Silchar Subsect. Conf. SILCON. 2022.*
- [5] P. Kumar, R. Bhatnagar, K. Gaur, and A. Bhatnagar, Classification of Imbalanced Data: Review of Methods and Applications, *IOP Conf. Ser. Mater. Sci. Eng.* 1099 (2021).
- [6] H. He and E. A. Garcia, Learning from imbalanced data, *IEEE Trans. Knowl. Data Eng.* 21 (2009) 1263–1284.
- [7] H. Kaur, H. S. Pannu, and A. K. Malhi, A systematic review on imbalanced data challenges in machine learning: Applications and solutions, *ACM Comput. Surv.* 52 (2019).
- [8] W. C. Lin, C. F. Tsai, Y. H. Hu, and J. S. Jhang, Clustering-based undersampling in class-imbalanced data, *Inf. Sci. (Ny)*. 409–410 (2017), 17–26.
- [9] M. Khushi *et al.*, A Comparative Performance Analysis of Data Resampling Methods on Imbalance Medical Data, *IEEE Access*, 9, (2021) 109960–109975.
- [10] J. Diz, G. Marreiros, and A. Freitas, Applying Data Mining Techniques to Improve Breast Cancer Diagnosis, *J. Med. Syst.*, 40 (2016).
- [11] S. Santiso, A. Casillas, and A. Pérez, The class imbalance problem detecting adverse drug reactions in electronic health records, *Health Informatics J.* 25 (2019) 1768–1778.
- [12] V. S. Spelman and R. Porkodi, A Review on Handling Imbalanced Data, *Proc. 2018 Int. Conf. Curr. Trends Toward Converging Technol. ICCTCT* (2018).
- [13] S. Maldonado, R. Weber, and F. Famili, Feature selection for high-dimensional class-imbalanced data sets using Support Vector Machines, *Inf. Sci. (Ny)*. 286 (2014) 228–246.
- [14] S. Dudoit, J. Fridlyand, and T. P. Speed, Comparison of discrimination methods for the classification of tumors using gene expression data, *J. Am. Stat. Assoc.* 97 (2002) 77–86.
- [15] F. Nie, H. Huang, X. Cai, and C. Ding, Efficient and robust feature selection via joint  $\ell_2/\ell_1$ -norms minimization, *Adv. Neural Inf. Process. Syst.* 23 24th Annu. Conf. Neural Inf. Process. Syst. 2010, NIPS. (2010) 1–9.
- [16] R. Sekar and C. Christopher Columbus, Enhancing Credit Card Application Approval through Data Scaling in Machine Learning Algorithms, *Int. Conf. Sustain. Commun. Networks Appl. ICSCNA 2023 - Proc.* (2023) 1388–1394.
- [17] M. Zeng, B. Zou, F. Wei, X. Liu, and L. Wang, Effective prediction of three common diseases by combining SMOTE with Tomek links technique for imbalanced medical data, *Proc. 2016 IEEE Int. Conf. Online Anal. Comput. Sci. ICOACS 2016.* (2016) 225–228.
- [18] I. Tomek, TWO MODIFICATIONS OF CNN, *IEEE Trans. Syst. Man Cybern.* vol. SMC-6 (1976) 769–772.
- [19] J. H. Lo, L. K. Lin, H. P. Cheng, and C. C. Hung, Deep Learning-based Image Recognition for Disaster Prevention Application, *Int. Exch. Innov. Conf. Eng. Sci.* (2022) 400–406.
- [20] J. Galupino and J. Dungca, Estimation of Permeability of Soil-Fly Ash Mix using Machine Learning Algorithms, *Int. Exch. Innov. Conf. Eng. Sci.* 9, (2023) 28–33.
- [21] N. Shweta and H. Nagendra, EEG signal classification using wavelet and fuzzy KNN classifier, *AIP Conf. Proc.* 2316 (2021).
- [22] (PDF) Application of K-nearest neighbor (KNN) approach for predicting economic events theoretical background. website: [https://www.researchgate.net/publication/304826093\\_Application\\_of\\_K-nearest\\_neighbor\\_KNN\\_approach\\_for\\_predicting\\_economic\\_events\\_theoretical\\_background](https://www.researchgate.net/publication/304826093_Application_of_K-nearest_neighbor_KNN_approach_for_predicting_economic_events_theoretical_background) (accessed 14.07. 2024).
- [23] S. Ray, A Quick Review of Machine Learning Algorithms, *Proc. Int. Conf. Mach. Learn. Big Data, Cloud Parallel Comput. Trends, Perspectives Prospect. Com.* 2019. (2019) 35–39.
- [24] G. Bansal and M. Singla, Ensembling of non-linear svm models with partial least square for diabetes prediction, *Lect. Notes Electr. Eng.* 569 (2020) 731–739.
- [25] B. X. Maochun, Idea of Knowledge-Based Engineering using CAD Model, *J. Mach. Comput.* 1 (2021) 198–205.

- [26] K. He and C. He, Housing Price Analysis Using Linear Regression and Logistic Regression: A Comprehensive Explanation Using Melbourne Real Estate Data, 2021 IEEE Int. Conf. Comput. ICOCO (2021) 241–246.
- [27] P. P. Singh, S. Prasad, B. Das, U. Poddar, and D. R. Choudhury, Classification of diabetic patient data using machine learning techniques, *Adv. Intell. Syst. Comput.*, 696 (2018) 427–436.
- [28] J. H. Friedman, Greedy function approximation: A gradient boosting machine, *Ann. Stat.* 29, (2001) 1189–1232.
- [29] T. Chen and C. Guestrin, XGBoost: A Scalable Tree Boosting System, *Proc. ACM SIGKDD Int. Conf. Knowl. Discov. Data Min.* 13 (2016) 785–794.
- [30] Q. Meng *et al.*, A communication-efficient parallel algorithm for decision tree, *Adv. Neural Inf. Process. Syst.* (2016) 1279–1287.
- [31] Y. Shen, Y. Jiang, W. Liu, and Y. Liu, Multi-class AdaBoost ELM. 2 (2015) 179–188.
- [32] X. Fei, Y. Fang, and Q. Ling, Discrimination of Excessive Exhaust Emissions of Vehicles based on Catboost Algorithm, *Proc. 32nd Chinese Control Decis. Conf. CCDC.* (2020) 4396–4401.
- [33] H. He, Y. Bai, E. A. Garcia, and S. Li, ADASYN: Adaptive synthetic sampling approach for imbalanced learning, *Proc. Int. Jt. Conf. Neural Networks.* (2008) 1322–1328.
- [34] A. Luque, A. Carrasco, A. Martín, and A. de las Heras, The impact of class imbalance in classification performance metrics based on the binary confusion matrix, *Pattern Recognit.* 91 (2019) 216–231.
- [35] C. Chen, X. Xu, X. Ma, L. Yang, G. Wang, and X. Ye, Comprehensive performance evaluation of intrusion detection model based on radar chart method, *Proc. - 2022 2nd Int. Conf. Electron. Inf. Eng. Comput. Technol. EIECT.* (2022) 44–47.
- [36] P. Sujatha and K. Mahalakshmi, Performance Evaluation of Supervised Machine Learning Algorithms in Prediction of Heart Disease, *IEEE Int. Conf. Innov. Technol. INOCON.* (2020).
- [37] E. M. Karabulut and T. Ibrikci, Effective automated prediction of vertebral column pathologies based on logistic model tree with SMOTE preprocessing, *J. Med. Syst.* 38 (2014).