International Journal of Mathematics and Computer Research

ISSN: 2320-7167

Volume 13 Issue 11 November 2025, Page no. – 5921-5927

Index Copernicus ICV: 57.55, Impact Factor: 8.615

DOI: 10.47191/ijmcr/v13i11.12

Handling Class Data Imbalance in Random Forest using the SMOTE and ADASYN Methods in Identify Economic Status

Fane Oktari[1], Alan Prahutama[2*], Tarno[3], Sudarno[4]

[1,2,3,4]Department of Statistics, Faculty of Science and Mathematics, Diponegoro University (UNDIP), Semarang, Indonesia

ARTICLE INFO

Published Online:
27 November 2025

ABSTRACT

Imbalanced class data refers to an imbalance in the amount of training data between two different classes, where one class represents a large amount of data (majority class) while the other class represents a very small amount of data (minority class). Oversampling is a technique of balancing data by generating data in the minority class so that the amount is balanced with the data in the majority class with the aim of improving the classification results for the better. The oversampling method was chosen to avoid losing important information in the imbalanced dataset. Synthetic Minority Over-Sampling Technique (SMOTE) and Adaptive Synthetic Approach (ADASYN) are oversampling techniques that produce synthetic data and use the concept of adjacency in the algorithm. SMOTE and ADASYN can reduce the possibility of overfitting, which is a disadvantage of ordinary oversampling. These two methods will be combined with the Random Forest classification algorithm. This research was conducted to solve the problem of imbalance class data in the dataset of the Indonesian economic financial crisis. The data in this study uses 9 independent variables based on macroeconomic aspects and 1 dependent variable. The dependent variable used is categorized as binary, namely crisis and non-crisis conditions. The results of this study indicate that handling class imbalance data in the random forest (RF) classification algorithm results in better classification performance. ADASYN-RF produces the best performance with Accuracy, recall, precision, F1 Score, and ROC AUC scores of 98.26%, 66.67%, 72.22%, 65.57%, and 82.93%, respectively.

Corresponding Author:
Alan Prahutama

KEYWORDS: Class Imbalance, SMOTE, ADASYN, Random Forest.

I. INTRODUCTION

A crisis is a major challenge for any country because it generates widespread disruptions that penetrate virtually all sectors of national life, particularly economic growth and financial stability. Economic crises often arise when internal vulnerabilities interact with external shocks, leading to a rapid deterioration of macroeconomic performance, declining investor confidence, and weakening fiscal capacity. Indonesia has faced several significant economic and financial crises that demonstrate how deeply such events can affect national development. Historically, Indonesia experienced two large-scale crises triggered by global economic downturns: the Asian Financial Crisis in 1997–1998 and the Global Financial Crisis in 2008. The 1998 crisis, characterized by currency collapse, hyper-inflation, and massive unemployment, is often cited as one of the most severe episodes in Southeast. A decade later, Indonesia again felt the repercussions of the 2008 international financial turmoil, although the impact was mitigated by stronger financial sector reforms and improved macroeconomic fundamentals.

In 2020, Indonesia once more confronted a major economic shock, this time caused by the spread of COVID-19, which disrupted global supply chains, halted mobility, and triggered a contraction in national economic output. The pandemic-induced crisis highlighted the country's vulnerability to sudden declines in consumption, investment, and trade, making it Indonesia's first recession in over two decades [1]. According to [2], no crisis occurs abruptly or without warning; rather, crises are preceded by identifiable patterns in key economic indicators. Early-warning signals such as excessive credit growth, exchange-rate misalignment, rising inflation, declining foreign reserves, and deteriorating fiscal balances often provide clues of impending

macroeconomic instability. These indicators play a crucial role in predicting vulnerabilities and guiding policymakers to formulate preemptive responses.

Therefore, determining economic conditions that have the potential to develop into a financial crisis in Indonesia requires a careful analysis of macroeconomic aspects. Monitoring variables such as GDP growth, inflation, exchange rates, capital flows, and government debt is essential to detect early signs of imbalance and to design timely policy interventions. By understanding these macroeconomic dynamics, Indonesia can strengthen its crisis-prevention strategies and enhance economic resilience in the face of global uncertainty.

Machine learning provides a flexible framework for solving classification problems in economics and finance, including the prediction of a country's financial condition. Among various classification algorithms, random forest is particularly suitable for modeling Indonesia's transition between crisis and non-crisis states. Random forest, introduced by [3], is an ensemble learning method that builds many decision trees on bootstrap samples and aggregates their outputs, which reduces variance and improves generalization performance. This algorithm offers several advantages over traditional single-model approaches such as logistic regression or a single decision tree [4]. First, random forest is relatively robust to noise and outliers because each tree is trained on a different subset of data and features, so individual extreme observations are less likely to dominate the final prediction. Second, it can handle high-dimensional macroeconomic data—such as exchange rates, interest rates, inflation, money supply, stock indices, and capital flows—without requiring extensive feature elimination, while still providing variable-importance measures to identify the most influential indicators [5]. In the context of financial stability analysis, these properties make random forest attractive for early-warning systems that classify Indonesia's economy into "crisis" or "non-crisis" conditions based on macroeconomic signals. Recent literature on AI and machine learning in financial services also highlights the growing role of ensemble methods, including random forests, in risk assessment and systemic risk monitoring [6]. Thus, using random forest to classify Indonesia's financial condition is justified both theoretically and empirically, as it combines robustness, relatively low computational cost, and high predictive accuracy.

Imbalanced data in classification problems occurs when one class (the majority class) contains far more observations than the other (the minority class). In this study, the response variable is imbalanced, so it cannot be processed directly using standard classification algorithms because they tend to be biased toward the majority class. As a result, the model may appear to have high overall accuracy but performs poorly in detecting the minority class, which is often the class of greatest interest, such as "crisis" versus "non-crisis" conditions or "default" versus "non-default" cases [7]. To address this problem, resampling techniques such as Synthetic Minority Over-sampling Technique (SMOTE) and Adaptive Synthetic Sampling (ADASYN) are widely used.

SMOTE generates synthetic samples by interpolating between existing minority class instances and their nearest neighbors in feature space, thereby increasing the representation of the minority class without simply duplicating observations [8]. This helps the classifier learn a more general decision boundary. ADASYN extends the idea of SMOTE by adaptively focusing on minority samples that are harder to learn—those located near the decision boundary or surrounded by majority samples—so that more synthetic data are generated in these difficult regions (He et al., 2008). By using SMOTE and ADASYN, the class distribution becomes more balanced, improving the classifier's ability to correctly identify both majority and minority classes and enhancing evaluation metrics such as recall, F1-score, and AUC for the minority class.

II. LITERATURE REVIEW

A. Early Warning System with the Exchange Market Pressure (EMP) Approach

An early warning system is a model for predicting the likelihood and timing of a crisis. The Exchange Market Pressure (EMP) index is used to measure international economic pressure on a country's finances. The EMP method from [9] is widely used because it is easy to calculate and can describe the extent of pressure on a country's foreign exchange. Exchange Market Pressure (EMP) is defined as:

$$EMP = \delta e_t + \left(\frac{\sigma\delta e}{\sigma\delta R}\right)\delta R_t \quad (1)$$

δe_t is the weighted average of exchange rate changes, δR_t is the rate of change in foreign exchange reserves, $\sigma\delta e$ is the standard deviation of the rate of change in exchange rates, and $\sigma\delta R$ s the standard deviation of the rate of change in foreign exchange reserves. Formally, a currency crisis is said to occur if $EMP > \mu EMP + mEMP$. Where μEMP is the average EMP index and $mEMP$ shows the standard deviation of the EMP index. The value of m used is 1.5, following the model used by the World Bank.

B. Random Forest

Random Forest is one of the most widely used machine learning methods for both classification and regression tasks due to its strong predictive performance and robustness. The algorithm was introduced by Breiman (2001) as an extension of the Classification and Regression Tree (CART) [10] framework by integrating two key techniques: bootstrap aggregating (bagging) and random feature selection. In the Random Forest process, multiple decision trees are constructed using different bootstrap samples drawn from the training dataset. At each node of every tree, only a random subset of predictor variables is considered for

splitting, which reduces correlation among trees and improves model generalization [3].

During tree construction, the algorithm typically employs the Gini index to determine the best split variable. The split chosen is the one that results in the greatest reduction in impurity, enabling the model to effectively separate classes from a randomly selected feature subset. In the classification stage, Random Forest aggregates the output of all trees using a majority voting mechanism, where the class chosen by most trees becomes the final prediction. Two important parameters in Random Forest include mtry, the number of randomly selected features evaluated at each split, and ntree, the number of trees grown in the forest.

Random Forest offers several advantages over single decision trees and many other machine learning methods. It is highly robust to noise and outliers, less prone to overfitting due to ensemble averaging, and capable of handling high-dimensional data with minimal preprocessing. Additionally, it provides useful estimates of feature importance, which helps identify the most influential variables in a dataset [11]. Its computational efficiency and strong predictive accuracy make Random Forest particularly suitable for financial risk modeling, crisis prediction, and other complex economic applications.

C. Grid Search

Grid Search is one of the most widely used algorithms for decision-making in model tuning, particularly in the context of hyperparameter optimization. Hyperparameters play a critical role in determining a model's performance, and selecting the optimal combination is essential for achieving the best predictive results. Grid Search provides a systematic approach for this task by exhaustively evaluating all possible combinations of hyperparameter values specified by the researcher. According to [12], Grid Search is frequently applied because its implementation is conceptually simple, easy to control, and suitable for models with a relatively small and well-defined hyperparameter space. In the Grid Search process, the researcher first defines a set of candidate values for each hyperparameter. The algorithm then constructs a parameter grid—essentially a Cartesian product of all defined values—and evaluates model performance for every combination. Each model is trained and validated, usually using cross-validation, to ensure reliable performance estimation. The best parameter configuration is the one that yields the highest evaluation score (such as accuracy, F1-score, or AUC) across all tested combinations. Although Grid Search can be computationally intensive when dealing with many parameters or large datasets, its exhaustive nature ensures that no combination within the predefined grid is overlooked [13].

D. SMOTE (Synthetic Minority Over Sampling Technique)

The SMOTE technique works by performing over-sampling on the minority class by creating synthetic samples. SMOTE generates data from the minority class using the [14]. Using Euclidean distance calculations, suppose there is data with p variables, then the distance between $\mathbf{x}' = \begin{bmatrix} x_1 & x_2 & \cdots & x_p \end{bmatrix}$ and $\mathbf{z}' = \begin{bmatrix} z_1 & z_2 & \cdots & z_p \end{bmatrix}$ is

$$d(x,z) = \sqrt{(x_1 - z_1)^2 + \cdots + (x_p - z_p)^2} \quad (2)$$

Using SMOTE, data will be generated using the following equation [15]:

$$\mathbf{x}_{syn} = \mathbf{x}_i + (\mathbf{x}_{kNN} - \mathbf{x}_i)\gamma \quad (3)$$

$\mathbf{x}_{syn}$ is synthetic data generated by SMOTE, $\mathbf{x}_i$ is the i-th data point from the minority class, $\mathbf{x}_{kNN}$ is the data point from the minority class that is closest to $\mathbf{x}_i$, and γ is a random number between 0 and 1. The procedure for generating artificial data for numerical and categorical data is as follows [16]:

1. Numeric Data
   a. Calculate the difference between the main vector and its k nearest neighbors.
   b. Multiply the differences by a random number between 0 and 1.
   c. Add the difference to the main value in the original main vector to obtain a new main vector.
2. Categorical Data
   a. Select the majority among the main vectors considered with their k nearest neighbors for nominal values. If there are identical values, then select randomly.
   b. Make that value the sample data for the new artificial class.

E. ADASYN (Adaptive Synthetic Technique)

The ADASYN approach to imbalanced data was first introduced by He et al in 2008 [7]. The main idea of ADASYN is to use minority class distribution weights based on the difficulty level of data learning by the model, where synthetic data is generated from minority classes that are difficult to learn compared to minority data that is easier to learn [17]. ADASYN has parameters to determine the expected balance level (β) and a threshold set as the maximum tolerance degree of class imbalance ratio (d_{th}). The ADASYN method uses density distribution $(\hat{r}_l)$ as a criterion to automatically decide how many synthetic samples are needed for each minority data.

F. Classification Performance Measure

The performance of the classification model is evaluated based on testing on objects that are predicted correctly or incorrectly. The confusion matrix is a place to tabulate the calculation results. The confusion matrix is a tool used to analyze how well the classifier recognizes tuples from different classes [18]. The following is a confusion matrix Tabel 1:

Table 1. ***Confusion Matrix***

Class	*Predicted Positive*	*Predicted Negative*
Actual Positive	*TP (True Positive)*	*FN (False egative)*
Actual Negative	*FP (False Positive)*	*TN (True Negative)*

There are various measures for evaluating model performance based on the confusion matrix, including accuracy, which is the overall accuracy of predictions; precision, which is a measure of the proportion of predictions from TP; and recall, which is a measure of the usefulness of a model that can be used to see the accuracy in a single class, especially in the case of unbalanced datasets. Accuracy, specificity, precision, recall, and F1 score are obtained as follows:

$$Accuracy = \frac{TP+TN}{TP+FP+TN+FN} \quad (4)$$

$$Specificity = \frac{TN}{FP+TN} \quad (5)$$

$$Precision = \frac{TP}{FP+TP} \quad (6)$$

$$Recall = \frac{TP}{TP+FN} \quad (7)$$

$$F1\_Score = \frac{2xPrecisionxRecall}{Precision+Recall} \quad (8)$$

G. Holdout and K-Fold Validation

In machine learning research, it is common not to use all of the data in the training process. To avoid overfitting, the model must be constructed in such a way that when new data is encountered, it can be predicted as well as the data used in the training process. The division of training and testing data can be done using the holdout validation method, which involves dividing the data into two parts (training and testing) with proportions determined by the researcher. The proportions used by researchers are 60%:40%, 70%:30%, or 80%:20% (Raschka, 2018) or can be determined by researchers with other numbers, provided that the proportion of training data is greater than that of testing data. In addition to holdout validation, data can also be divided using the K-Fold Cross Validation method. The performance of this method involves dividing the data into K groups of training and testing data, so that the training process will be repeated K times, and the model's performance is the average of all training processes.

III. RESEARCH METHOD

The data used is secondary data obtained from the Bank Indonesia (BI) website, International Financial Statistics (IFS), and the Central Statistics Agency (BPS). A total of 311 data points were collected, consisting of 9 independent variables and 1 dependent variable. The independent variables used were exports (X_1), foreign exchange reserves (X_2), the Jakarta Composite Index (X_3), the difference between lending and deposit interest rates (X_4), real deposit interest rates (X_5), real exchange rates (X_6), M1 (X_7), M2/Foreign Exchange Reserves (X_8), and M2 (X_9). The dependent variable has a binary categorical value, where 0 means no crisis and 1 means crisis.

The data in this study was processed using Python software with Google Colaboratory (Colab), which is a Python development environment that runs on Google Cloud and Microsoft Excel 2010. The data obtained was then analyzed using the following steps:

1. Data input
2. Perform pre-processing. Calculate Exchange Market Pressure (EMP) for the dependent variable with a value of 0 categorized for non-crisis periods and 1 categorized for crisis periods.
3. Performing descriptive analysis
4. Performing classification using random forest, either using the SMOTE and ADASYN data imbalance techniques or not.

4.1. Classification using the random forest technique with the following steps:
   a. Performing training and testing data division with holdout validation.
   b. Tune the random forest hyperparameters.
   c. Forming a random forest classification tree model.
   d. Performing predictions for the training data and testing data using the constructed random forest classification model.
   e. Measure classification performance.

4.2. Classification using the random forest technique with data imbalance handling using the SMOTE method.
   a. Performing the division of training data and testing data with holdout validation.
   b. Performing SMOTE random forest hyperparameter tuning.
   c. Forming the SMOTE random forest classification tree model.
   d. Measuring classification performance.

4.3. Classification using the random forest technique with data imbalance handling using the ADASYN method.
   a. Performing training data and testing data division with holdout validation.
   b. Performing hyperparameter tuning for SMOTE random forest.
   c. Building a SMOTE random forest classification tree model.
   d. Performing classification performance measurement.

IV. RESULTS AND DISCUSSION

A. Pre-Processing Data

Detecting Missing value of the dataset: Checking for missing values was performed on the initial dataset. Based on the table, there were no missing values in all independent and dependent variables, so there was no need to delete rows or estimate values for missing data.

Table 2. Number of *Missing Values for* Each Variable

Variable	Number of *Missing Values*
Exports	0
Foreign Exchange Reserves	0
IHSG	0
Interest rate differential between loans and deposits	0
Real deposit interest rate	0
Real exchange rate	0
M1	0
M2/Foreign exchange reserves	0
M2	0

Data Transformation: The dataset used consists of 9 variables that have numeric data types with different parameter units. Differences in parameter units will affect the data synthesis results because the SMOTE and ADASYN algorithms use Euclidean distance in their processes. Data transformation in this dataset uses min-max normalization by placing the data in the range of 0 to 1 so that the dataset has the same parameter units.

Early Warning System Calculation using the Exchange Market Pressure (EMP) Approach: Based on calculations using EMP, for a value of m=1.5, there were 9 months that experienced a crisis. The crisis periods experienced by Indonesia occurred in 2009 (March), 2008 (November), 2001 (June), 1999 (May & September), and 1998 (January, June, July, September).

B. Random Forest Modelling in Economic Status

The best parameters for the model were determined by simultaneous tuning with grid search. The table presents the parameters that were tested:

Table 3. Random Forest Parameters

Parameters	Parameter Value
Mtry	3, 4, 5, 6, 7, 8, 9
Ntree	50, 75, 100, 150, 200, 250, 500

This tuning process uses 5-fold cross validation and obtains the best parameter values, namely mtry = 5 and ntree = 75. The table presents the goodness measures in the random forest model.

Table 4. *Random Forest* Model Performance Measures (Training:Testing)

Performance	(70%:30%)	(75%:25%)	(80%:20%)
Accuracy	97.87%	97.43%	96.82%
Recall	0%	0%	0%
Precision	0%	0%	0%
ROC-AUC Score	50%	50%	50%
F1_score	0%	0%	0%

Table 4 shows accuracy values of over 90%, but recall, precision, and F1_score values are 0%, and roc_auc_score values are 50%. Low recall values indicate that the model is unable to accurately predict data at crisis levels. A low recall value is usually caused by an imbalance in the amount of data between classes, so an SMOTE random forest and ADASYN random forest model will be built.

C. SMOTE-Random Forest Modelling in Economic Status

The SMOTE algorithm increases the amount of data in the crisis class so that it is equal to the non-crisis class. The following table shows the crisis and non-crisis classes with a training data distribution of 70%, 75%, and 80%.

Table 5. Amount of Data for Each Class Before-After SMOTE

Training	Before	After
70	Crisis: 7 Non-Crisis: 210	Crisis: 210 Non-Crisis: 210
75	Crisis: 7 Non-Crisis: 226	Crisis: 226 Non-Crisis: 226
80%	Crisis: 7 Non-Crisis: 241	Crisis: 241 Non-Crisis: 241

After data balancing, hyperparameter tuning was performed using the same parameters as the previous random forest model. The parameters used for the SMOTE random forest were mtry = 5 and ntree = 75.

Table 6. Model Goodness Measures for SMOTE Random Forest (Training:Testing)

Performance	(70%:30%)	(75%:25%)	(80%:20%)
Accuracy	97.87%	97.43%	95.23%
Recall	100%	50%	50%
Precision	50%	50%	33.33%
ROC AUC Score	98.91%	74.34%	73.36%
F1_score	66.67%	50%	40%

Table 6 shows the model's performance with holdout validation. The accuracy value on the testing data generated is more than 95%, and the recall value has increased, with the best recall value being 100% on a 70%:30% holdout validation comparison. Precision also increased with the

best value of 50% in the 70%:30% and 75%:25% holdout validation comparisons. Roc_auc_score increased with the best value of 98.91% in the 70%:30% holdout validation comparison. The best F1 score was 66.67% in the 70%:30% holdout validation.

D. *ADASYN-Random Forest Modelling in Economic Status*

The ADASYN algorithm increases the amount of data in the crisis class so that it is equal to the non-crisis class. The following is a table of crisis and non-crisis classes with a training data distribution of 70%, 75%, and 80%.

Table 7. Amount of Data for Each Class Before-After ADASYN

Training	**Before**	**After**
70%	Crisis: 7 Non-Crisis: 210	Crisis: 156 Non-Crisis: 210
75%	Crisis: 7 Non-Crisis: 226	Crisis: 172 Non-Crisis: 226
80%	Crisis: 7 Non-Crisis: 241	Crisis: 179 Non-Crisis: 241

After data balancing, hyperparameter tuning was performed using the same parameters as the previous random forest model. The parameters used for the ADASYN random forest were mtry = 5 and ntree = 75.

Table 8. Model Goodness Measures for ADASYN Random Forest (Training:Testing)

Performance	**(70%:30%)**	**(75%:25%)**	**(80%:20%)**
Accuracy	98.94%	97.43%	98.41%
Recall	100%	50%	50%
Precision	66.67%	50%	100%
ROC AUC Score	99.46%	74.34%	75%
F1_score	80%	50%	66.67%

Table 8 shows the performance of the model with holdout validation. The accuracy value on the testing data generated is more than 95%, and the recall value has increased, with the best recall value being 100% on a 70%:30% holdout validation comparison. Precision also increased, with the best value being 100% in the 80%:20% holdout validation comparison. The ROC AUC score increased, with the best value being 99.46% in the 70%:30% holdout validation comparison. The best F1 score was 80% in the 70%:30% holdout validation.

V. CONCLUSION

Based on the results and discussion, it can be seen that using SMOTE and ADASYN successfully increased the proportion of minority class sample data, namely the crisis class, and made the proportion of the target variable balanced. Handling class imbalance with the SMOTE and ADASYN methods in classification using the Random Forest algorithm showed better results than not handling class imbalance. Based on the average testing data goodness measure, ADASYN Random Forest has a higher value than SMOTE Random Forest.

The models formed in this study are three, namely Random Forest (RF), SMOTE -RF, and ADASY-RF. The accuracy values of the three models are not much different, namely above 95%, but the recall, precision, ROC AUC score, and F1 score values between the models before and after balancing using SMOTE and ADASYN appear to be different.

Based on comparison using SMOTE RF and ADASYN RF, the recall value increased from 0% to 66.67%. The precision of SMOTE RF and ADASYN RF also increased to 44.44% and 72.22%. The F1 score of SMOTE RF and ADASYN RF increased to 52.22% and 65.67%. The ROC AUC score results of SMOTE RF and ADASYN RF increased to 82.20% and 82.93%. Overall, the ADASYN RF algorithm improved the classification results, as seen from the increase in recall, precision, F1 score, and ROC AUC score values.

REFERENCES

1. W. Bank, "Indonesia economic prospects: Boosting the recovery," *World Bank Publications*, 2021. https://www.worldbank.org/en/country/indonesia/publication/indonesia-economic-prospects
2. G. L. Kaminsky, S. Lizondo, and C. M. Reinhart, "Leading indicators of currency crises," *IMF Staff Pap.*, vol. 45, no. 1, pp. 1–48, 2000.
3. L. Breiman, "Random Forest," *Mach. Learn.*, vol. 45, no. 1, pp. 5–32, 2001.
4. O. Sagi and L. Rokach, "Ensemble learning: A survey," *Wiley Interdiscip. Rev. Data Min. Knowl. Discov.*, vol. 8, no. 4, pp. 1–18, 2018, doi: 10.1002/widm.1249.
5. V. Y. Kulkarni and P. K. Sinha, "Random forest classifiers: A survey and future research directions," *Int. J. Adv. Comput.*, vol. 36, no. 1, pp. 1144–1158, 2013.
6. D. B. Vuković, S. Dekpo-Adza, and S. Matović, "AI integration in financial services: A systematic review of trends and regulatory challenges," *Humanit. Soc. Sci. Commun.*, vol. 12, no. 562, 2025.
7. H. He, Y. Bai, E. A. Garcia, and S. Li, "ADASYN: Adaptive synthetic sampling approach for imbalanced learning," *Proc. Int. Jt. Conf. Neural Networks*, no. July 2008, pp. 1322–1328, 2008, doi: 10.1109/IJCNN.2008.4633969.
8. N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "SMOTE: Synthetic minority over-sampling technique," *J. Artif. Intell. Res.*, vol. 16,

no. February 2017, pp. 321–357, 2002, doi: 10.1613/jair.953.

9. J. Aizenman, J. Lee, and V. Sushko, "Exchange market pressure and its absorption: From the great moderation, to the global crisis (NBER Working Paper)," 2010.
10. L. Breiman, J. Friedman, R. Olshen, and C. Stone, *Classification and Regression Trees*. New York: hapman and Hall, Wadsworth, New York., 1984.
11. R. Piraei, M. Niazkar, S. H. Afzali, and A. Menapace, "Application of Machine Learning Models to Bridge Afflux Estimation," *Water (Switzerland)*, vol. 15, no. 12, pp. 1–19, 2023, doi: 10.3390/w15122187.
12. J. Bergstra and Y. Bengio, "Random search for hyper-parameter optimization," *J. Mach. Learn. Res.*, vol. 13, pp. 281–305, 2012.
13. S. Putatunda and K. Rama, "A comparative analysis of hyperopt as against other approaches for hyper-parameter optimization of XGBoost," *ACM Int. Conf. Proceeding Ser.*, pp. 6–10, 2018, doi: 10.1145/3297067.3297080.
14. T. Hasanin, T. M. Khoshgoftaar, J. L. Leevy, and R. A. Bauder, "Severely imbalanced Big Data challenges: investigating data sampling approaches," *J. Big Data*, vol. 6, no. 1, 2019, doi: 10.1186/s40537-019-0274-4.
15. S. Ahmed, F. Rayhan, A. Mahbub, M. Rafsan Jani, S. Shatabda, and D. M. Farid, "LIUboost: Locality informed under-boosting for imbalanced data classification," *Adv. Intell. Syst. Comput.*, vol. 813, pp. 133–144, 2019, doi: 10.1007/978-981-13-1498-8_12.
16. N. Cahyana, S. Khomsah, and A. S. Aribowo, "Improving Imbalanced Dataset Classification Using Oversampling and Gradient Boosting," *Proceeding - 2019 5th Int. Conf. Sci. Inf. Technol. Embrac. Ind. 4.0 Towar. Innov. Cyber Phys. Syst. ICSITech 2019*, pp. 217–222, 2019, doi: 10.1109/ICSITech46713.2019.8987499.
17. Z. Chen, "ADASYN – Random Forest Based Intrusion Detection Model," no. April 2020, pp. 1–13, 1999.
18. J. Han, M. Kamber, and J. Pei, *Data Mining Concepts and Techniques. 3rd edn*. New York: Waltham: Morgan Kaufmann Publishers., 2012.