Journal of Student Research Exploration

https://shmpublisher.com/index.php/josre

p-ISSN 2964-1691 | e-ISSN 2964-8246

Enhanced Out-of-Fold Stacking with Feature Grouping and Model-Specific Transformations for Diabetes Prediction Improvement

Ari Nugroho Putro[1], Sidiq Noor Kharisma[2], Gea Destadia Al-Zahra[3], Much Aziz Muslim[4], Dwika Ananda Agustina Pertiwi[5]

[1,2,3,4] Department of Computer Science, Universitas Negeri Semarang, Indonesia

[5] Faculty of Technology Management and Business, Universiti Tun Hussein Onn Malaysia, Malaysia

Article Info

Article history:

Received January 2, 2026
Revised January 9, 2026
Accepted January 15, 2026

Keywords:

Diabetes mellitus
Stacking ensemble
Out-of-fold
Feature grouping
Feature transformation

ABSTRACT

Diabetes mellitus is a chronic disease with serious implications for global health. Early detection is essential to reduce these risks, and machine learning methods are widely used in diabetes prediction. However, improving accuracy remains a major challenge in the development of predictive models. This study proposes a stacking-based ensemble learning approach with an out-of-fold (OOF) scheme to improve classification performance. The proposed method consists of several systematic steps, namely (1) data preprocessing via median imputation of invalid values and feature transformation according to model characteristics, (2) the creation of base learners comprising Logistic Regression, Gaussian Naïve Bayes, Support Vector Machine, Random Forest, and XGBoost, (3) model training using Stratified Cross Validation 5 Fold to generate OOF predictions, (4) combining all OOF predictions into a meta-feature matrix, and (5) training an XGBoost-based meta-model to generate the final prediction. This approach enables the meta-model to optimally learn the relationships among the outputs of the baseline models. Experimental results show that the proposed method achieves an accuracy of 91.15%, precision of 90.65%, recall of 83.21%, and an F1-score of 86.77%. These results indicate that stacking is effective in improving the accuracy of diabetes predictions.

1. Introduction

Health issues are critical concerns that require prompt and appropriate management [1]. In recent decades, diabetes mellitus has become one of the leading chronic diseases

[1] ***Corresponding Author:***
Ari Nugroho Putro,
Department of Computer Science,
State University of Semarang,
Sekaran, Gunungpati, Semarang City, Central Java 50229, Indonesia.
Email: arinugrohoputro@students.unnes.ac.id
DOI: https://doi.org/10.52465/josre.v4i1.674

contributing to rising global morbidity and mortality rates, as it leads to serious complications such as heart disease, kidney disorders, and nerve damage [2], [3]. The number of people with diabetes continues to rise each year, making early detection a critical step in reducing the risk of complications [4].

Symptoms of diabetes mellitus include blurred vision, fatigue, increased hunger, and frequent urination [5]. Risk factors such as abnormal blood pressure, obesity, and an unhealthy lifestyle contribute to the rise in diabetes cases. Pre-diabetes, characterized by insulin resistance, is an early stage of the disease's progression that can be managed if detected early [6].

The use of machine learning in diabetes prediction has advanced rapidly. Accuracy is the primary indicator for evaluating model performance. Several studies report high results, such as RFE-GRU with an accuracy of 90.70% [2] and SECNN at 89.47% [4], while other models like Random Forest achieve 79.57% [5]. These differences indicate that no single model is capable of consistently delivering optimal performance on diabetes data.

Previous research has focused on the use of single models or improvements in the preprocessing stage [7]. Previous research has not integrated a combination of heterogeneous models with feature grouping strategies and out-of-fold (OOF) prediction within a single systematic learning framework [8]. Consequently, these approaches have not previously addressed the limitations in capturing the variations of complex data patterns [9].

Given these limitations, this study proposes a stacking-based ensemble learning approach using heterogeneous models [10]. Each model is trained using a different subset of features and transformations tailored to its characteristics. Predictions from the base models are generated using an out-of-fold (OOF) scheme and then used as input for an XGBoost-based meta-model [11].

This approach is designed to improve prediction accuracy by combining the strengths of various models in capturing different data patterns [12]. The research stages include preprocessing, training base models with cross-validation, generating OOF predictions, training a meta-model, and evaluation.

The contributions of this research are: i) proposing an OOF-based stacking architecture with heterogeneous models, ii) integrating feature grouping strategies and model-specific transformations into a single learning framework, and iii) improving diabetes prediction accuracy through a structured multi-model approach.

2. Method

In detecting diabetes, this study employs several methodological steps, namely data preprocessing, the development of an ensemble-based classification model, out-of-fold (OOF) prediction, the development of a meta-model (meta-learner), and model evaluation [13]. The research workflow is shown in Figure 1.

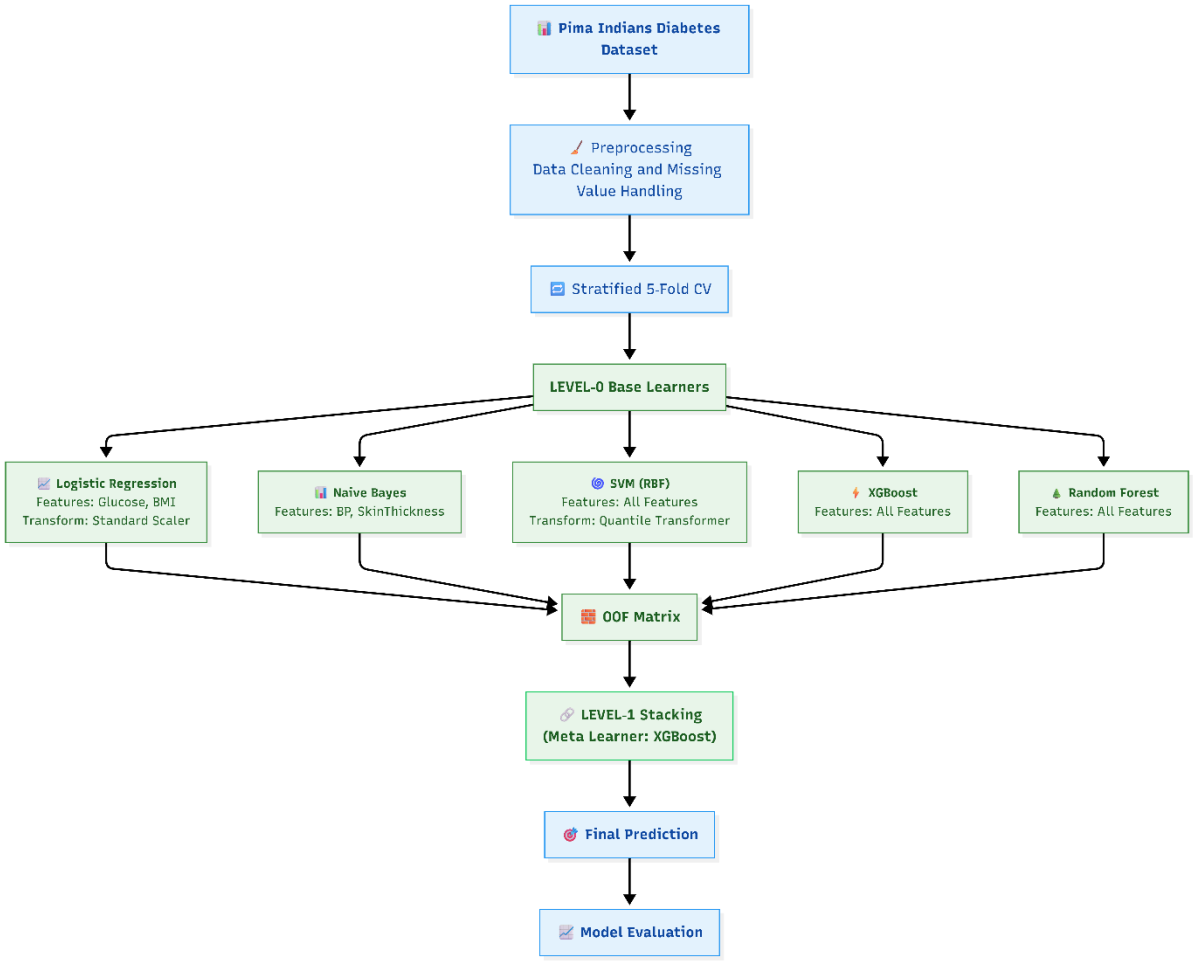

Figure 1. Research workflow

Data Description

This research dataset uses the Pima Indians Diabetes (PID) dataset, obtained from the Kaggle public repository via the link: https://www.kaggle.com/datasets/uciml/pima-indians-diabetes-database. This dataset consists of 768 samples with 8 predictor features and 1 class label (Outcome). The data distribution consists of 500 non-diabetes samples (class 0) and 268 diabetes samples (class 1).

Preprocessing

The preprocessing stage was conducted to ensure data quality prior to the modeling process [14]. In the PID dataset, several medical features—namely Glucose, BloodPressure, SkinThickness, and BMI—contained clinically invalid values of 0. These values were converted to missing values (NaN), and imputation was performed using the median value calculated from the training data in each fold [15].

To maintain the consistency of the data distribution in a specific model, feature transformations were performed according to the algorithm's characteristics. In the Logistic Regression model, StandardScaler was used for feature normalization [16]. In the Support Vector Machine model, the Quantile Transformer was used with a target normal distribution [17]. Meanwhile, the Gaussian Naïve Bayes, Random Forest, and XGBoost models used raw data without transformation because they are not sensitive to feature scale.

Base Learners

This study uses five classification models as base learners: Logistic Regression, Gaussian Naïve Bayes, Support Vector Machine (SVM), Random Forest, and Extreme Gradient Boosting (XGBoost). Logistic Regression is a linear model that models class probabilities using the sigmoid function, as shown in equation (1).

$$P(y = 1 \mid x) = \frac{1}{1+e^{-(w^T x+b)}} \quad (1)$$

This model was trained using a subset of features (Glucose, BMI) selected to represent key metabolic conditions.

Gaussian Naïve Bayes (GNB) is a probabilistic model that assumes independence among features and a Gaussian distribution for each class [18]. The conditional distribution is expressed as in equation (2).

$$P(x_i \mid y) = \frac{1}{\sqrt{2\pi\sigma_y^2}} \exp\left(-\frac{(x_i-\mu_y)^2}{2\sigma_y^2}\right) \quad (2)$$

This model was trained using the features (BloodPressure, SkinThickness).

A Support Vector Machine (SVM) is a model that seeks the optimal hyperplane with the maximum margin [19], as shown in equation (3).

$$\min_{w,b} \frac{1}{2} \| w \|^2 \text{ s.t.} y_i(w^T x_i + b) \geq 1, \forall i \quad (3)$$

SVM utilizes all features with the Quantile Transformer to improve data separability and generates probabilities using Platt scaling.

Random Forest (RF) is a decision tree-based ensemble method that combines the predictions of multiple trees [20], as shown in equation (4).

$$\hat{y}(x) = \text{mode}\{h_t(x)\}_{t=1}^{T} \quad (4)$$

This model uses all features without transformation and employs 300 decision trees to improve prediction stability.

Extreme Gradient Boosting (XGBoost) is a tree-based boosting method that is built iteratively [21], as shown in equation (5).

$$F_{m+1}(x) = F_m(x) + \eta h_m(x) \quad (5)$$

Out-of-Fold Prediction

To improve model generalization, a 5-fold stratified K-fold cross-validation scheme was used. In each fold, the data was split into training and validation sets with balanced class distributions. Each model is trained using the training data, then generates probability predictions on the validation data. This process is repeated across all folds so that each sample receives exactly one prediction on the validation data. All predictions from each model are combined into an $N \times M$ out-of-fold (OOF) matrix, where N is the number of samples and M is the number of models.

Meta Learner

The OOF matrix is used as a new feature for training XGBoost-based meta-models. This model combines all predictions from the base models to generate a final prediction [21]. With this approach, the meta-model learns the relationships between the base model's predictions to improve classification performance.

Model Evaluation

Model evaluation was performed using the prediction results from the meta-model. The data was split using a 5-fold stratified K-fold cross-validation scheme [22]. Model performance was measured using a confusion matrix consisting of true positive (TP), true negative (TN), false positive (FP), and false negative (FN). Based on these values, evaluation metrics were calculated, including accuracy, precision, recall, F1-score [23].

3. Results and Discussion

This section presents the results of experiments conducted to evaluate the performance of the proposed model in predicting diabetes. The evaluation was conducted in stages to analyze the contribution of each component in the developed architecture, namely a single model as a baseline, the use of stacking, the impact of feature transformations [24], and feature grouping strategies [25]. This step-by-step approach aims to provide a clearer understanding of the factors contributing to improved model performance.

3.1. Single Model Performance

As a first step, an evaluation was conducted on the individual models used as a baseline. The models tested included Logistic Regression, Gaussian Naïve Bayes, Support Vector Machine, Random Forest, and XGBoost, using all features without any transformations. The performance of the single model is shown in table 1.

Table 1. Single model performance

Model	Accuracy	Precision	Recall	F1-score
Logistic Regression	76.82%	71.84%	55.22%	62.45%
Gaussian Naïve Bayes	74.61%	64.78%	59.70%	62.14%
Support Vector Machine	75.65%	70.98%	51.12%	59.44%
Random Forest	76.43%	68.35%	60.45%	64.16%
XGBoost	73.83%	63.79%	57.84%	60.67%

Based on these results, Logistic Regression demonstrated the best performance with an accuracy of 76.82%. However, overall, all single models still exhibit limitations in achieving optimal performance. This indicates that single models are not yet capable of optimally capturing the complexity of patterns in diabetes data.

3.2. The Impact of Stacking

Next, we evaluated the impact of using stacking on model performance. In this phase, stacking was applied using all features without any transformations. The accuracy of the stacking model can be seen in the table 2.

Table 2. Stacking model accuracy

Model	Accuracy
Logistic Regression (Baseline Terbaik)	76.82%
Stacking (All Features, Without Transformation)	90.89%

The results show that the use of stacking yields a highly significant improvement in accuracy, from 0.7682 to 0.9089. This improvement indicates that stacking can combine the strengths of various models to produce more accurate predictions. By leveraging the outputs of multiple models with different characteristics, the meta-model can construct a more comprehensive representation of the data.

3.3. The Effect of Feature Transformation

The following analysis was conducted to evaluate the impact of feature transformations on the performance of the stacking model. The transformations were applied according to the characteristics of each model. The accuracy of stacking with feature transformation can be seen in table 3.

Table 3. Stacking with feature transformation accuracy

Model	Accuracy
Stacking (All Features, Without Transformation)	90.89%
Stacking (All Features with Transformation)	90.36%

The results show that applying transformations to all features does not improve performance; in fact, it slightly reduces accuracy. This indicates that transformations do not always have a positive effect when applied universally to all models. The effectiveness of transformations depends heavily on the alignment between the data distribution and the specific requirements of each algorithm.

3.4. The Effect of Feature Grouping

Next, an analysis was conducted on the impact of the feature grouping strategy, which involves dividing features into several subsets used by a specific model. The accuracy of stacking with feature grouping is shown in table 4.

Table 4. Stacking with feature grouping accuracy

Model	Accuracy
Stacking (All Features, Without Transformation)	90.89%
Stacking (Feature Grouping, Without Transformation)	89.97%

The results show that using feature grouping without transformation reduces the model's performance. This may be due to the reduced amount of information available in each model because of feature partitioning, meaning that some important patterns in the data cannot be captured optimally.

3.5. Combination of Feature Grouping and Transformation

The next step is to evaluate the combination of feature grouping and feature transformation. The accuracy of stacking with feature transformation and feature grouping is shown in table 5.

Table 5. Stacking with feature transformation and feature grouping accuracy

Model	Accuracy
Stacking (All Features with Transformation)	90.36%
Stacking (Feature Grouping with Transformation)	91.15%

The results show that the combination of feature grouping and transformation yields the best performance, with an accuracy of 91.15%. This indicates that feature grouping is effective when combined with transformations that align with the model's characteristics. Transformations help adjust the data distribution, while feature grouping allows each model to focus on the most relevant subset of features.

3.6. Overall Comparison of Models

A summary of the performance of all models is presented in table 6.

Table 6. Summary of the performance of all models

Model	Accuracy	Precision	Recall	F1-score
Logistic Regression	76.82%	71.84%	55.22%	62.45%
Gaussian Naïve Bayes	74.61%	64.78%	59.70%	62.14%
Support Vector Machine	75.65%	70.98%	51.12%	59.44%
Random Forest	76.43%	68.35%	60.45%	64.16%
XGBoost	73.83%	63.79%	57.84%	60.67%
Stacking (All Features, Without Transformation)	90.89%	90.24%	82.84%	86.38%
Stacking (All Features with Transformation)	90.36%	89.11%	82.46%	85.66%
Stacking (Feature Grouping, Without Transformation)	89.97%	89.63%	80.60%	84.87%
Stacking (Feature Grouping with Transformation)	91.15%	90.65%	83.21%	86.77%

The results of this study indicate that model performance is influenced not only by the choice of algorithm but also by the learning strategy employed. The out-of-fold (OOF) stacking approach was found to yield a significant performance improvement compared to a single model. Furthermore, the combination of feature grouping and transformations tailored to the model's characteristics contributes to further improvements in accuracy, whereas feature grouping without transformations tends to reduce performance due to limitations in feature representation.

Compared to previous studies, the approach proposed in this study demonstrates superior performance. A summary of the comparison with previous studies is presented in Table 7.

Table 7. Comparison with previous studies

Study	Dataset	Best Model	Accuracy
Bhattacharya & Datta (2025)	Pima Indians Diabetes	Gradient Boosting	76.00%
Chang et al. (2021)	Pima Indians Diabetes	Random Forest	79.57%
Zhao et al. (2024)	Pima Indians Diabetes	SECNN	89.47%
Shams et al. (2025)	Pima Indians Diabetes	RFE-GRU	90.70%
Our Model	**Pima Indians Diabetes**	**OOF Stacking**	**91.15%**

These results demonstrate that the proposed approach outperforms previous methods by leveraging a combination of heterogeneous models, feature grouping, and feature transformations within an OOF-based stacking framework. Thus, the structured learning architecture has proven to play a crucial role in improving the accuracy of diabetes prediction.

4. Conclusion

This study proposes a stacking-based ensemble approach with an out-of-fold (OOF) scheme for diabetes prediction using heterogeneous models, feature grouping, and feature transformations. Experimental results show that the proposed method significantly improves accuracy compared to a single model, with an increase from 76.82% to 91.15%. Evaluation results indicate that stacking is the primary component contributing to performance improvement through the combination of outputs from various base learners. Furthermore, feature transformation does not always yield improvements when applied globally but can enhance performance when combined with feature grouping. Conversely, feature grouping without transformation tends to degrade performance due to limitations in information representation.

Overall, the combination of feature grouping and transformations in the stacking framework yielded the best performance, indicating that alignment between model characteristics, feature subsets, and data distribution is crucial for improving a model's generalization ability. This approach proved effective in capturing complex patterns in diabetes data and has the potential to be applied to other classification problems with similar characteristics.

REFERENCES

[1] C. Xu, F. Shi, W. Ding, C. Fang, and C. Fang, "Development and validation of a machine learning model for cardiovascular disease risk prediction in type 2 diabetes patients," *Sci. Rep.*, vol. 15, no. 1, Dec. 2025, doi: 10.1038/s41598-025-18443-7.

[2] M. Y. Shams, Z. Tarek, and A. M. Elshewey, "A novel RFE-GRU model for diabetes classification using PIMA Indian dataset," *Sci. Rep.*, vol. 15, no. 1, Dec. 2025, doi: 10.1038/s41598-024-82420-9.

[3] J. Zhao, H. Gao, C. Yang, T. An, Z. Kuang, and L. Shi, "Attention-Oriented CNN Method for Type 2 Diabetes Prediction," *Applied Sciences (Switzerland)*, vol. 14, no. 10, May 2024, doi: 10.3390/app14103989.

[4] J. Zhao, H. Gao, C. Yang, T. An, Z. Kuang, and L. Shi, "Attention-Oriented CNN Method for Type 2 Diabetes Prediction," *Applied Sciences (Switzerland)*, vol. 14, no. 10, May 2024, doi: 10.3390/app14103989.

[5] V. Chang, J. Bailey, Q. A. Xu, and Z. Sun, "Pima Indians diabetes mellitus classification based on machine learning (ML) algorithms," *Neural Comput. Appl.*, vol. 35, no. 22, pp. 16157–16173, Aug. 2023, doi: 10.1007/s00521-022-07049-z.

[6] M. Bhattacharya and D. Datta, "Intelligent Models for Diabetic Prediction Using Conventional Machine Learning Techniques and Ensemble Learning Algorithms," *SN Comput. Sci.*, vol. 6, no. 1, Jan. 2025, doi: 10.1007/s42979-024-03479-9.

[7] S. Shafi and G. A. Ansari, "Heart Disease Prediction Using Machine Learning with Metaheuristic Feature Selection Approaches," *Biomedical Materials and Devices*, 2025, doi: 10.1007/s44174-025-00507-x.

[8] Y. Yuan, J. Wei, H. Huang, W. Jiao, J. Wang, and H. Chen, "Review of resampling techniques for the treatment of imbalanced industrial data classification in equipment condition monitoring," Nov. 01, 2023, *Elsevier Ltd*. doi: 10.1016/j.engappai.2023.106911.

[9] Q. A. Hidayaturrohman and E. Hanada, "Impact of Data Pre-Processing Techniques on XGBoost Model Performance for Predicting All-Cause Readmission and Mortality Among Patients with Heart Failure," *BioMedInformatics*, vol. 4, no. 4, pp. 2201–2212, Dec. 2024, doi: 10.3390/biomedinformatics4040118.

[10] S. Siraj, F. H. Dahri, J. A. Chandio, A. H. Jalbani, and A. A. Laghari, "Comparison of machine learning techniques to predict students' CGPA by using course learning outcomes datasets," *Human-Intelligent Systems Integration*, Apr. 2025, doi: 10.1007/s42454-025-00063-1.

[11] I. P. Nguemkam Tebou, N. Tsopze, and D. Tchuente, "Hybrid Method to Explain Predictions of Stacking Ensemble Model," *Information Systems Frontiers*, Feb. 2026, doi: 10.1007/s10796-025-10684-1.

[12] S. Shafieian and M. Zulkernine, “Multi-layer stacking ensemble learners for low footprint network intrusion detection,” *Complex and Intelligent Systems*, vol. 9, no. 4, pp. 3787–3799, Aug. 2023, doi: 10.1007/s40747-022-00809-3.

[13] M. A. Muslim *et al.*, “New model combination meta-learner to improve accuracy prediction P2P lending with stacking ensemble learning,” *Intelligent Systems with Applications*, vol. 18, May 2023, doi: 10.1016/j.iswa.2023.200204.

[14] M. Sagming, R. Heymann, and M. V. Visaya, “Using topological data analysis and machine learning to predict customer churn,” *J. Big Data*, vol. 11, no. 1, Dec. 2024, doi: 10.1186/s40537-024-01020-6.

[15] E. Alsharif and M. Alharby, “An Ensemble Machine Learning Approach for Detecting and Classifying Malware Attacks on Mobile Devices,” *Arab. J. Sci. Eng.*, vol. 50, no. 19, pp. 15825–15841, Oct. 2025, doi: 10.1007/s13369-025-10011-5.

[16] C. T. Doan and H. Du Nguyen, “Robust water quality prediction across multiple indicator formulations using an explainable ensemble learning model,” *Water Resour. Ind.*, vol. 34, Dec. 2025, doi: 10.1016/j.wri.2025.100329.

[17] D. A. Debal and T. M. Sitote, “Chronic kidney disease prediction using machine learning techniques,” *J. Big Data*, vol. 9, no. 1, Dec. 2022, doi: 10.1186/s40537-022-00657-5.

[18] K. Kevin, M. Enjeli, and A. Wijaya, “Analisis Sentimen Pengunaaan Aplikasi Kinemaster Menggunakan Metode Naive Bayes,” *Jurnal Ilmiah Computer Science*, vol. 2, no. 2, pp. 89–98, Jan. 2024, doi: 10.58602/jics.v2i2.24.

[19] B. Chao and H. Guangqiu, “Innovative SVM optimization with differential gravitational fireworks for superior air pollution classification,” *Sci. Rep.*, vol. 14, no. 1, Dec. 2024, doi: 10.1038/s41598-024-75839-7.

[20] S. Singh, M. Kumar, B. K. Verma, and S. Kumar, “Optimizing Air Pollution Prediction With Random Forest Algorithm,” *Aerosol Science and Engineering*, 2025, doi: 10.1007/s41810-025-00292-6.

[21] M. A. Muslim and Y. Dasril, “Company bankruptcy prediction framework based on the most influential features using XGBoost and stacking ensemble learning,” *International Journal of Electrical and Computer Engineering (IJECE)*, vol. 11, no. 6, p. 5549, Dec. 2021, doi: 10.11591/ijece.v11i6.pp5549-5557.

[22] J. Sadaiyandi, P. Arumugam, A. K. Sangaiah, and C. Zhang, “Stratified Sampling-Based Deep Learning Approach to Increase Prediction Accuracy of Unbalanced Dataset,” *Electronics (Basel).*, vol. 12, no. 21, p. 4423, Oct. 2023, doi: 10.3390/electronics12214423.

[23] P. M. Vieira and F. Rodrigues, “An automated approach for binary classification on imbalanced data,” *Knowl. Inf. Syst.*, vol. 66, no. 5, pp. 2747–2767, May 2024, doi: 10.1007/s10115-023-02046-7.

[24] A. Masood, M. Niazkar, M. Zakwan, and R. Piraei, “A Machine Learning-Based Framework for Water Quality Index Estimation in the Southern Bug River,” *Water (Switzerland)*, vol. 15, no. 20, Oct. 2023, doi: 10.3390/w15203543.

[25] H. Ahmad, B. Kasasbeh, B. Aldabaybah, and E. Rawashdeh, “Class balancing framework for credit card fraud detection based on clustering and similarity-based selection (SBS),” *International Journal of Information Technology (Singapore)*, vol. 15, no. 1, pp. 325–333, Jan. 2023, doi: 10.1007/s41870-022-00987-w.