Optimization of Random Forest for Health Data Classification Using PCA and K-Means SMOTE-ENN

Dadang Priyanto
Department of Computer Science, Universitas Bumigora, Mataram, Indonesia
dadang.priyanto@universitasbumigora.ac.id (corresponding author)

Hairani Hairani
Department of Computer Science, Universitas Bumigora, Mataram, Indonesia
hairani@universitasbumigora.ac.id

Khairan Marzuki
Department of Computer Science, Universitas Bumigora, Mataram, Indonesia
khairan@universitasbumigora.ac.id

Muhammad Innuddin
Department of Computer Science, Universitas Bumigora, Mataram, Indonesia
inn@universitasbumigora.ac.id

Received: 1 July 2025 | Revised: 21 July 2025 and 3 August 2025 | Accepted: 15 August 2025

 DOI: https://doi.org/10.48084/etasr.12976

ABSTRACT

Health data classification is a significant challenge in the healthcare field, particularly due to the inherent characteristics of health data, which typically exhibit high dimensionality and imbalanced class distributions. These factors can complicate the training process of classification models and adversely affect their performance and accuracy. Consequently, a method is required to address data complexity and class imbalance, ensuring that the resulting information is both accurate and reliable. This study aims to improve the performance of the Random Forest (RF) classification model when processing health data by integrating two primary approaches: Principal Component Analysis (PCA) and K-Means SMOTE-ENN. PCA is instrumental in reducing data dimensions while extracting the most informative features, thus minimizing noise and reducing computational demands. Meanwhile, K-Means SMOTE-ENN serves to balance class distribution through a combination of clustering-based oversampling and Edited Nearest Neighbors-based data cleaning, effectively addressing the issue of overfitting caused by unrepresentative synthetic data. The RF classification model was chosen, recognized for its strong performance in managing data with high dimensions and complex variable interactions. Experimental results indicate that the joint application of PCA and K-Means SMOTE-ENN significantly enhances the model performance. In the Pima Indians Diabetes dataset, accuracy rose to 98.41%, and the Area Under Curve (AUC) value reached 98.33%. For the Heart Disease dataset, an accuracy of 97.56% and an AUC of 97.73% were achieved. Compared with previous methods, the proposed approach achieves 2.91% accuracy improvement with SMOTE and Stacking Ensemble on the Pima Indians Diabetes dataset and 6.26% accuracy improvement and 14.73% AUC improvement compared with XGBoost on the Heart Disease dataset. These results show that combining PCA with K-Means SMOTE-ENN significantly improves the performance of RF on imbalanced healthcare data.

Keywords-K-Means SMOTE-ENN; PCA; health data; data imbalance; data reduction

I. INTRODUCTION

In recent years, advances in computer technology have significantly driven progress in data mining and machine learning technologies [1, 2]. These technologies are increasingly being applied to extract valuable information from data in various fields, particularly in the healthcare sector [3, 4]. Their role in healthcare is essential to solving a range of problems. One key application is disease classification, which supports early detection, more accurate diagnosis, and more

effective treatment planning. However, in practice, this process often faces challenges due to the high dimensionality and imbalanced class distribution of health data, which can negatively impact the performance of classification models. For example, the Pima dataset for the classification of diabetes contains an imbalanced class distribution. A similar issue exists in heart disease data, such as the Heart Disease dataset, which suffers from class imbalance and high dimensionality. These issues highlight the importance of applying data preprocessing techniques, such as dimensionality reduction and class balancing, to improve classification model performance.

In [5], Principal Component Analysis (PCA) was applied for feature extraction in breast cancer classification using several machine learning methods and Deep Neural Networks (DNN), finding that the combination of PCA and DNN outperformed conventional DNNs and traditional machine learning classifiers, with PCA improving accuracy by 3.68% over conventional DNNs. However, this study only performed feature extraction without addressing the class imbalance problem. In [6], the Radius-SMOTE method was applied, while in [7], imbalanced health data were handled without performing feature reduction on the dataset.

In [8], the Outlier-SMOTE method was used to address the issue of imbalanced data without performing feature reduction on health data. In [9], the RN-SMOTE method was applied to handle imbalanced data without selecting the relevant features to be processed. In [10], various machine learning methods were applied for early classification of Parkinson's Disease, with the results showing that the Random Forest (RF) method performed best compared to others, achieving an accuracy of 91.83%. In [11], the NR-Clustering SMOTE method was developed to address the issue of imbalanced health data. This method achieved an accuracy of 89.56% and an AUC of 89.56% on the PIMA dataset, as well as an accuracy of 89.84% and an AUC of 89.84% on the Haberman dataset.

Table I presents a comparison between this and some previous studies. Previous studies have focused on addressing class imbalance through various sampling techniques, such as SMOTE, NR-Clustering SMOTE, and NR-Modified SMOTE. However, no approach integrates feature extraction methods for dimensionality reduction before the resampling process. This lack of integration has the potential to degrade model performance due to the presence of noisy or redundant features.

TABLE I. COMPARISON BETWEEN THIS AND PREVIOUS RELATED STUDIES.

Ref	Dataset	Feature dimensionality	Hybrid sampling	Method
[11]	Pima	-	NR-Clustering SMOTE	RF
[12]		-	NR-Modified SMOTE	RF
[13]		-	SMOTE	Stacking ensemble
[14]	Heart	-	-	XGBoost
[15]		-	-	ANN
Proposed method	Pima and Heart Disease	PCA	K-Means SMOTE	K-Means SMOTE

This study addresses this gap using a different approach, integrating feature extraction with PCA and handling class imbalance using K-Means SMOTE-ENN. This is a novel integration design that has not been explored in previous studies. The PCA method is used as a dimensionality reduction technique to extract the most relevant key features from the data [16-18]. On the other hand, to handle class imbalance, the K-Means SMOTE-ENN method is applied to add synthetic data to the minority class [19], while ENN cleans the data from outliers and noise, resulting in a more balanced and cleaner data distribution [20]. The RF classification method is used on health data due to its ability to handle high-dimensional data and its robustness against overfitting [21]. Therefore, this study aims to integrate PCA and K-Means SMOTE-ENN methods to improve the performance of the RF model in classifying health data. The results of this study are expected to contribute to the development of more effective health data classification methods, particularly through the integration of PCA and K-Means SMOTE-ENN to address high-dimensional data and imbalanced class issues, resulting in more accurate, reliable, and applicable models in health data analysis.

II. RESEARCH METHODS

Figure 1 shows the proposed method to produce an optimal classification model for health data with an imbalanced class distribution. The method consists of several main steps: data collection, data preprocessing, building a classification model using RF, and evaluating its performance using classification evaluation metrics. This study uses two primary datasets from Kaggle: the Pima [22] and Heart Disease datasets [23]. The Pima dataset consists of 768 instances with eight input attributes and one output, and an imbalance ratio of 1.87. The Heart Disease dataset contains 303 instances with 13 input attributes and one output, and an imbalance ratio of 1.20.

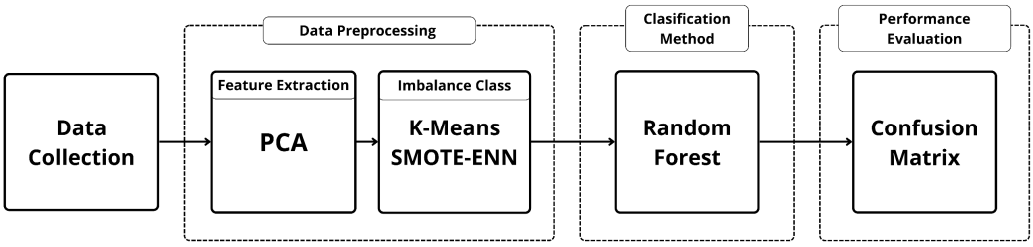

Fig. 1. Research method.

Data preprocessing is a crucial step in the data analysis and machine learning pipeline, as data quality greatly influences the model's performance. The preprocessing in this study involves two main processes: PCA for dimensionality reduction and K-Means SMOTE-ENN for handling class imbalance. PCA is a widely used statistical method for reducing the number of variables in a dataset while retaining essential information [24]. The PCA procedure consists of six main steps [25]. First, compute the covariance matrix of the normalized d -dimensional dataset using (1). Second, the eigenvalues and corresponding eigenvectors of the covariance matrix are determined using (2). Third, sort the eigenvalues in descending order. Fourth, select k eigenvectors corresponding to the k largest eigenvalues, where k represents the dimensions in the new feature subspace. Fifth, construct a projection matrix using the selected k eigenvectors using (3). Sixth, transform the original data into the new k-dimensional feature space.

$$C = \frac{1}{N-1} X^T X \quad (1)$$

$$C * v = \gamma * v \quad (2)$$

$$T = X * v \quad (3)$$

where C is the covariance matrix and X is the feature data. γ is the eigenvalue, v is the eigenvector value, and T is the principal component score.

K-Means SMOTE-ENN is a technique that addresses class imbalance more effectively than conventional resampling methods [26, 27]. It integrates K-Means SMOTE for oversampling and ENN for cleaning both synthetic and original majority class data that may be noisy [28]. K-Means SMOTE is an extension of the SMOTE method that uses the K-Means algorithm to first cluster minority data before generating synthetic samples. This approach focuses the sampling process on minority clusters, reducing class overlap. The number of clusters in K-Means SMOTE is set to the default value of 100. ENN is then applied to remove the majority class data misclassified by its k-nearest neighbors, aiming to reduce noise and clarify class boundaries using $k = 3$. The variable k is the number of nearest neighbors used to evaluate each instance in the dataset. Combining these methods produces a more balanced, clean, and representative dataset, improving classification model performance. This technique significantly enhances model accuracy and generalization by addressing class imbalance and reducing noise. The dataset was divided into training and testing with a ratio of 80:20.

RF is an ensemble method that works by constructing many decision trees, each trained on a randomly selected subset of the training data using bootstrap sampling. At each tree split, only a random subset of features is considered, increasing diversity among trees and effectively reducing the risk of overfitting. Each tree makes an independent prediction, and the final result is determined by majority voting for classification tasks. This method is more stable and accurate than a single decision tree, as it reduces variance, improves generalization, and is more resistant to complex or noisy data. Another advantage is its ability to handle high-dimensional data and directly estimate the importance of each feature, which is helpful for feature selection and model interpretation.

The confusion matrix is an evaluation tool used to assess the performance of classification methods (see Table II). It consists of four key components: True Positives (TP), True Negatives (TN), False Positives (FP), and False Negatives (FN) [29]. TN represents the number of negative instances correctly classified, while FP indicates the number of negative instances incorrectly classified as positive. Based on these values, the confusion matrix is used to calculate various evaluation metrics such as Accuracy and AUC. This study evaluates both Accuracy and AUC, where Accuracy is calculated using (4) [30, 31] and AUC using (5) [32].

TABLE II. CONFUSION MATRIX

Actual	**Prediction**	
	Negative	**Positive**
Negative	TN	FP
Positive	FN	TP

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

$$\text{AUC} = \frac{(\text{Recall} * \text{Specificity})}{2} \quad (5)$$

III. RESULTS AND DISCUSSION

This section discusses the results and findings of this study. The research used two health-related datasets, namely Pima [22] and Heart Disease [23], obtained from Kaggle. The Pima dataset contains 768 samples, while the Heart Disease dataset has 303 samples. Initially, feature extraction was performed using the PCA method to reduce data dimensions while retaining the most important information. This helps simplify the data structure, reduce model complexity, and speed up the training process without significantly affecting classification performance. However, both datasets have imbalanced class distributions, which can affect model performance. To address this issue, the K-Means SMOTE-ENN method was applied, which performs oversampling on the minority class by generating synthetic samples, while ENN removes potentially noisy samples from the majority class based on comparisons with the nearest neighbors. Table III shows the data distribution before and after balancing.

TABLE III. DATA DISTRIBUTION BEFORE AND AFTER HYBRID SAMPLING

Dataset	**Original**		**Hybrid sampling**	
	Negative	**Positive**	**Negative**	**Positive**
Pima	500	268	340	290
Heart Disease	165	138	102	102

After the model training process was complete using the default hyperparameters of the classification method, performance evaluation was carried out using a confusion matrix to assess how well the model could distinguish between positive and negative classes. Figures 2-7 show the confusion matrices. For the Pima dataset (Figure 2), the RF method produced 77 correct predictions for the positive class (TP) and 34 for the negative class (TN), with 43 misclassifications (22 FP and 21 FN). When PCA was used with RF (Figure 3), performance improved, with 93 TPs and 28 TNs, and a reduction in misclassifications to 33 cases. Combining PCA

and K-Means SMOTE-ENN with RF (Figure 4) gave the best results. The model classified 58 positive cases and 66 negative cases with only two errors (FN).

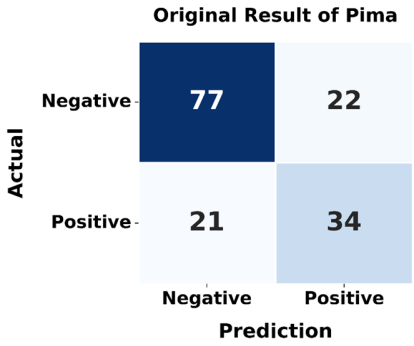

Fig. 2. Confusion matrix results of Pima original data.

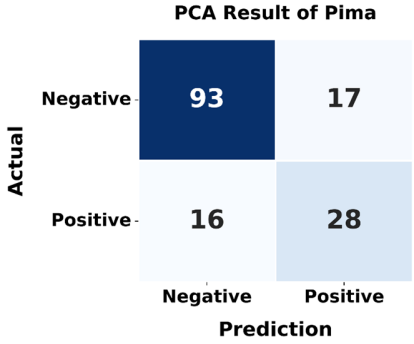

Fig. 3. Confusion matrix results of Pima data with PCA.

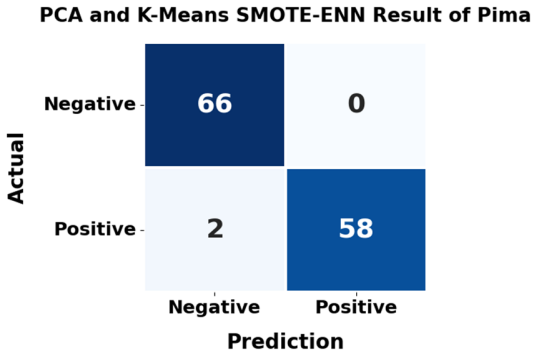

Fig. 4. Confusion matrix results of Pima data with the proposed method.

RF produced fairly good results for the Heart Disease dataset with 24 TPs and 27 TNs, and 10 misclassifications (Figure 5). Using PCA with RF (Figure 6) increased the TP count to 30, although FP also increased. Meanwhile, combining PCA and K-Means SMOTE-ENN with RF (Figure 7) achieved the best performance. The model correctly classified 21 positive cases with only 1 FP, and all 19 negative cases with no FN, indicating high accuracy. This demonstrates that well-balanced data and appropriately reduced features can lead to more accurate and reliable classifications.

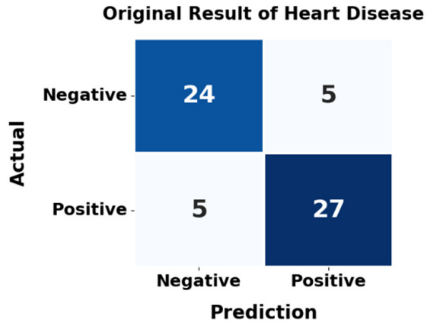

Fig. 5. Confusion matrix results of Heart Disease original data.

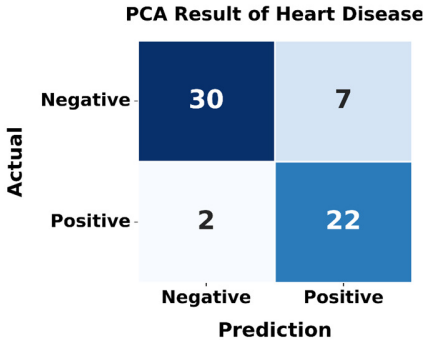

Fig. 6. Confusion matrix results of Heart Disease data with PCA.

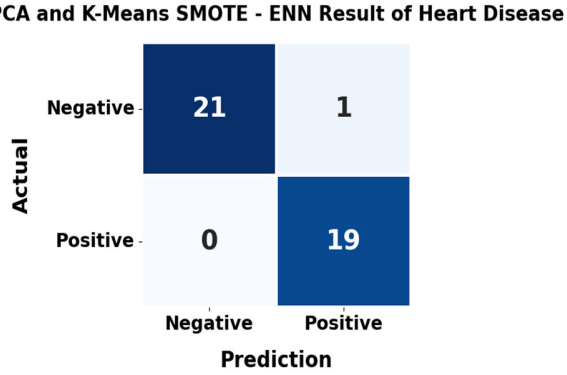

Fig. 7. Confusion matrix results of Heart Disease data with the proposed method.

As shown in Table IV, on the Pima dataset, the RF model without preprocessing achieved an accuracy of 72.08% and an AUC of 69.8%. After applying PCA for feature extraction, performance improved, with accuracy increasing to 78.57% and AUC to 74.1%. This indicates that PCA effectively removed less relevant features while preserving essential information. When data balancing using K-Means SMOTE-ENN was applied alongside PCA, model performance improved significantly, with accuracy reaching 98.41% and AUC rising to 98.33%. These results suggest that balancing the class distribution and removing noisy data from both minority and majority classes helps reduce model bias and enable better generalization across classes [11]. In the Heart Disease dataset, the RF model without optimization achieved 83.61% accuracy and 83.57% AUC. After applying PCA, accuracy increased to

85.25%, and AUC to 86.38%. The combined use of PCA and K-Means SMOTE-ENN improved significantly, yielding 97.56% accuracy and 97.73% AUC.

TABLE IV. PERFORMANCE OF CLASSIFICATION WITH RF

Dataset	Method	Accuracy	AUC
Pima	Original	74.08%	69.80%
	PCA	78.57%	74.10%
	PCA and K-Means SMOTE-ENN (Proposed method)	**98.41%**	**98.33%**
Heart Disease	Original	83.61%	83.57%
	PCA	85.25%	86.38%
	PCA and K-Means SMOTE-ENN (Proposed method)	**97.56%**	**97.73%**

Using PCA and K-Means SMOTE-ENN together can significantly improve the performance of both datasets. PCA helps reduce data dimensions and remove irrelevant features, while K-Means SMOTE-ENN addresses class imbalance that can cause model bias. Proper preprocessing strategies have increased accuracy and generalization, as higher AUC values indicate.

This study found that the combination of PCA and K-Means SMOTE-ENN positively affects the performance of the RF model, as seen in the significant increase in accuracy and AUC scores. However, K-Means SMOTE-ENN has a greater influence on performance improvement than PCA alone, which is evident from the sharp increase in accuracy and AUC after applying K-Means SMOTE-ENN. PCA helps simplify data and improve performance to a certain level, but the largest contribution to accuracy and AUC comes from the data balancing process. This is consistent with previous studies, such as [33], which showed that PCA can improve classification model performance, and [34, 35], which proved that K-Means SMOTE-ENN is effective in enhancing classification performance. These studies confirm that PCA and K-Means SMOTE-ENN can extract features effectively and handle class imbalance optimally, improving classification models.

The proposed method outperforms previous approaches. It combines PCA for dimensionality reduction, K-Means SMOTE-ENN for data balancing, and the RF algorithm as the classification model. This method significantly increases accuracy and AUC [36], as clearly shown in Table IV. Based on Table V, the proposed method consistently outperforms other methods in terms of both accuracy and AUC. PCA is proven to reduce the dimensions of the data without losing important information. Meanwhile, K-Means SMOTE-ENN handles class imbalance more effectively than conventional resampling techniques. This combination, together with the RF algorithm—known for its robustness and ability to manage non-linear relationships—produces a reliable and high-performance classification model [37-39].

However, a limitation of this study lies in using datasets with a low imbalance ratio. Therefore, further experiments are needed with datasets that have an imbalance ratio above 1:10 to test the reliability of the proposed method.

TABLE V. COMPARISON OF THE RESULTS OF THE PROPOSED METHOD WITH PREVIOUS STUDIES

Study	Dataset	Method	Accuracy	AUC
[40]	Pima	CNN	85.00%	89.00%
[13]		SMOTE Stacking Ensemble	95.50%	-
[12]		NR-Modified SMOTE	89.36%	89.36%
[41]		Backward Elimination and SVM	85.71%	-
[15]	Heart Disease	ANN	86.00%	-
[14]		XGBoost	91.30%	83.00%
[42]		MLP-PSO	84.61%	-
Proposed method	Pima	PCA and K-Means SMOTE-ENN with RF	**98.41%**	**98.38%**
	Heart Disease		**97.56%**	**97.73%**

IV. CONCLUSION

Based on the results of the study, the application of feature extraction using PCA has been proven to improve the performance of the RF classification method. Additionally, using K-Means SMOTE-ENN to address class imbalance in a dataset that has undergone PCA significantly boosts the performance of the RF model. Although PCA contributes to performance improvement, K-Means SMOTE-ENN has a greater impact in this study, especially because both datasets are imbalanced. The application of RF with PCA and K-Means SMOTE-ENN on the Pima dataset achieved an accuracy of 97.5% and an AUC of 97.47%. On the Heart Disease dataset, the same method achieved an accuracy of 95.12% and an AUC of 95.45%. Therefore, it can be concluded that using PCA and K-Means SMOTE-ENN together in the RF method is the most optimal approach to improve classification performance for healthcare data with imbalanced class characteristics. Future research could further explore the effect of the number of clusters in the K-Means algorithm and the value of the parameter k in the ENN method on model performance, to find the most optimal configuration for data balancing.

ACKNOWLEDGMENT

The authors express their deepest gratitude to KEMDIKTISAINTEK for the financial support provided through the Fundamental Research (PFR) scheme in 2025.

REFERENCES

[1] B. Charbuty and A. Abdulazeez, "Classification Based on Decision Tree Algorithm for Machine Learning," *Journal of Applied Science and Technology Trends*, vol. 2, no. 01, pp. 20–28, Mar. 2021, https://doi.org/10.38094/jastt20165.

[2] Y. Zeng and F. Cheng, "Medical and Health Data Classification Method Based on Machine Learning," *Journal of Healthcare Engineering*, vol. 2021, pp. 1–5, Nov. 2021, https://doi.org/10.1155/2021/2722854.

[3] R. F. Mansour, A. E. Amraoui, I. Nouaouri, V. G. Diaz, D. Gupta, and S. Kumar, "Artificial Intelligence and Internet of Things Enabled Disease Diagnosis Model for Smart Healthcare Systems," *IEEE Access*, vol. 9, pp. 45137–45146, 2021, https://doi.org/10.1109/ACCESS.2021.3066365.

[4] E. Šabić, D. Keeley, B. Henderson, and S. Nannemann, "Healthcare and anomaly detection: using machine learning to predict anomalies in heart rate data," *AI & SOCIETY*, vol. 36, no. 1, pp. 149–158, Mar. 2021, https://doi.org/10.1007/s00146-020-00985-1.

[5] P. Rani, R. Kumar, A. Jain, R. Lamba, R. Kumar Sachdeva, and T. Choudhury, "PCA-DNN: A Novel Deep Neural Network Oriented

System for Breast Cancer Classification," *EAI Endorsed Transactions on Pervasive Health and Technology*, vol. 9, Oct. 2023, https://doi.org/10.4108/eetpht.9.3533.

[6] G. A. Pradipta, R. Wardoyo, A. Musdholifah, and I. N. H. Sanjaya, "Radius-SMOTE: A New Oversampling Technique of Minority Samples Based on Radius Distance for Learning From Imbalanced Data," *IEEE Access*, vol. 9, pp. 74763–74777, 2021, https://doi.org/10.1109/ACCESS.2021.3080316.

[7] L. Yuningsih, G. A. Pradipta, D. Hermawan, P. D. W. Ayu, D. P. Hostiadi, and R. R. Huizen, "IRS-BAG-Integrated Radius-SMOTE Algorithm with Bagging Ensemble Learning Model for Imbalanced Data Set Classification," *Emerging Science Journal*, vol. 7, no. 5, pp. 1501–1516, Oct. 2023, https://doi.org/10.28991/ESJ-2023-07-05-04.

[8] V. P. K. Turlapati and M. R. Prusty, "Outlier-SMOTE: A refined oversampling technique for improved detection of COVID-19," *Intelligence-Based Medicine*, vol. 3–4, Dec. 2020, Art. no. 100023, https://doi.org/10.1016/j.ibmed.2020.100023.

[9] A. Arafa, N. El-Fishawy, M. Badawy, and M. Radad, "RN-SMOTE: Reduced Noise SMOTE based on DBSCAN for enhancing imbalanced data classification," *Journal of King Saud University - Computer and Information Sciences*, vol. 34, no. 8, pp. 5059–5074, Sep. 2022, https://doi.org/10.1016/j.jksuci.2022.06.005.

[10] A. Govindu and S. Palwe, "Early detection of Parkinson's disease using machine learning," *Procedia Computer Science*, vol. 218, pp. 249–261, 2023, https://doi.org/10.1016/j.procs.2023.01.007.

[11] D. D. Prasetya, T. Widiyaningtyas, H. Hairani, and A. Aminuddin, "Addressing Imbalance in Health Datasets: A New Method NR-Clustering SMOTE and Distance Metric Modification," *Computers, Materials & Continua*, vol. 82, no. 2, pp. 2931–2949, 2025, https://doi.org/10.32604/cmc.2024.060837.

[12] T. Widiyaningtyas, H. Hairani, D. D. Prasetya, U. Pujianto, and W. Caesarendra, "A Modified SMOTE with Noise Filtering and Manhattan Distance Metric Approach to Address Imbalanced Health Datasets," *Engineering, Technology & Applied Science Research*, vol. 15, no. 4, pp. 25452–25459, Aug. 2025, https://doi.org/10.48084/etasr.11925.

[13] M. S. Reza, R. Amin, R. Yasmin, W. Kulsum, and S. Ruhi, "Improving diabetes disease patients classification using stacking ensemble method with PIMA and local healthcare data," *Heliyon*, vol. 10, no. 2, Jan. 2024, Art. no. e24536, https://doi.org/10.1016/j.heliyon.2024.e24536.

[14] R. C. Das, M. C. Das, Md. A. Hossain, Md. A. Rahman, M. H. Hossen, and R. Hasan, "Heart Disease Detection Using ML," in *2023 IEEE 13th Annual Computing and Communication Workshop and Conference (CCWC)*, Las Vegas, NV, USA, Mar. 2023, pp. 0983–0987, https://doi.org/10.1109/CCWC57344.2023.10099294.

[15] K. M. Jha, V. Velaga, K. Routhu, G. Sadaram, S. B. Boppana, and N. Katnapally, "Evaluating the Effectiveness of Machine Learning for Heart Disease Prediction in Healthcare Sector," *Journal of Cardiobiology*, vol. 9, no. 1, 2025.

[16] T. K. N. Fariz and S. S. Basha, "Enhancing solar radiation predictions through COA optimized neural networks and PCA dimensionality reduction," *Energy Reports*, vol. 12, pp. 341–359, Dec. 2024, https://doi.org/10.1016/j.egyr.2024.06.025.

[17] A. Razzaque and D. A. Badholia, "PCA based feature extraction and MPSO based feature selection for gene expression microarray medical data classification," *Measurement: Sensors*, vol. 31, Feb. 2024, Art. no. 100945, https://doi.org/10.1016/j.measen.2023.100945.

[18] T. M. Usman, Y. K. Saheed, D. Ignace, and A. Nsang, "Diabetic retinopathy detection using principal component analysis multi-label feature extraction and classification," *International Journal of Cognitive Computing in Engineering*, vol. 4, pp. 78–88, Jun. 2023, https://doi.org/10.1016/j.ijcce.2023.02.002.

[19] D. Elreedy, A. F. Atiya, and F. Kamalov, "A theoretical distribution analysis of synthetic minority oversampling technique (SMOTE) for imbalanced learning," *Machine Learning*, vol. 113, no. 7, pp. 4903–4923, Jul. 2024, https://doi.org/10.1007/s10994-022-06296-4.

[20] M. Muntasir Nishat *et al.*, "A Comprehensive Investigation of the Performances of Different Machine Learning Classifiers with SMOTE-ENN Oversampling Technique and Hyperparameter Optimization for Imbalanced Heart Failure Dataset," *Scientific Programming*, vol. 2022, pp. 1–17, Mar. 2022, https://doi.org/10.1155/2022/3649406.

[21] X. Wang *et al.*, "Exploratory study on classification of diabetes mellitus through a combined Random Forest Classifier," *BMC Medical Informatics and Decision Making*, vol. 21, no. 1, Dec. 2021, Art. no. 105, https://doi.org/10.1186/s12911-021-01471-4.

[22] "Pima Indians Diabetes Database." Kaggle, [Online]. Available: https://www.kaggle.com/datasets/uciml/pima-indians-diabetes-database.

[23] "Heart Disease Dataset." Kaggle, [Online]. Available: https://www.kaggle.com/datasets/johnsmith88/heart-disease-dataset.

[24] Y. Han and I. Joe, "Enhancing Machine Learning Models Through PCA, SMOTE-ENN, and Stochastic Weighted Averaging," *Applied Sciences*, vol. 14, no. 21, Oct. 2024, Art. no. 9772, https://doi.org/10.3390/app14219772.

[25] R. Oktafiani, "Breast Cancer Classification with Principal Component Analysis and Smote using Random Forest Method and Support Vector Machine," *International Journal of Computer Applications*, vol. 186, no. 16, Apr. 2024.

[26] G. Douzas, F. Bacao, and F. Last, "Improving imbalanced learning through a heuristic oversampling method based on k-means and SMOTE," *Information Sciences*, vol. 465, pp. 1–20, Oct. 2018, https://doi.org/10.1016/j.ins.2018.06.056.

[27] Z. Xu, D. Shen, T. Nie, Y. Kou, N. Yin, and X. Han, "A cluster-based oversampling algorithm combining SMOTE and k-means for imbalanced medical data," *Information Sciences*, vol. 572, pp. 574–589, Sep. 2021, https://doi.org/10.1016/j.ins.2021.02.056.

[28] R. Bounab, K. Zarour, B. Guelib, and N. Khlifa, "Enhancing Medicare Fraud Detection Through Machine Learning: Addressing Class Imbalance With SMOTE-ENN," *IEEE Access*, vol. 12, pp. 54382–54396, 2024, https://doi.org/10.1109/ACCESS.2024.3385781.

[29] U. Ependi, A. F. Rochim, and A. Wibowo, "A Hybrid Sampling Approach for Improving the Classification of Imbalanced Data Using ROS and NCL Methods," *International Journal of Intelligent Engineering and Systems*, vol. 16, no. 3, pp. 345–361, Jun. 2023, https://doi.org/10.22266/ijies2023.0630.28.

[30] I. Saifudin and T. Widiyaningtyas, "Systematic Literature Review on Recommender System: Approach, Problem, Evaluation Techniques, Datasets," *IEEE Access*, vol. 12, pp. 19827–19847, 2024, https://doi.org/10.1109/ACCESS.2024.3359274.

[31] H. M. Khasanah, A. Aminuddin, F. F. Abdulloh, M. Rahardi, H. Hairani, and B. Pramudya, "Optimizing mushroom classification through machine learning and hyperparameter tuning," *Engineering and Applied Science Research*, vol. 51, 2024, Art. no. 651660, https://doi.org/10.14456/EASR.2024.61.

[32] S. Rezvani and X. Wang, "A broad review on class imbalance learning techniques," *Applied Soft Computing*, vol. 143, Aug. 2023, Art. no. 110415, https://doi.org/10.1016/j.asoc.2023.110415.

[33] M. A. Salam, A. Taher, M. Samy, and K. Mohamed, "The Effect of Different Dimensionality Reduction Techniques on Machine Learning Overfitting Problem," *International Journal of Advanced Computer Science and Applications*, vol. 12, no. 4, 2021, https://doi.org/10.14569/IJACSA.2021.0120480.

[34] M. Lamari *et al.*, "SMOTE–ENN-Based Data Sampling and Improved Dynamic Ensemble Selection for Imbalanced Medical Data Classification," in *Advances on Smart and Soft Computing*, vol. 1188, F. Saeed, T. Al-Hadhrami, F. Mohammed, and E. Mohammed, Eds. Springer Singapore, 2021, pp. 37–49.

[35] M. Lin, X. Zhu, T. Hua, X. Tang, G. Tu, and X. Chen, "Detection of Ionospheric Scintillation Based on XGBoost Model Improved by SMOTE-ENN Technique," *Remote Sensing*, vol. 13, no. 13, Jul. 2021, Art. no. 2577, https://doi.org/10.3390/rs13132577.

[36] L. G. R. Putra, K. Marzuki, and H. Hairani, "Correlation-based feature selection and Smote-Tomek Link to improve the performance of machine learning methods on cancer disease prediction," *Engineering and Applied Science Research*, vol. 50, 2023, Art. no. 577583, https://doi.org/10.14456/EASR.2023.59.

[37] M. W. Huang, C. H. Chiu, C. F. Tsai, and W. C. Lin, "On Combining Feature Selection and Over-Sampling Techniques for Breast Cancer

Prediction," *Applied Sciences*, vol. 11, no. 14, Jul. 2021, Art. no. 6574, https://doi.org/10.3390/app11146574.

[38] S. Sreejith, H. Khanna Nehemiah, and A. Kannan, "Clinical data classification using an enhanced SMOTE and chaotic evolutionary feature selection," *Computers in Biology and Medicine*, vol. 126, Nov. 2020, Art. no. 103991, https://doi.org/10.1016/j.compbiomed.2020.103991.

[39] P. Mooijman, C. Catal, B. Tekinerdogan, A. Lommen, and M. Blokland, "The effects of data balancing approaches: A case study," *Applied Soft Computing*, vol. 132, Jan. 2023, Art. no. 109853, https://doi.org/10.1016/j.asoc.2022.109853.

[40] A. Mousa, W. Mustafa, and R. B. Marqas, "A Comparative Study of Diabetes Detection Using The Pima Indian Diabetes Database," *The Journal of University of Duhok*, vol. 26, no. 2, pp. 277–288, Sep. 2023, https://doi.org/10.26682/suod.2023.26.2.24.

[41] F. Maulidina, Z. Rustam, S. Hartini, V. V. P. Wibowo, I. Wirasati, and W. Sadewo, "Feature optimization using Backward Elimination and Support Vector Machines (SVM) algorithm for diabetes classification," *Journal of Physics: Conference Series*, vol. 1821, no. 1, Mar. 2021, Art. no. 012006, https://doi.org/10.1088/1742-6596/1821/1/012006.

[42] A. Al Bataineh and S. Manacek, "MLP-PSO Hybrid Algorithm for Heart Disease Prediction," *Journal of Personalized Medicine*, vol. 12, no. 8, Jul. 2022, Art. no. 1208, https://doi.org/10.3390/jpm12081208.