Stacked Ensemble Learning for Early Heart Disease Prediction Using PCA–LDA and SMOTE

Santosh Kumar[1], Bharat Bhushan[2], Debabrata Singh[3], Narendra Kumar[4]
{*santoshkumar@soa.ac.in*[1], bharat_bhushan1989@yahoo.com[2], debabratasingh@soa.ac.in[3],
narendra298@gmail.com[4]}
Siksha 'O' Anusandhan University, Bhubaneswar, Odisha[1,3]
Sharda University, Greater Noida[2]
Amity University, Ranchi, Jharkhand[4]

Abstract. Heart disease emerges as the leading cause of death around the world. The rate of mortality crosses more than that of any disease. This study focuses on an ensemble approach to deal with the right prediction of heart diseases where Principal Component Analysis (PCA) and Linear Discriminant Analysis (LDA) feature selection techniques are applied to detect the informative and unique features. The next step deals with the class imbalance problem and uses the Synthetic Minority Oversampling Technique (SMOTE) technique to create a balanced dataset for further processing. After that the proposed model selects the most promising standard classifiers such as Random Forest (RF), Support Vector Machine (SVM), k-Nearest Neighbours (kNN), Logistic Regression (LR), Decision Tree (DT), and Gaussian Naïve Bayes (GNB). Furthermore, the proposed model, Stacking-Based Ensemble Learning Model (SELM), combines all the classifiers and uses it as a meta-learner to improve the predictive accuracy and increase the stability. Moreover, the proposed model was tested on three benchmark heart datasets, namely Framingham Heart Study dataset, CDC Heart Disease Indicators (2020), and Cleveland Heart Disease dataset. The results show the superiority in terms of predictive accuracy over the traditional standalone classifiers as well as ensemble classifiers. The SELM model is evaluated on three standard metrics such as accuracy, F1-score, and AUC where accuracy achieved is 96.8% and subsequently higher F1-score and AUC values. Finally, results illustrate the efficiency and reliability of the SELM model for heart disease prediction.

Keywords: Heart disease prediction, PCA, LDA, Feature selection, SMOTE, Classification.

1 Introduction

Heart disease emerges as one of the main causes of death worldwide where more than 18 million people die due to it. Therefore, there is a need of early screening to save the people's life. In this context traditional diagnosis does not help the patient on time due to its time-consuming medical tests and also expensive treatment for poor people. Sometimes people are also living in remote areas

IICTDS 2025, December 19-20, Chandigarh, India
Copyright © 2026 EAI
DOI 10.4108/eai.19-12-2025.2362402

where adequate medical facilities are not available. To overcome these limitations, computational methods can help the patients for early heart disease detection and diagnosis. Finally, the main advantage of computational methods is early diagnosis and affordability for every section of society.

Machine learning (ML) techniques are more prominent to detect the useful patterns from medical datasets that's helpful in disease classification. However, medical datasets usually contain many anomalies like missing values, imbalanced classes, and irrelevant features that affect the performance of the prediction model. To overcome one of the problems, PCA and LDA are applied to identify the relevant features and reduce the dimensions from the given datasets [1, 2, 3]. These relevant and informative features help in the construction of a reliable and accurate prediction system.

Class imbalance is another common issue in heart disease datasets, where healthy cases outnumber disease-positive cases. This can cause the model to favour the majority class. To address this, the SMOTE is used to generate additional minority samples and balance the dataset [4, 5, 6].

Ensemble methods have proven their efficiency in the area of the medical domain, especially helpful in improving prediction performance. The effectiveness of the ensemble method is due to the integration of multiple models and that leads to accurate and stable results. Popular methods such as bagging, boosting, voting, and stacking are generally considered during the construction of ensemble methods. Many studies suggest the usability and effectiveness of ensemble methods in heart disease prediction [7, 8, 9, 10]. Out of these four methods, stacking performs better due to the integration of outputs of multiple classifiers using a meta-learner. This approach performs better to capture the complex and informative patterns that other models fail to detect.

Finally, the stacking method is deployed for the construction of the proposed model (SELM) for heart disease prediction. At the first step, PCA and LDA feature selection methods are applied for feature reduction and SMOTE is used to handle the class imbalances for multiple heart datasets in the subsequent second step. The next steps utilise six state-of-the-art classifiers such as RF, SVM, kNN, LR, DT, and GNB as base models. The output of these classifiers is combined with stacking with a meta-learner to enhance the overall prediction ability of the proposed model. Furthermore, the SELM model is evaluated on three heart datasets that show the efficiency in cardiovascular risk assessment.

2 Literature Review

ML is popularly used for heart disease prediction due to the availability of clinical datasets from various hospitals and clinical settings. Many studies suggest the superiority of machine learning models for disease prediction over the traditional statistical methods when feature selection techniques are applied [1, 2, 7]. These studies indicate that the model performance is based on the quality of datasets. The performance of prediction accuracy may be affected because of noisy and redundant features of the respective datasets.

Feature reduction techniques play an important role in improving model performance. PCA is widely used as a feature reduction technique where useful and informative features are preserved with reduced data size [1]. Sharma et al. [11] suggested the importance of selecting relevant features that lead to improved accuracy and reduced computational time. LDA technique is also useful in improving the class separation especially in case of high-dimensional datasets [2, 12]. Many studies

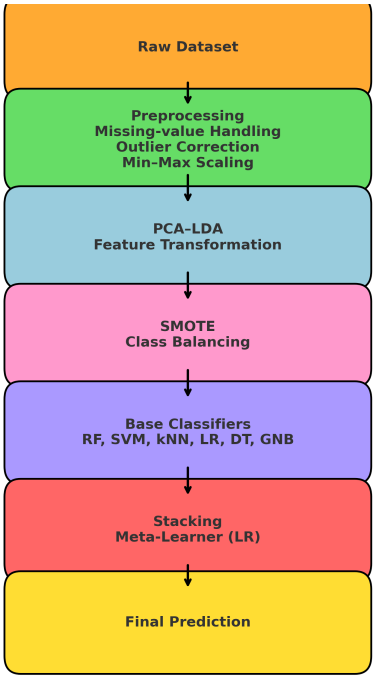


Fig. 1. Proposed SELM framework showing preprocessing, feature transformation, SMOTE balancing, and stacked ensemble construction.

also combined PCA, LDA with other feature selection techniques to reduce the feature correlation and improve the model understanding [13].

Class imbalance is another major problem in heart disease datasets. Healthy samples are usually higher than disease cases, which causes model bias toward the majority class. This reduces the model's ability to detect disease cases. Khushi et al. [14] showed that classifiers perform poorly on minority classes without proper balancing. SMOTE is a popular method used to solve this problem. Kavitha [15] showed that SMOTE improves prediction performance. Nwe et al. [16] reported that combining SMOTE with kNN improves recall and F1-score.

Ensemble learning methods are widely used to improve prediction accuracy. RF and bagging methods reduce variance and improve model stability [17, 18]. Boosting methods such as AdaBoost and Gradient Boosting also improve performance by focusing on difficult samples [7]. Voting ensemble methods combine multiple classifiers and produce better results than individual models [17].

Stacking is one of the most effective ensemble methods. It combines multiple classifiers using a meta-learner. Bhagat et al. [19] showed that stacking improves prediction accuracy compared to single models. Uplopwar et al. [20] reported that combining SMOTE with stacking further improves accuracy and AUC. Other studies also improved stacking performance using parameter optimization

[21].

However, some limitations still exist. Many studies use only one dataset, which limits general performance. Also, feature reduction, class balancing, and stacking are often used separately. Only a few studies combine all these methods in one framework [19, 20]. Therefore, this study proposes a SELM that integrates PCA, LDA, SMOTE, and stacking. The model is evaluated on three heart disease datasets.

3 Proposed Methodology

The proposed SELM is designed as a unified pipeline. It addresses common problems in clinical heart disease datasets such as missing values, noisy or redundant features, different feature scales, and class imbalance. The methodology is presented in a clear and connected manner. It is described as a continuous process rather than separate steps.

The proposed SELM is developed as a unified pipeline to address several challenges commonly encountered in clinical heart-disease datasets, including missing values, noisy or redundant attributes, inconsistent feature scales, and substantial class imbalance. The methodology is presented as a clear and connected explanation instead of separate individual steps to improve readability.

The proposed methodology starts with data preprocessing to ensure consistent input values. After preprocessing, the model is trained using the prepared dataset. In the dataset, numerous missing values were encountered, which were dealt with by the mean and median, depending on the data distribution. The missing values in categorical attributes are replaced with the most frequently occurring values. The outliers are identified using interquartile range and z-score methods, and later these values are adjusted so they do not affect the model training. This step is essential when using distance-based algorithms such as k-NN and margin-based classifiers like SVM. For better feature utilization, min-max normalization is utilized so that all the features contribute equally during the training procedures.

Dimension reduction is the next stage after applying preprocessing. In this step, dimension reduction is conducted using two-stage PCA-LDA techniques. In the first step, PCA is applied to reduce the features while keeping the most relevant information. This step is very prominent in removing the redundancy and reducing the noise. In the second step, LDA is applied for clear class separation between disease-positive and disease-negative samples. The final feature set is obtained after completion of the two-stage PCA-LDA process. These obtained features are compact and informative, which helps in reducing overfitting.

Class imbalance is a common problem in heart disease datasets. In most cases, healthy samples greatly outnumber disease-positive cases. To handle this issue, the SMOTE is used. Instead of copying existing minority samples, SMOTE generates new synthetic samples. It does this by interpolating between nearby minority-class instances. This creates a more balanced data distribution. As a result, classifiers can better learn the patterns associated with disease-positive cases.

After transformation and balancing, the dataset is used to train multiple base classifiers. These include RF, SVM, k-NN, LR, DT, and GNB. Each model brings a different strength. Tree-based models are good at learning non-linear patterns. Distance and margin-based models focus on spatial boundaries. Probabilistic models estimate class likelihoods, while linear models offer a stable

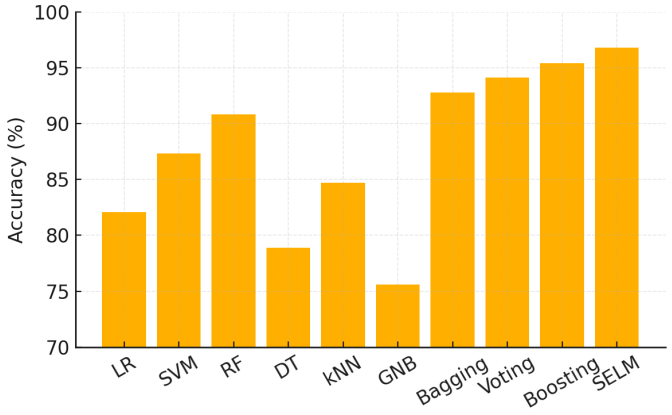


Fig. 2. Accuracy comparison of individual classifiers and ensemble models.

baseline. This variety helps the ensemble learn from different decision views.

In the final step, the stacking technique is incorporated in the proposed model. In this step, the outputs of all the individual base classifiers are accumulated into a new feature set. A LR model is used as the meta-learner. It learns how the strengths of different models complement each other and corrects inconsistent predictions. This second-level learning produces a more stable and accurate output than any single classifier. The complete SELM workflow is shown in Fig. 1.

4 Results and Discussion

This section analyses the performance of the proposed SELM framework and compare it with standalone classifiers and traditional ensemble techniques. The experiments were performed on three widely used datasets namely Clevland Heart Disease [22], Framingham Heart Study [23], and CDC Indicators (2020) [24]. All datasets were handled using the same SELM framework. The process involved data preprocessing, PCA–LDA transformation, SMOTE oversampling, training of base models, and construction of the stacked ensemble. Each dataset was divided into an 80–20% split ratio for training and testing evaluation, respectively, to maintain class balance. Model optimization for both the base classifiers and the meta-learner was performed using five-fold cross-validation. The proposed SLEM model was evaluated using standard metrics, namely accuracy, precision, recall, F1-score, and the area under the ROC curve (AUC).

4.1 Overall Performance Comparison

Table 1 shows the average accuracy of all models across the three datasets. Among the individual classifiers, Random Forest achieved the highest accuracy of 90.8%, followed by SVM and kNN. LR, DT, and GNB performed comparatively worse, which indicates that these models struggle

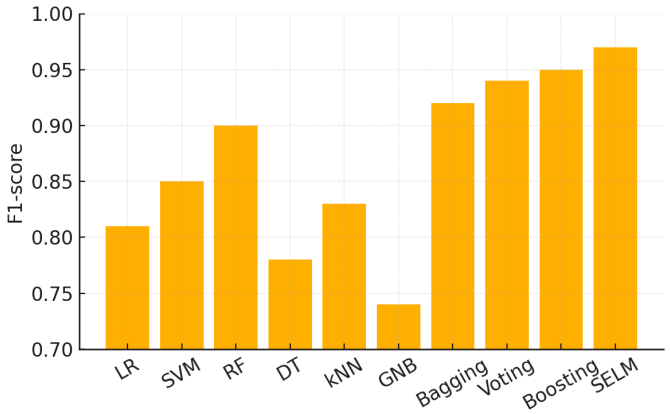


Fig. 3. F1-score comparison of competing models.

to capture complex clinical patterns. Ensemble techniques further improved performance. Bagging reached 92.8%, Voting achieved 94.1%, and Boosting performed even better at 95.4%. The proposed SELM model surpassed all baselines with an accuracy of 96.8%. This improvement of over 6% compared with the strongest individual classifier demonstrates the benefit of combining heterogeneous learning signals. Fig. 2 shows that SELM consistently outperforms competing approaches across all datasets.

Because accuracy alone may mask weak minority-class performance, F1-score and AUC were also examined. As shown in Fig. 3, SELM achieved the highest F1-score (0.97), followed by Boosting and Voting. Lower F1-scores for GNB and DT reflect higher misclassification rates. ROC curve analysis provided deeper insight into classifier behaviour. SELM achieved the highest AUC value of 0.98, outperforming Boosting (0.96) and RF (0.94). This demonstrates SELM's strong discriminative capability, which is essential for threshold-based decision-making in clinical screening.

SELM demonstrated strong generalization across all datasets. On the smaller Cleveland dataset, SELM maintained high accuracy despite the increased risk of overfitting. On the larger and more diverse Framingham dataset, SELM achieved strong accuracy and F1-scores, highlighting its adaptability. The CDC Indicators dataset, which is more imbalanced and noisier, exhibited the greatest improvement in recall and AUC when using SELM. Fig. 5 shows the effect of SMOTE in balancing class distribution. Models trained without SMOTE often produced lower recall and F1-scores for minority cases despite reasonable accuracy. PCA reduced redundancy and noise, while LDA enhanced class separability. Together with SMOTE, these transformations greatly contributed to SELM's stability and superior performance.

Bagging reduces variance by averaging predictions and Voting aggregates model outputs without learning interactions. Boosting focuses on difficult samples but may become sensitive to noisy data. SELM uses a meta-learner to learn from multiple classifiers. It learns better decision patterns. Therefore, it consistently achieves better results across all datasets.

Table 1: Accuracy comparison of individual classifiers and ensemble models.

Model	**Accuracy (%)**
Logistic Regression (LR)	82.1
Support Vector Machine (SVM)	87.3
Random Forest (RF)	90.8
Decision Tree (DT)	78.9
k-Nearest Neighbors (kNN)	84.7
Gaussian Naïve Bayes (GNB)	75.6
Bagging Ensemble	92.8
Voting Ensemble	94.1
Boosting Ensemble	95.4
SELM	**96.8**

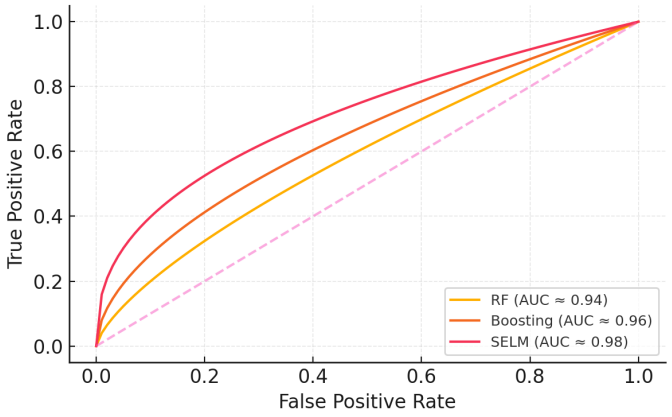


Fig. 4. ROC curves of RF, Boosting, and SELM.

Table 2: AUC comparison of high-performing models.

Model	**AUC**
Random Forest (RF)	0.94
Boosting Ensemble	0.96
SELM	**0.98**

Table 3: Dataset-wise accuracy (mean ± standard deviation) of top-performing models.

Dataset	RF	Boosting	SELM
Cleveland Heart Disease	90.2 ± 1.7	93.5 ± 1.4	96.5 ± 1.1
Framingham Heart Study	91.7 ± 2.0	94.7 ± 1.6	96.9 ± 1.3
CDC Indicators (2020)	88.9 ± 2.3	94.1 ± 1.8	97.1 ± 1.4

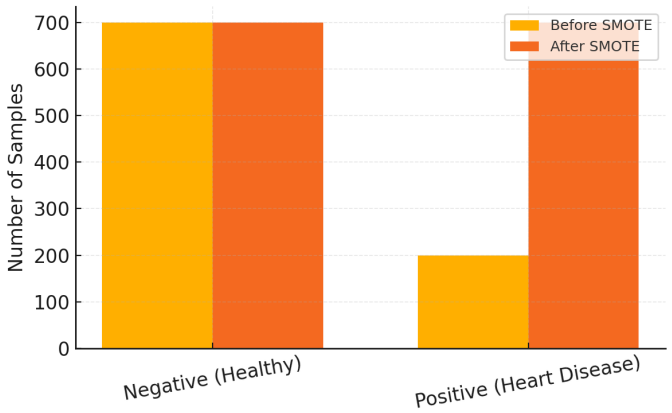


Fig. 5. Class distribution before and after SMOTE.

4.2 Discussion and Limitations

The outcome shows the performance of the model in terms of accuracy and reliability for heart disease prediction. The three major architectural components of the proposed model include PCA–LDA feature transformation, SMOTE-based class balancing, and stacked ensemble learning that work well towards enhancing the model performance. Some limitations still exist that can be overcome by the inclusion of real-world clinical datasets. Synthetic oversampling can also introduce artificial patterns, especially in high-dimensional data. In addition, stacked models demand more computational resources than simpler classifiers. SELM still achieves high accuracy, strong F1-scores, and AUC values. It appears suitable for clinical decision-support systems.

5 Conclusion and Future Work

This work proposes a unified stacking-based ensemble learning model (SELM) designed for the early prediction of heart disease. The proposed framework uses PCA–LDA for feature transformation, applies SMOTE to handle class imbalance, and adopts a heterogeneous stacked ensemble of different classifiers.

The proposed SELM model was tested on three standard benchmark datasets. The model delivered consistently higher performance than individual classifiers and standard ensemble techniques on given three datasets. The model recorded 96.8% accuracy, supported by high F1-score and AUC values, demonstrating strong predictive capability. The results confirm the robustness and reliability of SELM as a component of a heart disease clinical decision support system.

The outcomes are encouraging, but there is still room for improvement. Future research may include temporal patient data. It can track risk variations and enhance prediction quality. Furthermore, validation on real-world datasets and in multi-centric clinical environments could provide a clearer picture for practical deployment and scalability.

References

[1] Gárate-Escamila AK, El Hassani AH, Andrès E. Classification models for heart disease prediction using feature selection and PCA. Inform Med Unlocked. 2020;19:100330.

[2] Ricciardi C, Valente AS, Edmund K, Cantoni V, Green R, Fiorillo A, et al. Linear discriminant analysis and principal component analysis to predict coronary artery disease. Health Informatics J. 2020;26(3):2181-2192.

[3] Ahmed M, Hassan R, Datto S, Saleh S, Islam S, Redwan K, Mahmood S. A comparative study of machine learning models for cardiovascular risk prediction using PCA-transformed Framingham data. In: Proceedings of the 2025 International Conference on Quantum Photonics, Artificial Intelligence, and Networking (QPAIN). IEEE; 2025. p. 1-6.

[4] Aish MA, Ghafoor AA, Nasim F, Ali KI, Akhter S, Azeem S. Improving stroke prediction accuracy through machine learning and synthetic minority oversampling. J Comput Biomed Inform. 2024;7(2).

[5] Zheng H, Sherazi SWA, Lee JY. A stacking ensemble prediction model for the occurrences of major adverse cardiovascular events in patients with acute coronary syndrome on imbalanced data. IEEE Access. 2021;9:113692-113704.

[6] Brandt J, Lanzén E. A comparative review of SMOTE and ADASYN in imbalanced data classification. 2021.

[7] Alqahtani A, Alsubai S, Sha M, Vilcekova L, Javed T. Cardiovascular disease detection using ensemble learning. Comput Intell Neurosci. 2022;2022:5267498.

[8] Theerthagiri P. Predictive analysis of cardiovascular disease using gradient boosting based learning and recursive feature elimination technique. Intell Syst Appl. 2022;16:200121.

[9] Nguyen K, Lim JWY, Lee KP, Lin T, Tian J, Do TT, et al. Heart disease classification using novel heterogeneous ensemble. In: Proceedings of the 2021 IEEE EMBS International Conference on Biomedical and Health Informatics (BHI). IEEE; 2021. p. 1-4.

[10] Patro SP, Padhy N, Sah RD. An improved ensemble learning approach for the prediction of cardiovascular disease using majority voting prediction. Int J Model Ident Control. 2022;41(1-2):68-86.

[11] Sharma A, Dhanka S, Kumar A, Maini S. A comparative study of heterogeneous machine learning algorithms for arrhythmia classification using feature selection technique and multi-dimensional datasets. Eng Res Express. 2024;6(3):035209.

[12] Isnanto RR, Rashad I, Widodo CE. Classification of heart disease using linear discriminant analysis algorithm. In: E3S Web of Conferences. EDP Sciences; 2023. Vol. 448. p. 02053.

[13] Jain D, Singh V. An efficient hybrid feature selection model for dimensionality reduction. Procedia Comput Sci. 2018;132:333-341.

[14] Khushi M, Shaukat K, Alam TM, Hameed IA, Uddin S, Luo S, et al. A comparative performance analysis of data resampling methods on imbalance medical data. IEEE Access. 2021;9:109960-109975.

[15] Kavitha M. Comparative analysis of SMOTE techniques and machine learning models for imbalanced medical datasets. Integration. 2024;16:18.

[16] Nwe MM, Lynn KT. KNN-based overlapping samples filter approach for classification of imbalanced data. In: International Conference on Software Engineering Research, Management and Applications. Cham: Springer International Publishing; 2019. p. 55-73.

[17] Bashir S, Qamar U, Khan FH. A multicriteria weighted vote-based classifier ensemble for heart disease prediction. Comput Intell. 2016;32(4):615-645.

[18] Tiwari A, Chugh A, Sharma A. Ensemble framework for cardiovascular disease prediction. Comput Biol Med. 2022;146:105624.

[19] Bhagat M, Sharma A, Agarwal P. An efficient stacking-based ensemble technique for early heart attack prediction. Multimed Tools Appl. 2025;84(30):36351-36375.

[20] Uplopwar A, Vashisth R, Kushwaha A. Cardiovascular disease detection using stacking meta-learner with pipeline machine learning. Neural Comput Appl. 2025:1-14.

[21] Berikol GB, Yildiz O, Özcan İT. Diagnosis of acute coronary syndrome with a support vector machine. J Med Syst. 2016;40(4):84.

[22] Detrano, R., A. Jánosi, W. Steinbrunn, M. Pfisterer, J. Schmid, S. Sandhu, K. Guppy, S. Lee, and V. Froelicher. "Heart Disease—UCI Machine Learning Repository." American Journal of Cardiology: New York, NY, USA (1989).

[23] Dawber, Thomas R., Gilcin F. Meadors, and Felix E. Moore Jr. "Epidemiological approaches to heart disease: the Framingham Study." American Journal of Public Health and the Nations Health 41, no. 3 (1951): 279-286.

[24] Pytlak, K. "Indicators of heart disease (2022 update)." Kaggle. Available online: https://www. kaggle. com/datasets/kamilpytlak/personal-key-indicators-of-heart-disease/data (accessed on 23 October 2023) (2023).