International Journal of Academic Research
P-ISSN: 2348-7666: Impact Factor: 7.215
Homepage: https://www.ijar.org.in/
Volume-13, Issue-3, July-Sep-2026

Research Article

A Comparative Study of Machine Learning and Deep Ensemble Architectures for Cardiovascular Disease Risk Prediction on Imbalanced Health-Indicator Data

Sri Veda Aryan Pallapu[1], A. Jithin Chowdary Atluri[2]

[1] Dept. of CSE, SRM University – AP, Amaravathi, India.
[2]Dept. of CSE, SRM University – AP, Amaravathi, India.
Corresponding Author: Sri Veda Aryan pallapu.

Abstract
Cardiovascular disease (CVD) remains the single largest contributor to global mortality, yet early screening is often out of reach in resource-constrained settings where clinical testing is costly or unavailable. This paper investigates whether inexpensive, self-reported health indicators can support reliable automated CVD risk screening. Fifteen classifiers spanning classical machine learning, deep neural architectures, and two purpose-built deep ensembles—EnsCVDD-Net and BlCVDD-Net—are evaluated on the publicly available Heart Disease Health Indicators dataset. To counter the strong class skew inherent in population-level survey data, the pipeline couples point-biserial correlation-driven feature selection with Adaptive Synthetic (ADASYN) minority oversampling. Across accuracy, precision, recall, and F1-score, a soft-voting ensemble of heterogeneous base learners delivers the strongest overall performance (91.7% accuracy, 0.918 F1), outperforming every individual model as well as the deep architectures by a wide margin, while remaining light enough for real-time inference. The selected model is packaged into a browser-based screening application that returns instant risk estimates from user-entered indicators. The results suggest that, for tabular health-indicator data of this kind, carefully balanced classical ensembles remain a stronger and more deployable choice than deeper networks, and that accessible pre-clinical screening tools can be built entirely from open-source components.
Keywords: Cardiovascular disease, ensemble learning, class imbalance, ADASYN, voting classifier, deep learning, health informatics, clinical decision support.

1. Introduction

The World Health Organization attributes roughly one in three deaths worldwide to cardiovascular conditions, with ischaemic events and strokes dominating the toll. A large share of this burden is preventable: when elevated risk is flagged early, lifestyle changes and low-cost pharmacological interventions measurably improve outcomes. The bottleneck is detection. Conventional cardiac work-ups depend on specialist interpretation, laboratory testing, and imaging infrastructure—resources that are scarce precisely in the regions where CVD incidence is rising fastest.

Data-driven risk models offer a complementary path. Rather than replacing clinical diagnosis, a screening model trained on routinely collectable indicators (blood pressure history, cholesterol status, body mass index, general health self-assessment, and similar variables) can

triage individuals toward or away from formal evaluation at negligible marginal cost. Two practical obstacles, however, recur throughout the literature. First, population health surveys are heavily imbalanced—positive CVD cases are a small minority—which biases naively trained classifiers toward the majority class and suppresses sensitivity, the metric that matters most in screening. Second, published work frequently reports a single model family in isolation, making it difficult to judge whether added architectural complexity actually pays off on tabular survey data.

This work addresses both gaps through a controlled, single-pipeline comparison. The contributions are fourfold:

- A unified evaluation of fifteen classifiers—eight classical machine learning models and seven deep architectures—trained under identical preprocessing, sampling, and validation conditions on the Heart Disease Health Indicators dataset.
- A preprocessing strategy that pairs point-biserial correlation feature screening with ADASYN adaptive oversampling, and an analysis of its effect on minority-class sensitivity.
- Two custom deep ensemble designs, EnsCVDD-Net and BlCVDD-Net, which aggregate heterogeneous neural learners through ensembling and blending respectively, included to test whether deep ensembling closes the gap with classical ensembles on tabular data.
- An end-to-end deployment of the best-performing model as a lightweight web application providing real-time risk feedback, demonstrating that the full pipeline runs on commodity hardware.

The remainder of the paper is organised as follows. Section II reviews related work. Section III details the dataset, preprocessing, and model designs. Section IV describes the experimental protocol. Section V presents and discusses results, and Section VI concludes with limitations and future directions.

2. Related Work

2.1 Classical Learners for Cardiac Risk Modelling

Logistic regression has long served as the interpretable baseline for binary cardiac outcomes, and tree-based methods extend it to non-linear feature interactions. Shukur and Mijwil (**shukur2023?**) benchmarked a spectrum of supervised learners for heart disease diagnosis, while Dhanka *et al.* (**dhanka2023?**) performed a broad audit of supervised algorithms for coronary artery disease detection. Kapila *et al.* (**kapila2023?**) introduced a Quine–McCluskey-based binary classifier tailored to the same task, and Wang *et al.* (**wang2023?**) combined cloud modelling with random forests for coronary risk assessment. Gradient-boosted trees have repeatedly proven competitive on structured medical data, as demonstrated by Jawalkar *et al.* (**jawalkar2023?**) using stochastic gradient boosting for early prediction.

2.2 Deep Architectures

Deep models dominate physiological-signal tasks: Bizopoulos and Koutsouris (**bizopoulos2019?**) survey deep learning across cardiology, Liao *et al.* (**liao2021?**) fuse CNN and LSTM stages for multivariate physiological signal recognition, and Phan *et al.* (**phan2022?**) apply self-supervised learning to multi-lead ECG arrhythmia classification. Transformer-style pretraining on structured health records has also emerged, notably Med-BERT (**rasmy2021?**). For tabular risk indicators, however, evidence is more mixed—recurrent and convolutional inductive biases confer less advantage when inputs lack spatial or temporal ordering, a tension this study examines directly.

2.3 Ensembles and Imbalance Handling

Ensemble strategies consistently improve robustness in clinical prediction. Noor *et al.* (**noor2023?**) pair stacking with balancing and dimensionality reduction for heart disease,

Chaurasia and Chaurasia (**chaurasia2023?**) apply sequential feature selection within an ensemble, Hymavathi *et al.* (**hymavathi2024?**) integrate ensemble meta-features, and Shen *et al.* (**shen2023?**) propose EnsDeepDP, an ensemble deep approach for disease prediction. On the imbalance side, oversampling remains the standard countermeasure; ADASYN in particular concentrates synthetic generation on minority samples that are hardest to learn, adapting the sampling density to local classification difficulty. Optimisation-augmented networks such as the GA-tuned ANN of Arroyo and Delima (**arroyo2022?**) and data-assimilation approaches for ICU forecasting (**albers2023?**) improve specific niches but either struggle with heterogeneous survey data or target streaming clinical settings rather than population screening.

The present study differs from the above in holding the entire pipeline fixed while varying only the classifier, thereby isolating the contribution of model choice on balanced health-indicator data.

3. Materials and Methods

3.1 Dataset

Experiments use the Heart Disease Health Indicators dataset (derived from the CDC BRFSS survey and distributed via Kaggle), in which each record describes one respondent through demographic attributes, lifestyle factors, and self-reported medical history, with a binary label indicating a prior heart disease or heart attack diagnosis. Because the data are survey-based, no invasive or laboratory measurements are required at inference time—an intentional property, since the target deployment is pre-clinical screening.

3.2 Feature Screening

Redundant and weakly informative attributes inflate variance without improving discrimination. Two complementary filters were applied. First, pairwise correlation analysis over the upper triangle of the feature correlation matrix identified strongly collinear attribute pairs, from which one member was discarded. Second, point-biserial correlation—appropriate for measuring association between continuous predictors and a dichotomous outcome—ranked the remaining attributes against the CVD label, and low-association variables (including smoking status, fruit and vegetable consumption, healthcare access indicators, and several demographic fields) were removed. The retained set comprises the indicators with the strongest marginal relationship to the outcome, such as hypertension history, cholesterol status, BMI, stroke history, diabetes status, general health rating, and mobility difficulty.

3.3 Class Balancing with ADASYN

Positive cases form a small fraction of the corpus, so a classifier can achieve deceptively high accuracy by defaulting to the negative class. The Adaptive Synthetic Sampling (ADASYN) algorithm was applied to the training partition only. Unlike uniform oversampling, ADASYN weights synthetic-sample generation by the proportion of majority-class neighbours around each minority point, producing more synthetic examples in regions where the decision boundary is hardest to learn. The rebalanced data were then split into training and held-out test sets with a fixed random seed for reproducibility.

3.4 Candidate Models

3.4.1 Classical machine learning

Eight learners were trained: logistic regression, naïve Bayes, decision tree, k-nearest neighbours, support vector machine with probability calibration, random forest, XGBoost, and AdaBoost. Hyperparameters were tuned by grid search with cross-validation on the training partition.

3.4.2 Deep learning

Seven neural architectures were implemented in TensorFlow/Keras: a LeNet-style convolutional network, a GRU network, an LSTM, a bidirectional LSTM, a hybrid CNN+LSTM that feeds convolutional feature maps into a recurrent stage, and the two proposed ensembles

described next. All deep models used early stopping on validation accuracy (patience of five epochs) to limit overfitting.

3.4.3 Proposed deep ensembles

EnsCVDD-Net aggregates the calibrated outputs of multiple trained neural learners and fuses them through averaging, seeking variance reduction across architecturally diverse members. *BlCVDD-Net* instead adopts a blending scheme: base-learner predictions on a held-out fold become inputs to a shallow meta-learner that learns how to weight each member. Both designs test the hypothesis that ensembling can lift deep models to parity with classical ensembles on tabular inputs.

3.4.4 Voting classifier

The deployed classical ensemble is a soft-voting combination of logistic regression, a decision tree, and a probability-calibrated SVM. Soft voting averages predicted class probabilities rather than hard labels, which allows confident members to outweigh uncertain ones and typically yields smoother decision boundaries than majority voting.

3.5 Evaluation Protocol

All models were assessed on the same held-out test partition using accuracy, precision, recall, and F1-score, with confusion matrices inspected to verify per-class behaviour. F1-score is treated as the headline metric because it balances the cost of missed positives (dangerous in screening) against false alarms (which erode user trust and burden downstream clinical capacity).

3.6 Deployment

The selected model was serialised and wrapped in a Streamlit web application. Users enter their health indicators through a form; the backend applies the identical preprocessing transformations used in training and returns a binary risk estimate in real time. The full stack—Python, scikit-learn, imbalanced-learn, TensorFlow, and Streamlit—is open source, and inference runs comfortably on a standard laptop-class CPU, keeping the tool deployable in low-resource environments.

4. Experimental Setup

Development was carried out in Python 3.11 on Jupyter and Google Colab, with GPU acceleration used only for deep-model training. Classical models used scikit-learn, XGBoost, LightGBM, and CatBoost implementations; balancing used imbalanced-learn; deep models used TensorFlow/Keras. A consistent random seed governed the resampling and train–test split so that every classifier saw identical data.

5. Results and Discussion

5.1 Classical Model Performance

Table 1 reports the classical results. Individual learners cluster between 66% and 72% accuracy, with logistic regression the strongest single model—consistent with the near-linear separability of several retained indicators. The soft-voting ensemble, however, moves the operating point dramatically, reaching 91.7% accuracy with balanced precision (0.920) and recall (0.917). The roughly twenty-point jump over its own base learners indicates that the members err on substantially different subsets of the test data, so probability averaging cancels a large fraction of individual mistakes.

Performance of Classical Machine Learning Models

Model	Accuracy	Precision	Recall	F1
Voting Classifier	**0.917**	**0.920**	**0.917**	**0.918**
Logistic Regression	0.716	0.717	0.716	0.717
AdaBoost	0.708	0.710	0.708	0.709
SVM	0.702	0.704	0.702	0.703
XGBoost	0.695	0.696	0.695	0.695

Naïve Bayes	0.690	0.694	0.690	0.689
Decision Tree	0.674	0.683	0.674	0.673
k-NN	0.662	0.662	0.662	0.662

5.2 Deep Model Performance

Table 2 summarises the neural results. The LeNet-style CNN attains the best deep-model accuracy (0.716), matching logistic regression, while BiLSTM posts the highest recall (0.791) among all deep models—useful where sensitivity is prioritised. The proposed EnsCVDD-Net and BlCVDD-Net achieve strong precision (0.757) but at the cost of recall (0.576), yielding conservative classifiers that under-flag positives. The hybrid CNN+LSTM improves precision over its individual components without surpassing the ensembles overall.

Performance of Deep Learning Models

Model	Accuracy	Precision	Recall	F1
LeNet	0.716	0.735	0.765	0.750
GRU	0.704	0.749	0.703	0.725
BiLSTM	0.699	0.704	0.791	0.745
LSTM	0.686	0.754	0.645	0.695
CNN+LSTM	0.684	0.760	0.631	0.689
EnsCVDD-Net	0.662	0.757	0.576	0.654
BlCVDD-Net	0.662	0.757	0.576	0.654

5.3 Analysis

Three findings stand out. First, on flat tabular indicators, the inductive biases that make CNNs and recurrent networks powerful on images and sequences confer little benefit: no deep model exceeded the best linear baseline by a meaningful margin, and all trailed the classical voting ensemble by roughly twenty accuracy points. Second, imbalance handling was decisive; ablation without ADASYN sharply degraded minority-class recall across every family, confirming that sampling strategy, not architecture, was the dominant lever for sensitivity. Third, heterogeneity of base learners mattered more than depth—combining a linear model, an axis-aligned tree, and a kernel machine produced greater error decorrelation than combining several neural networks that share gradient-based training dynamics, which helps explain why the classical ensemble outperformed the deep ensembles.

From a deployment standpoint, the voting classifier is also the pragmatic winner: it trains in minutes on a CPU, requires no accelerator at inference, and its sub-second prediction latency suits an interactive web front end. These properties directly serve the accessibility goal that motivates the work.

5.4 Limitations

The evaluation rests on a single self-reported survey dataset; label noise from self-reporting and cohort-specific biases may limit generalisation to other populations. Synthetic oversampling balances the training distribution but cannot create genuinely new clinical information. Finally, the system outputs a screening signal, not a diagnosis, and any clinical use would require prospective validation and regulatory compliance.

6. Conclusion and Future Work

This paper presented a controlled comparison of fifteen classifiers for cardiovascular risk screening on imbalanced health-indicator data, unified under a single preprocessing pipeline built on point-biserial feature screening and ADASYN balancing. A soft-voting ensemble of heterogeneous classical learners proved decisively superior—91.7% accuracy and 0.918 F1—to both its individual members and to all deep architectures evaluated, including two purpose-built deep ensembles. The resulting model powers a lightweight, open-source web application delivering real-time risk estimates, illustrating a practical route to affordable pre-clinical screening.

Future work will pursue external validation on independent cohorts, calibrated probability outputs with uncertainty estimates, explainability via SHAP-style attributions to make individual predictions auditable by clinicians, and federated training to incorporate institutional data without centralising sensitive records.

REFERENCES

00 B. S. Shukur and M. M. Mijwil, "Involving machine learning techniques in heart disease diagnosis: A performance analysis," *Int. J. Electr. Comput. Eng. (IJECE)*, vol. 13, no. 2, p. 2177, Apr. 2023. R. Kapila, T. Ragunathan, S. Saleti, T. J. Lakshmi, and M. W. Ahmad, "Heart disease prediction using novel Quine McCluskey binary classifier (QMBC)," *IEEE Access*, vol. 11, pp. 64324–64347, 2023. S. Dhanka, V. K. Bhardwaj, and S. Maini, "Comprehensive analysis of supervised algorithms for coronary artery heart disease detection," *Expert Syst.*, vol. 40, no. 7, p. e13300, Aug. 2023. J. Wang, C. Rao, M. Goh, and X. Xiao, "Risk assessment of coronary heart disease based on cloud-random forest," *Artif. Intell. Rev.*, vol. 56, no. 1, pp. 203–232, Jan. 2023. A. Noor, N. Javaid, N. Alrajeh, B. Mansoor, A. Khaqan, and S. H. Bouk, "Heart disease prediction using stacking model with balancing techniques and dimensionality reduction," *IEEE Access*, vol. 11, pp. 116026–116045, 2023. A. P. Jawalkar *et al.*, "Early prediction of heart disease with data analysis using supervised learning with stochastic gradient boosting," *J. Eng. Appl. Sci.*, vol. 70, no. 1, p. 122, Dec. 2023. Y. Shen *et al.*, "EnsDeepDP: An ensemble deep learning approach for disease prediction through metagenomics," *IEEE/ACM Trans. Comput. Biol. Bioinf.*, vol. 20, no. 2, pp. 986–998, Mar. 2023. J. C. T. Arroyo and A. J. P. Delima, "An optimized neural network using genetic algorithm for cardiovascular disease prediction," *J. Adv. Inf. Technol.*, vol. 13, no. 1, 2022. D. Albers *et al.*, "Interpretable physiological forecasting in the ICU using constrained data assimilation and electronic health record data," *J. Biomed. Inform.*, vol. 145, Sep. 2023. V. Chaurasia and A. Chaurasia, "Novel method of characterization of heart disease prediction using sequential feature selection-based ensemble technique," *Biomed. Mater. Devices*, vol. 1, no. 2, pp. 932–941, Sep. 2023. K. Hymavathi *et al.*, "Advancing heart disease detection by using ensemble meta-features integration," in *Proc. IEEE Int. Conf. Comput., Power Commun. Technol. (IC2PCT)*, Feb. 2024, pp. 888–892. P. Bizopoulos and D. Koutsouris, "Deep learning in cardiology," *IEEE Rev. Biomed. Eng.*, vol. 12, pp. 168–193, 2019. J. Liao *et al.*, "Recognizing diseases with multivariate physiological signals by a DeepCNN-LSTM network," *Appl. Intell.*, vol. 51, no. 11, pp. 7933–7945, Nov. 2021. T. Phan *et al.*, "Multimodality multi-lead ECG arrhythmia classification using self-supervised learning," in *Proc. IEEE-EMBS Int. Conf. Biomed. Health Inform. (BHI)*, Sep. 2022, pp. 1–4. L. Rasmy *et al.*, "Med-BERT: Pretrained contextualized embeddings on large-scale structured electronic health records for disease prediction," *NPJ Digit. Med.*, vol. 4, no. 1, p. 86, May 2021.

Citation: Sri Veda Aryan Pallapu and A. Jithin Chowdary Atluri. 2026. "A Comparative Study of Machine Learning and Deep Ensemble Architectures for Cardiovascular Disease Risk Prediction on Imbalanced Health-Indicator Data". International Journal of Academic Research, 13(3): 22-27.