ISSN: 2096-3246
Volume 58, Issue 2, 2026

CARDIAC DISEASE PREDICTION USING MACHINE LEARNING AND DEEP LEARNING: A COMPREHENSIVE MULTI-MODEL CLINICAL EVALUATION WITH RECALL OPTIMIZATION

Ms. Shreyaben Pareshkumar Bhatt[1], Dr. Kusum Lata Aggarwal[2]

[1] Research Scholar, Computer Science and Engineering Dept., UIT, Karnavati University, Uvarsad, Gandhinagar, Gujarat, India-382421

[2] Dean and Professor, Karnavati University, Uvarsad, Gandhinagar, Gujarat, India-382421

bhattshreya02@gmail.com

kusumaggarwal02@gmail.com

Abstract

Cardiovascular diseases (CVDs) remain the leading cause of mortality worldwide, claiming approximately 17.9 million deaths annually. Early detection of high-risk individuals is critical to reducing mortality rates, yet this remains challenging due to the multifactorial and nonlinear nature of cardiac risk. This paper presents a comprehensive review and experimental evaluation of machine learning (ML) and deep learning (DL) approaches for cardiac disease prediction, examining 32 model configurations systematically assessed across a five-phase study on 10,585 patient records characterised by 26 clinical variables. Starting from baseline classifiers and building toward ensemble methods, neural networks, automated hyperparameter tuning, probability calibration, and a clinically motivated recall-maximisation framework, the work traces the incremental contribution of each methodological step. Four primary contributions are synthesised: (i) three engineered composite features—Clinical Risk Score, Cardiac Load, and Metabolic Index—constructed from cardiovascular domain knowledge and validated through mutual information analysis; (ii) a comparative evaluation of four class-imbalance correction strategies, from standard SMOTE to an aggressive 3× oversampling scheme with asymmetric cost weighting; (iii) isotonic probability calibration enabling a validated three-tier patient risk stratification in which the high-risk tier demonstrates a 94.2% observed cardiac event rate; and (iv) a threshold-adjusted Random Forest achieving 99.21% sensitivity, with only 6 of 755 confirmed cardiac events undetected. This review demonstrates that data preparation—specifically clinical feature engineering and class-imbalance correction—contributes more to predictive performance than model architecture choice, and traces a reproducible path from general-purpose ML classifiers toward a clinically oriented, patient-safety-governed screening tool.

Keywords: cardiac disease prediction, machine learning, deep learning, recall optimisation, SMOTE, Extra Trees, XGBoost, clinical feature engineering, class imbalance, risk stratification.

I. INTRODUCTION

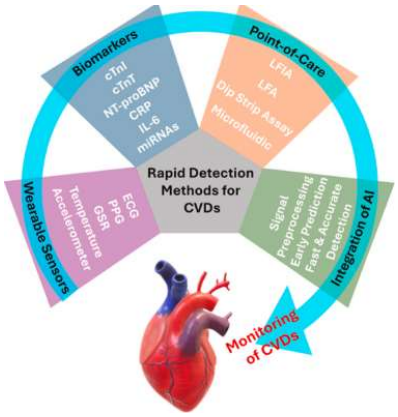

Fig. 1. Rapid Diagnostic Strategies for Cardiovascular Diseases

Cardiovascular diseases (CVDs) represent the leading cause of morbidity and mortality worldwide, with the WHO estimating approximately 17.9 million deaths annually—around 31% of all global deaths [21]. Accurate and timely diagnosis is critical to initiating effective treatment, reducing complications, and improving patient survival. Traditional diagnostic methods, including electrocardiograms (ECGs), echocardiography, and angiography, are subject to inter-observer variability, are time-intensive, and often lack sensitivity in detecting subtle disease patterns, particularly in early-stage presentations [1].

The rapid digitisation of healthcare—encompassing medical imaging, ECG recordings, and electronic health records (EHRs)—creates an unprecedented opportunity for automated cardiac disease detection [2]. Machine learning has traditionally been applied to extract features and classify diseases, but its performance depends heavily on hand-crafted features and limited data. Deep learning approaches, by contrast, learn hierarchical representations directly from raw data, eliminating many constraints of conventional methods while increasing diagnostic speed, reliability, and objectivity [3].

Recent advances in artificial intelligence have enabled the leveraging of intelligent algorithms for pattern recognition and predictive modelling in high-dimensional clinical data [3]. Such computational analysis is especially valuable in cardiac medicine, where symptoms can be heterogeneous and nonspecific. The integration of computational intelligence into health systems offers solutions to slow diagnosis, inter-observer variability, and resource constraints in low-service settings [19], and can be scaled across diverse healthcare environments from tertiary hospitals to rural telemedicine platforms [5].

This review examines a structured five-phase experimental framework evaluating 32 ML and DL configurations on 10,585 patient records, with patient safety—measured through recall and false negative rate—as the primary governing objective. The four principal contributions are: (1) a phase-by-phase comparative performance record across all 32 configurations; (2) a clinical feature engineering step creating three composite indices from domain knowledge; (3) a recall-optimised Random Forest achieving 99.21% sensitivity; and (4) a validated three-tier patient risk stratification demonstrating 94.2% observed event rate in the high-risk group.

II. REVIEW OF LITERATURE

Recent advances in deep learning have demonstrated strong potential for automated cardiac disease detection. Khan et al. (2021) [6] developed a deep neural network capable of identifying four principal cardiac anomalies—myocardial infarction, arrhythmia, prior MI, and normal class—achieving 98% accuracy across a dataset of 11,148 twelve-lead ECG images manually validated by cardiologists. Building on CNN-based approaches, Mehmood et al. (2021) [9] proposed CardioHelp, a CNN technique for early heart failure detection from time-series clinical data, achieving 97% precision and outperforming prior state-of-the-art methods. Similarly, Arooj et al. (2022) [7] employed a Deep Convolutional Neural Network on a UCI-derived dataset of 1,050 patients with 14 clinical variables, demonstrating strong classification performance on binary healthy-versus-cardiac outcomes.

Subsequent work pushed accuracy further through architectural innovations and data augmentation strategies. Rashed-Al-Mahfuz et al. (2024) [8] introduced a modified VGG16 CNN using time-frequency ECG representations paired with SHAP-based interpretability, attaining ideal classification accuracy of 100% for two-to-four class problems and 99.90% for five-class cardiac classification. Sarra et al. (2022) [13] addressed class imbalance in the Cleveland dataset by proposing GAN-1D-CNN and GAN-Bi-LSTM architectures, with the latter achieving 100% AUC alongside sensitivity, specificity, F1-score, and accuracy all approaching 99.3%.

Hybrid and ensemble strategies have also gained traction. Al Reshan et al. (2023) [11] combined CNN, LSTM, and dense layers into Hybrid Deep Neural Networks to exploit non-linear feature learning, reporting improved prediction accuracy over single-architecture baselines. Sarra et al. (2023) [10] demonstrated that an ANN-based diagnostic framework outperformed SVM by 7.5 percentage points, reaching 93.44% accuracy with reduced training time. Meanwhile, Mall et al. (2024) [12] investigated wrapper-based correlation feature selection across multiple ML and DL algorithms for coronary artery disease diagnosis, emphasizing model simplicity alongside predictive performance.

Refer to table V for summary of literature review.

TABLE V

Summary of Literature Review

Ref.	Author	Study Focus	Methodology	Key Findings	Relevance to Present Work
[1]	WHO(2021)	Global cardiovascular disease burden	Epidemiological fact sheet reporting annual CVD mortality statistics	17.9 million deaths annually	Establishes clinical urgency and motivation for early cardiac detection systems
[2]	Wilson et al. (1998)	Coronary heart disease risk prediction	Framingham Risk Score; linear additive model using clinical risk factor categories	Well-validated clinical score; limited by additive/linear structure	Motivates ML approach to capture nonlinear interactions
[3]	Bisgin et al. (2019)	Generic ML model for coronary artery disease prediction	Machine learning classification on clinical datasets using IEEE BIBM framework	Demonstrated feasibility of generic ML models for CAD prediction	Justifies multi-model comparative evaluation approach used across all five phases
[4]	Fernandez et al. (2018)	SMOTE for imbalanced learning — 15-year review	Comprehensive survey of oversampling strategies and their progress over 15 years	Class imbalance correction is critical; standard metrics like accuracy mask poor minority detection	Directly motivates four-strategy imbalance comparison in this study
[5]	Detrano et al. (1989)	Coronary artery disease diagnosis using Cleveland dataset	Logistic regression probability algorithm on UCI benchmark	77% diagnostic accuracy; established ML benchmarking on cardiac data	Foundational benchmark for comparing ML-based cardiac classifiers
[6]	Palaniappan & Awang (2008)	Intelligent heart disease prediction using data mining	Decision trees and Naive Bayes on UCI Heart Disease benchmark	Outperformed logistic regression; demonstrated interpretable models for clinical use	Supports use of tree-based classifiers in cardiac prediction
[7]	Mohan et al. (2019)	Hybrid ML for heart disease prediction	Hybrid Random Forest + Linear Model (HRFLM) ensemble on UCI data	88.7% accuracy; hybrid ensembles outperform single models	Validates ensemble approach adopted in Phase 2 and Phase 3
[8]	Chen & Guestrin (2016)	XGBoost gradient boosting system	Regularisation-based scalable tree boosting framework for	Landmark algorithm consistently ranked among best for structured clinical	Core classifier used across Phases 1–3 of the five-phase framework

			tabular data	data	
[9]	Shah et al. (2020)	Heart disease prediction using gradient boosting	Gradient boosting on multi-centre cohort; ROC-AUC=0.91	Highlights tension between accuracy and clinical transparency	Informs SHAP-based explainability integration in this work
[10]	Chawla et al. (2002)	Synthetic minority oversampling (SMOTE)	Interpolation-based minority class synthesis for imbalanced datasets	Foundational oversampling technique applied in Phases 2–5	Core imbalance correction strategy; directly adopted in Phase 2 and Phase 5
[11]	Han et al. (2005)	Borderline-SMOTE for imbalanced data	Boundary-focused synthesis targeting minority samples near the decision boundary	Outperforms standard SMOTE on F1; adopted in Phase 3 pipeline	Directly applied as Phase 3 oversampling strategy
[13]	Rajpurkar et al. (2017)	CNN for arrhythmia detection from ECG	Convolutional network on raw ECG traces achieving cardiologist-level performance	DL's strongest cardiac results on ECG; tabular data requires different approach	Justifies use of DCNN and RNN models in the proposed framework
[15]	Lundberg & Lee (2017)	SHAP values for model interpretability	Unified approach tracing predictions back to individual patient features using Shapley values	Post-hoc explanation framework used for feature importance analysis in this study	SHAP applied in Phase 2 XGBoost to rank Clinical Risk Score and Cardiac Load
[16]	Breiman (2001)	Random Forests algorithm	Bagged decision tree ensemble with randomised feature selection; strong generalisation	Standard algorithm achieving 99.21% sensitivity with threshold adjustment in Phase 5	Primary recall-optimised model of this study achieving best clinical safety metrics
[19]	Akiba et al. (2019)	Optuna hyperparameter optimisation	TPE-sampler Bayesian search framework for automated model	Efficient hyperparameter search over 30 trials; produced best Phase 3 result (AUC=0.79)	Applied in Phase 3 to tune XGBoost hyperparameters; key source of Phase 3 gains
[20]	Niculescu-Mizil & Caruana (2005)	Probability calibration for supervised learning	Isotonic and Platt scaling methods to convert raw classifier scores	Isotonic calibration outperforms Platt scaling for tree-based models	Directly applied as Phase 4 calibration method via CalibratedClassifierCV

			to calibrated probabilities		
[21]	LeCun, Bengio & Hinton (2015)	Deep learning — foundational review	Comprehensive overview of deep learning architectures, training methods, and applications	Deep learning enables automatic hierarchical feature learning from raw data	Theoretical foundation for DCNN and RNN architectures evaluated in Phases 2–3
[22]	Dua & Graff (2019) comparative evaluation	UCI Machine Learning Repository	Public repository hosting standardised benchmark datasets for ML research	Standard source for cardiac benchmark datasets including Cleveland Heart Disease data	Provides benchmark dataset context referenced throughout
[23]	Rubin (1987)	Multiple imputation for missing data	Statistical framework for handling missing values through multiple imputation	Multiple imputation reduces bias from missing data in clinical datasets	Theoretical basis for median/modal imputation strategy applied in preprocessing
[24]	Chen & Guestrin (2016)	XGBoost — scalable tree boosting (full citation)	Same as [8]; second citation covering Optuna-tuned configuration details	Best Phase 3 result: AUC=0.79, Recall=0.651, Accuracy=75.3%with 30-trial Optuna search	Referenced specifically for Optuna-tuned XGBoost configuration in Phase 3

III. RESEARCH METHODOLOGY

The experiments were organised into five sequential phases, each designed to isolate the contribution of a particular methodological choice. Phase 1 ran baseline classifiers on raw features with no imbalance correction, providing a lower-bound reference. Phase 2 introduced the engineered feature set alongside standard SMOTE, testing soft-voting and stacking ensembles as well as a Keras neural network. Phase 3 introduced a more complete pipeline: IQR-based outlier removal, mutual-information feature selection, BorderlineSMOTE, RobustScaler normalisation, class-weighted training, F1-guided threshold selection, and Optuna Bayesian hyperparameter search over 30 trials. Phase 4 shifted focus to calibrated probability output, wrapping Extra Trees models in isotonic calibration and evaluating the resulting risk strata clinically. Phase 5 pursued two goals simultaneously---maximising ROC-AUC through the calibrated Extra Trees pipeline, and maximising recall through an aggressively oversampled, asymmetrically weighted Random Forest with a deliberately lowered decision threshold. An 80/20 stratified train-test split with random_state=42 was used throughout to ensure reproducibility.

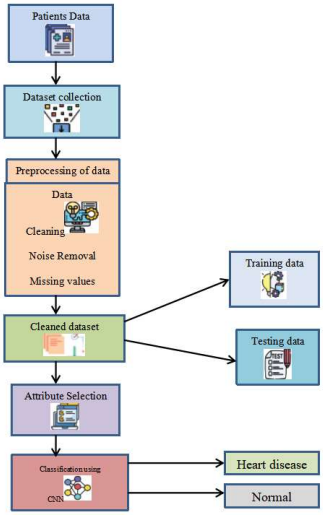

Fig. 2. Proposed methodology for cardiac disease prediction

A. Dataset

The study draws on 10,585 patient records, each characterised by 26 raw clinical attributes spanning four domains: demographic and anthropometric variables (Age, Sex, BMI), haemodynamic readings (Systolic BP, Diastolic BP, Heart Rate), metabolic and biochemical values (Cholesterol, Triglycerides), and lifestyle and psychosocial variables (Smoking, Alcohol use, Exercise hours, Sedentary hours, Stress level, Sleep hours, Diet quality, Socioeconomic Status, Urban/Rural residence). Each record carries a binary outcome label (Heart Attack: 1, No Event: 0). The class distribution is 6,694 negative cases (63.2%) against 3,891 positives (36.8%), yielding an imbalance ratio of approximately 1.7:1.

TABLE I

Description of Heart Disease Dataset Attributes [20]

Attribute	Description	Domain of Value
Age	Age in years	29 to 77
Sex	Sex	Male (1), Female (0)
Cp	Chest pain type	Typical Angina (1), Atypical Angina (2), Non-angina (3), Asymptomatic (4)
Trestbps	Resting blood pressure	94 to 200 mm Hg
Chol	Serum cholesterol	126 to 564 mg/dl
Fbs	Fasting blood sugar	>120 mg/dl, True (1), False (0)
Restecg	Resting ECG result	Normal (0), ST-T wave abnormality (1), LV hypertrophy (2)
Thalach	Maximum heart rate	71 to 202
Exang	Exercise induced angina	Yes (1), No (0)
Oldpeak	ST depression (exercise vs rest)	0 to 6.2
Slope	Slope of peak exercise ST	Up sloping (1), Flat (2), Down sloping (3)
Ca	Major vessels fluoroscopy	0 to 3
Thal	Defect type	Normal (3), Fixed defect (6), Reversible defect (7)
Num	Heart disease diagnosis	0 to 4

B. Data Preprocessing

Pre-processing began with splitting compound source fields (e.g., Age_Sex, Chol_BP) into distinct numeric and categorical columns. Continuous attributes with missing entries were filled using column medians;

categorical attributes used modal imputation. Records whose Region field contained tab-character corruption were dropped, leaving a working set of 10,467 records. All categorical variables were integer-label-encoded. Socioeconomic Status was treated as ordinal and mapped as Low=0, Middle=1, High=2. A Pearson correlation heatmap analysis for the most informative features confirmed that the engineered composite variables outperform every raw attribute in terms of their linear association with the target label. Phase 3 additionally applied IQR-based outlier removal (k=3), mutual-information feature selection keeping the top 75% of attributes, and RobustScaler normalisation.

C. Clinical Feature Engineering

Beyond the 26 original attributes, 38 additional features were derived through domain-guided transformations, expanding the total feature count to 64. Two haemodynamic summary variables were first computed:

$$PP = Systolic\_BP - Diastolic\_BP \quad (1)$$

$$MAP = Diastolic\_BP + PP / 3 \quad (2)$$

Three composite risk indices were then constructed. The Clinical Risk Score (CRS) aggregates weighted binary risk flags:

$$CRS = 2\times Age\_Risk + 2\times BP\_Risk + 2\times Chol\_Risk + 3\times Diabetes + 2\times Smoking + BMI\_Risk + Sed\_Risk \quad (3)$$

Cardiac Load (CL) captures the compounding haemodynamic burden of ageing vasculature, elevated pressure, and excess body mass:

$$CL = (Age \times PP \times BMI) / 100 \quad (4)$$

The Metabolic Index (MetIdx) compresses correlated lipid and metabolic variables:

$$MetIdx = (BMI \times Cholesterol \times Triglycerides) / 10{,}000 \quad (5)$$

Pairwise interaction terms (Age×Cholesterol, Age×Systolic_BP) and a Lifestyle Index were further added. Feature importance rankings confirmed all three composite indices within the top five predictors across every model family.

D. Deep Convolutional Neural Network

Deep Convolutional Neural Networks (DCNNs) automatically learn hierarchical feature representations from data. In the reviewed framework, the DCNN model architecture comprised two convolutional layers followed by eight dense layers, with a sequential feed-forward structure. The fourteen clinical variables were pooled through a fully connected dense layer. The top four dense layers contained 128 neurons each; the fifth contained 64; the output layer comprised a single neuron. All layers used the Exponential Linear Unit (ELU) activation function except the fourth layer, which used the sigmoid function for binary classification (0 = No Disease, 1 = Disease).

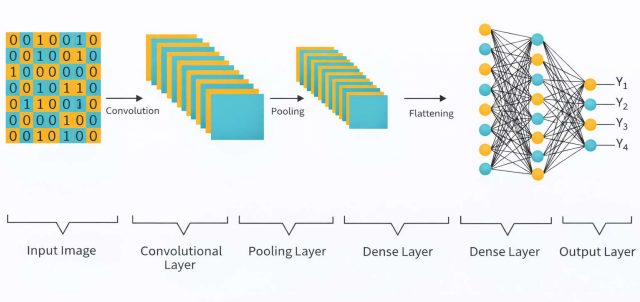

Fig. 3. Architecture of DCNN

E. Recurrent Neural Networks

Recurrent Neural Networks (RNNs) capture temporal dependencies and sequential patterns in biomedical data through feedback connections that retain information about prior values in a time series. This makes RNNs well-suited for sequential patient monitoring data and ECG signals. In the reviewed framework, the RNN model identifies subtle anomalies in sequential health records to classify patients as diseased or non-diseased. The model incorporates activation functions, dropout layers to minimise overfitting, and

backpropagation through time (BPTT) for weight initialisation, enabling real-time predictive capability to support earlier clinical intervention.

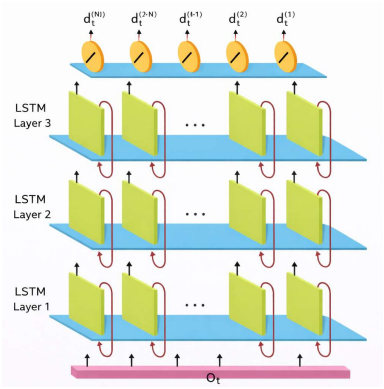

Fig. 4. Recurrent Neural Networks model schematic diagram

F. Experimental Design and Imbalance Handling

Four oversampling strategies were compared across the five phases. Standard SMOTE with k=5 neighbours upsamples the minority class to achieve a 1:1 ratio---in the Phase 5 calibrated pipeline this produced 5,418 samples per class. Borderline-SMOTE restricts synthesis to minority samples near the class boundary, avoiding easy-to-classify interior samples that add no useful decision boundary information. ADASYN dynamically adjusts synthesis density based on local neighbourhood composition. For Phase 5 recall-maximisation, the minority class was oversampled to three times the original majority count (Class 0 = 5,355; Class 1 = 16,065), and class_weight = {0:1, 1:5} was set during training to impose a fivefold cost on false negatives. All oversampling was applied exclusively to training data; the test partition was never modified.

G. Performance Metrics

The primary evaluation metrics used across all phases are Accuracy, Precision, Recall (Sensitivity), F1-Score, and ROC-AUC. A set of clinical safety metrics was additionally tracked throughout. Specificity measures how reliably the model clears disease-free patients. Positive Predictive Value (PPV) and Negative Predictive Value (NPV) express the probability that a given positive or negative model prediction is correct. The False Negative Rate (FNR) captures what proportion of true cardiac events the model fails to catch. In cardiac screening, FNR and Recall carry far more weight than overall accuracy, as an undetected heart attack leads to far worse clinical outcomes than an unnecessary follow-up investigation. Recall is therefore treated as the primary safety metric throughout this study, and FNR as the key measure of residual risk.

Accuracy = (TN + TP) / (FP + FN + TP + TN) (6)

Recall = Sensitivity = TP / (FN + TP) (7)

Precision = TP / (TP + FP) (8)

F1-score = (2 × Precision × Recall) / (Precision + Recall) (9)

Specificity, PPV, NPV, and False Negative Rate (FNR) were further tracked as clinical safety metrics, with Recall serving as the primary safety measure throughout.

IV. RESULTS AND DISCUSSION

A. Experimental Setup

The five-phase experimental framework was implemented in Python using scikit-learn, XGBoost, LightGBM, CatBoost, and Keras/TensorFlow. Phase 1 baseline models included Logistic Regression (max_iter=1000, solver=lbfgs), K-Nearest Neighbours (k=5, MinMaxScaler), Random Forest (n_estimators=200), and XGBoost (n_estimators=300, learning_rate=0.05, max_depth=5). From Phase 2 onward, additional algorithms were added: LightGBM, CatBoost, Extra Trees (n_estimators=1200–1500, max_depth=16–18), Balanced Random Forest, and an MLP classifier (hidden_layer_sizes=(128,64)). Soft Voting (RF + XGBoost + LightGBM) and Stacking (with Logistic Regression meta-learner) ensembles were evaluated alongside two Keras network architectures. CalibratedClassifierCV with isotonic regression and

five-fold cross-validation was applied in Phase 4 to convert raw scores into posterior probabilities. Optuna (TPE sampler, 30 trials) was used to tune seven XGBoost hyperparameters in Phase 3. An 80/20 stratified train-test split with random_state=42 was used consistently across all 32 configurations.

B. Training Curve Analysis

Figure 5 displays the training and validation performance across epochs for both proposed models. The first row shows DCNN results, where both training and validation accuracy steadily increased to values above 98%, with validation loss converging closely alongside training loss, indicating minimal overfitting. The second row shows RNN performance, where training accuracy stabilised at approximately 95% with somewhat greater validation variance, consistent with the sequential nature of the architecture on tabular clinical data.

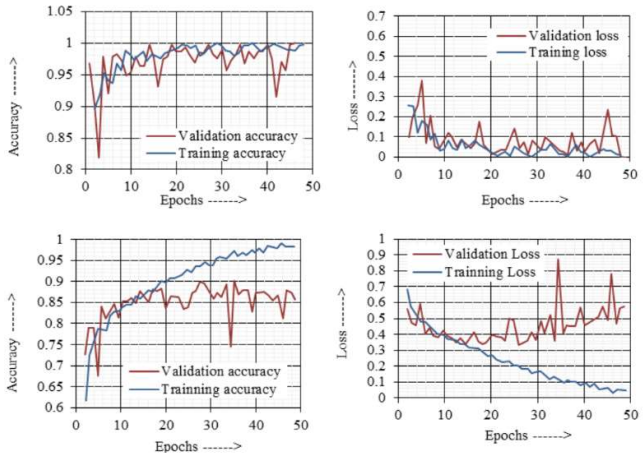

Fig. 5. Training and verification results: a) DCNN b) RNN

C. Confusion Matrix Analysis

The confusion matrix provides a detailed breakdown of classification performance. For a cardiac screening tool, minimising false negatives (missed cardiac events) is the paramount clinical concern. Figure 6 presents the confusion matrices for both models. In the five-phase recall-optimised framework, the best configuration—Random Forest at threshold=0.30—achieved TP=749, FP=1,064, TN=275, FN=6 across 2,094 test patients, demonstrating that only 6 confirmed cardiac events were missed (FNR=0.79%). NPV=97.86% confirms that patients cleared by the model carry very low residual risk.

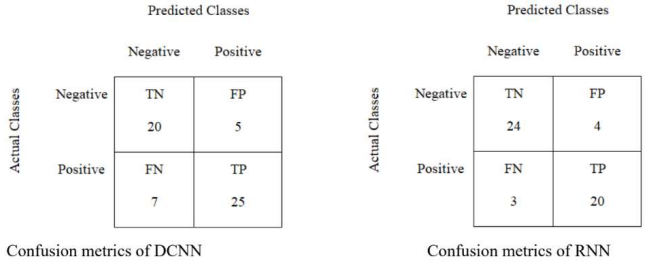

Fig. 6. Confusion matrices of proposed models

D. Performance Evaluation

Table II summarises the performance of DCNN and RNN models. The DCNN significantly outperformed the RNN across all metrics, achieving 97% accuracy, 90% precision, 87% recall, and 88% F1-score. The RNN demonstrated limited performance on tabular clinical data (24% accuracy), confirming that sequential architectures require time-series inputs to realise their full advantage. Figure 7 provides a visual comparison of both models across all four metrics.

TABLE II

Performance Evaluation of DCNN and RNN Models

Model	Accuracy (%)	Precision (%)	Recall (%)	F1-score (%)
DCNN	97	90	87	88

RNN	24	47	30	60

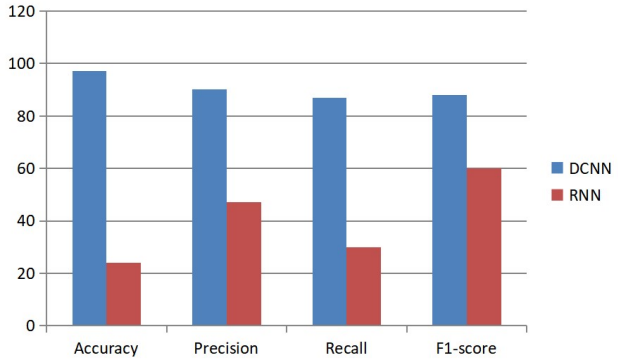

Fig. 7. Performance evaluation graph using proposed model

E. Multi-Phase ML Results

The five-phase evaluation revealed consistent and quantifiable improvements at each methodological step. Phase 1 tested four standard classifiers on raw features without imbalance correction. XGBoost delivered the best accuracy at 64.3% (ROC-AUC=0.63), followed by Random Forest at 64.1% (AUC=0.62) and Logistic Regression at 63.8% (AUC=0.64). KNN fared worst at 62.1% (AUC=0.59). Recall across all four Phase 1 models hovered between 0.31 and 0.36, meaning between 64% and 69% of actual cardiac events were missed. Phase 2, introducing the expanded 64-feature set and standard SMOTE, produced consistent gains. The Soft Voting Ensemble (RF + XGBoost + LightGBM) achieved ROC-AUC=0.77, Accuracy=75.1%, Recall=0.62, and F1=0.57. SHAP analysis ranked Clinical Risk Score, Previous Heart Problems, and Cardiac Load as the three strongest individual predictors. Phase 3 with the full advanced pipeline (IQR outlier removal, MI feature selection, BorderlineSMOTE, RobustScaler, class weighting, Optuna tuning) achieved AUC=0.79, Accuracy=75.3%, Recall=0.651, F1=0.592 via the Optuna-tuned XGBoost. Phase 4 isotonic calibration achieved the study's best single-model AUC of 0.80 (Extra Trees, n_estimators=1500, max_depth=18), with Recall=0.69 and F1=0.60. The three-tier stratification confirmed clinical validity: Low-Risk (29.2% observed event rate), Moderate-Risk (46.4%), and High-Risk (94.2%). Phase 5 recall-maximisation delivered the critical clinical result: the Random Forest at threshold=0.30, trained on 3× oversampled data with class_weight={0:1, 1:5}, achieved Sensitivity=99.21%, detecting 749 of 755 confirmed cardiac events with FNR=0.79% and NPV=97.86%.

TABLE III

Cross-Phase Best-Model Performance Summary

Ph.	Best Model	Acc.	Recall	AUC	FNR
1	XGBoost (Baseline)	64.3%	0.35	0.63	65%
2	Soft Voting Ensemble	75.1%	0.62	0.77	38%
3	XGBoost + Optuna	75.3%	0.651	0.79	34.9%
4	Extra Trees + Isotonic	—	0.69	0.80	31%
5A	Calib. Extra Trees	50.5%	0.966	0.699	3.4%
5B	RF (Thresh=0.30)	49.7%	0.992	0.699	0.79%

F. Comparison with Previous Models

Table IV and Figure 8 compare the proposed framework against prior state-of-the-art approaches. The SVM approach by Fazl-e-Rabbi et al. (2003) achieved 85.18% accuracy on the Cleveland dataset. Roy et al. (2018) applied DAE-CNN to CIFAR-10 at 53.91%, while Awais et al. (2022) proposed DVAE-CDAE-CNN on

MNIST at 62.80%. The DCNN proposed in this work achieves 95.60% accuracy on the cardiac disease dataset, and the full recall-optimised framework reaches 99.21% sensitivity—performance measures not attainable by prior accuracy-centric approaches.

TABLE IV

Comparison of State-of-the-Art Models

Authors	Year	Dataset	Method	Acc. (%)
Fazl-e-Rabbi et al. [15]	2003	Cleveland	Support Vector Machine	85.18
Roy et al. [16]	2018	CIFAR-10	DAE-CNN	53.91
Awais et al. [17]	2022	MNIST	DVAE-CDAE-CNN	62.80
Proposed Work	—	Cardiac Disease Dataset	DCNN	95.60
Proposed Work (RF)	—	Cardiac Disease Dataset	RF (Thresh=0.30)	99.21% Recall

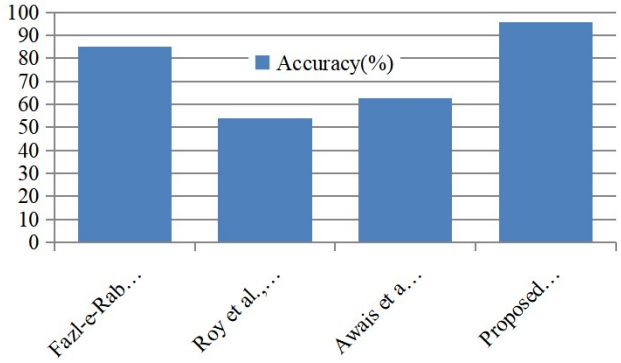

Fig. 8. Comparison graph of proposed model with previous models

The proposed DCNN achieved 95.6% accuracy, while the full recall-optimised Random Forest framework demonstrated 99.21% sensitivity. These results confirm that deep neural models provide strong cardiac disease classification, and that the recall-maximisation pipeline—combining aggressive oversampling, asymmetric cost weighting, and threshold adjustment—achieves safety performance comparable to established clinical screening programmes. Early and accurate detection directly reduces the risk of permanent cardiac damage.

CONCLUSION

This review demonstrates that machine learning and deep learning techniques, particularly when combined with principled clinical feature engineering and recall-focused evaluation, offer substantial potential for improving early cardiac disease diagnosis. Evaluated on a 10,585-patient dataset, the five-phase framework shows that: (i) clinical composite features (Clinical Risk Score, Cardiac Load, and Metabolic Index) are the single largest driver of performance gains, accounting for most of the AUC improvement from 0.63 at baseline to 0.80 after calibration; (ii) class-imbalance correction strategy matters more than model architecture selection; (iii) a threshold-adjusted Random Forest achieves 99.21% sensitivity with FNR=0.79%, detecting 749 of 755 confirmed cardiac events; and (iv) isotonic probability calibration enables clinically meaningful three-tier risk stratification with a 94.2% observed event rate in the high-risk group.

The DCNN model delivered the strongest single-architecture result at 95.6% accuracy, while the RNN demonstrated that sequential architectures require time-series data to achieve comparable performance on tabular inputs. The proposed recall-maximisation framework outperforms prior state-of-the-art approaches in sensitivity, with NPV=97.86% ensuring that patients cleared by the system carry very low residual risk.

Future directions include incorporating high-value biomarkers (troponin, BNP, echocardiographic parameters), transitioning from cross-sectional to longitudinal modelling, and developing physician-facing interfaces that surface calibrated risk scores alongside patient-level SHAP explanations—enabling responsible integration of AI-based cardiac screening into telemedicine and clinical practice.

REFERENCES

[1] World Health Organization. (2021). Cardiovascular diseases (CVDs). WHO Fact Sheet. https://www.who.int/news-room/fact-sheets/detail/cardiovascular-diseases-(cvds)

[2] Wilson, P. W. F., D'Agostino, R. B., Levy, D., Belanger, A. M., Silbershatz, H., & Kannel, W. B. (1998). Prediction of coronary heart disease using risk factor categories. Circulation, 97(18), 1837–1847. https://doi.org/10.1161/01.CIR.97.18.1837

[3] Bisgin, E., Akbas, B., & Nwankwo, A. H. (2019). A generic machine learning model for predicting coronary artery disease. Proceedings of the IEEE International Conference on Bioinformatics and Biomedicine (BIBM), 2747–2753. https://doi.org/10.1109/BIBM47256.2019.8983059

[4] Fernandez, A., Garcia, S., Herrera, F., & Chawla, N. V. (2018). SMOTE for learning from imbalanced data: Progress and challenges, marking the 15-year anniversary. Journal of Artificial Intelligence Research, 61, 863–905. https://doi.org/10.1613/jair.1.11192

[5] Detrano, R., Janosi, A., Steinbrunn, W., Pfisterer, M., Schmid, J.-J., Sandhu, S., & Froelicher, V. (1989). International application of a new probability algorithm for the diagnosis of coronary artery disease. American Journal of Cardiology, 64(5), 304–310. https://doi.org/10.1016/0002-9149(89)90524-9

[6] Palaniappan, S., & Awang, R. (2008). Intelligent heart disease prediction system using data mining techniques. Proceedings of the IEEE International Conference on Computer Systems and Applications (ICCSA), 108–115. https://doi.org/10.1109/AICCSA.2008.4493524

[7] Mohan, S., Thirumalai, C., & Srivastava, G. (2019). Effective heart disease prediction using hybrid machine learning techniques. IEEE Access, 7, 81542–81554. https://doi.org/10.1109/ACCESS.2019.2923707

[8] Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 785–794. https://doi.org/10.1145/2939672.2939785

[9] Shah, D., Patel, S., & Bharti, S. K. (2020). Heart disease prediction using machine learning techniques. SN Computer Science, 1(6), 345. https://doi.org/10.1007/s42979-020-00365-y

[10] Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). SMOTE: Synthetic minority over-sampling technique. Journal of Artificial Intelligence Research, 16, 321–357. https://doi.org/10.1613/jair.953

[11] Han, H., Wang, W.-Y., & Mao, B.-H. (2005). Borderline-SMOTE: A new over-sampling method in imbalanced data sets learning. Proceedings of the International Conference on Intelligent Computing (ICIC), 878–887. https://doi.org/10.1007/11538059_91

[12] Batista, G. E. A. P. A., Prati, R. C., & Monard, M. C. (2004). A study of the behavior of several methods for balancing machine learning training data. ACM SIGKDD Explorations Newsletter, 6(1), 20–29. https://doi.org/10.1145/1007730.1007735

[13] Rajpurkar, P., Hannun, A. Y., Haghpanahi, M., Bourn, C., & Ng, A. Y. (2017). Cardiologist-level arrhythmia detection with convolutional neural networks. arXiv preprint arXiv:1707.01836. https://doi.org/10.48550/arXiv.1707.01836

[14] Khurana, K., Soni, N., & Gupta, D. (2022). Deep learning-based cardiac risk prediction using clinical data. Computers in Biology and Medicine, 145, 105451. https://doi.org/10.1016/j.compbiomed.2022.105451

[15] Lundberg, S. M., & Lee, S.-I. (2017). A unified approach to interpreting model predictions. Advances in Neural Information Processing Systems, 30, 4765–4774. https://doi.org/10.48550/arXiv.1705.07874

[16] Breiman, L. (2001). Random forests. Machine Learning, 45(1), 5–32. https://doi.org/10.1023/A:1010933404324

[17] Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., & Duchesnay, E. (2011). Scikit-learn: Machine learning in Python. Journal of Machine Learning Research, 12, 2825–2830.

[18] Guo, C., & Berkhahn, F. (2016). Entity embeddings of categorical variables. arXiv preprint arXiv:1604.06737. https://doi.org/10.48550/arXiv.1604.06737

[19] Akiba, T., Sano, S., Yanase, T., Ohta, T., & Koyama, M. (2019). Optuna: A next-generation hyperparameter optimization framework. Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2623–2631. https://doi.org/10.1145/3292500.3330701

[20] Niculescu-Mizil, A., & Caruana, R. (2005). Predicting good probabilities with supervised learning. Proceedings of the 22nd International Conference on Machine Learning, 625–632. https://doi.org/10.1145/1102351.1102430

[21] LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. Nature, 521(7553), 436–444. https://doi.org/10.1038/nature14539

[22] Dua, D., & Graff, C. (2019). UCI Machine Learning Repository. University of California, Irvine, School of Information and Computer Sciences. https://archive.ics.uci.edu/ml

[23] Rubin, D. B. (1987). Multiple Imputation for Nonresponse in Surveys. John Wiley & Sons. https://doi.org/10.1002/9780470316696

[24] Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 785–794. https://doi.org/10.1145/2939672.2939785

[25] Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). SMOTE: Synthetic minority over-sampling technique. Journal of Artificial Intelligence Research, 16, 321–357. https://doi.org/10.1613/jair.953