

Cardio Fusion: A Dual-Stream Regularized Fusion Architecture for Generalizable Cardiovascular Disease Prediction Under Dataset Heterogeneity

K. Padmini

padmini24u@gmail.com

Vikrant University

Nilesh V. Ingale²

Vikrant University

V. Aruna³

sreenidhi Institute of Science and Technology

Research Article

Keywords: Heartdiseaseprediction, Hybriddeeplearning, Optimized featureselection, Multi-datasetbenchmarking, Regularized neural network, Decision fusion, Clinical decision support

Posted Date: September 11th, 2026

DOI: <https://doi.org/10.21203/rs.3.rs-10995403/v1>

License:   This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

Additional Declarations:

No competing interests reported.

Figures 1 is not available with this version

Cardio Fusion: A Dual-Stream Regularized Fusion Architecture for Generalizable Cardiovascular Disease Prediction Under Dataset Heterogeneity.

K.Padmini^{1,4}, Dr. Nilesh V. Ingale², Dr. V. Aruna³

¹Research Scholar, School of Computer Science, Vikrant University, Gwalior, M.P.

padmini24u@gmail.com

²Research Supervisor, School of Computer Science, Vikrant University, Gwalior, M.P.

nileshingale@gmail.com

³Co-Supervisor, Department of CSE, Sreenidhi Institute of Science and Technology, Ghatkesar, Hyderabad, India,

aruna.v@sreenidhi.edu.in

⁴Asst. Professor, Department of CSE, Sreenidhi Institute of Science and Technology, Ghatkesar, Hyderabad, India,

padmini.k@sreenidhi.edu.in

ABSTRACT

Computational risk-prediction systems have become an indispensable adjunct to conventional clinical diagnosis, yet the field continues to be constrained by a persistent methodological obstacle: benchmark datasets remain small, heterogeneous in feature composition, and frequently outdated, while single-classifier models exhibit limited generalizability across diverse patient populations. To address this limitation, the present study proposes a dual-branch hybrid deep learning architecture comprising (i) a tabular branch, in which an L1/L2-regularized deep neural network is ensemble with Random Forest, Support Vector Machine, and K-Nearest Neighbour classifiers over a feature subset optimized through Neighbourhood Component Analysis [13], and (ii) an imaging/ECG branch constructed upon ResNet-50 [14], DenseNet-121 [15], and EfficientNet-B0 [16], with the outputs of both branches integrated via a weighted soft-voting meta-learner. The proposed framework is systematically benchmarked across six contemporary datasets, spanning the canonical Cleveland and Statlog repositories, the 920-instance four-hospital combined UCI dataset that has gained prominence in recent (2024–2025) federated learning and interpretability research, a merged 1,190-record UCI–Kaggle cohort, the feature-rich Z-Alizadeh Sanicohort [18], and the longitudinal Framingham Heart Study cohort [19]. On the standard Cleveland benchmark, the proposed regularized hybrid classifier attains an accuracy of 96.5%; on a second, independently re-implemented empirical dataset — the 68,612-record Kaggle cardiovascular cohort (cardio_train.csv) — the fused model achieves an accuracy of 73.95% and an AUC-ROC of 80.57% (95% CI: 79.96–81.14), a result attributable to the dataset's exclusively lifestyle- and vitals-based feature composition rather than to any deficiency in the proposed architecture. Across the broader multi-dataset benchmark, the framework demonstrates competitive performance relative to recent state-of-the-art systems, including stacked TabNet–XGBoost ensembles (95.20% accuracy, 91.92% F1-score) [2] and applied artificial-intelligence pipelines reporting accuracies of up to 99.48% on the Statlog cohort [4]. Collectively, these findings substantiate the premise that the integration of optimized feature selection, architectural hybridization, and explicit regularization yields a classifier that is concurrently more accurate and more robust across heterogeneous clinical data sources than conventional single-algorithm baselines, thereby reinforcing its viability as a deployable clinical decision-support tool.

Keywords: Heart disease prediction; Hybrid deep learning; Optimized feature selection; Multi-dataset benchmarking; Regularized neural network; Decision fusion; Clinical decision support

1. Introduction

Cardiovascular disease (CVD) is responsible for an estimated 17.9 million deaths annually and remains the foremost cause of mortality worldwide [1]. Its aetiology is governed by an interacting constellation of

demographic, clinical, and behavioural determinants — including age, hypertension, dyslipidaemia, diabetes mellitus, obesity, and tobacco use — while its early clinical manifestations frequently overlap with those of other conditions, thereby complicating timely and accurate diagnosis. Although clinical examination, electrocardiography (ECG), and laboratory investigation continue to constitute the diagnostic gold standard, these modalities are resource-intensive and contingent upon specialist availability, a constraint that has motivated sustained investigation into machine learning (ML)-based decision-support systems as a scalable complement to conventional practice.

Two structural impediments have historically constrained the clinical translation of ML-based heart disease prediction. The first concerns feature quality: clinical datasets characteristically comprise a heterogeneous mixture of strongly predictive and weakly informative attributes, and the indiscriminate inclusion of all available features increases dimensionality, computational overhead, and susceptibility to overfitting — a risk that is especially pronounced in the small benchmark cohorts (approximately 270–920 instances) that predominate in the literature. The second concerns dataset fragmentation: the field has historically concentrated its empirical efforts on a single, decades-old cohort — the 1988 Cleveland dataset [10] — raising legitimate concerns regarding generalizability to the more diverse, larger, and more recently curated cohorts now available, including the four-hospital combined UCI dataset [5], merged UCI–Kaggle cohorts, and extensive longitudinal cohorts such as the Framingham Heart Study [19].

The present study addresses both impediments in a unified framework. First, an optimized feature-selection stage grounded in Neighbourhood Component Analysis (NCA) [13] and genetic-algorithm-guided search is employed to identify the clinically dominant attribute subset for each constituent dataset, in preference to a fixed, hand-curated feature list. Second, a dual-branch hybrid architecture is proposed, comprising a tabular branch that ensembles a regularized deep neural network with classical classifiers, and an imaging/ECG branch constructed upon modern convolutional backbones, the two being integrated via a weighted decision-fusion layer. Third, rather than confining evaluation to a single legacy cohort, the proposed framework is benchmarked across six contemporary datasets representative of the breadth of cohorts now employed in 2024–2025 heart disease prediction research and, as a substantive extension introduced in this revision, is empirically re-implemented on a second, independent dataset (`cardio_train.csv`) rather than relying exclusively on literature-cited comparators. The principal contributions of this paper may be summarized as follows: (i) a unified hybrid architecture, both diagrammatically specified and implemented, accommodating structured clinical attributes and ECG/imaging data alike; (ii) a comparative survey and tabulation of the datasets currently employed in the field, encompassing their size, feature composition, and provenance; (iii) empirical results obtained on two independently re-implemented datasets (Cleveland and `cardio_train.csv`), literature-benchmarked results across four additional datasets, and a multi-metric comparison against recent (2024–2025) state-of-the-art hybrid and ensemble models; and (iv) a fully documented and reproducible training configuration for the second empirical dataset, thereby resolving the hyperparameter-transparency deficiency identified in earlier drafts of this work.

2. Related Work

Early ML-based cardiac risk studies established the feasibility of classifiers such as CART, SVM, and Random Forest on small clinical cohorts, with reported accuracies in the range of 85–95% once optimized feature selection — via Genetic Algorithms, Particle Swarm Optimization, or Information Gain — was introduced as a precursor to classification. On the Cleveland benchmark specifically, reported single-classifier accuracies include Random Forest (89.2%), Decision Tree (89.1%), Artificial Neural Network (92.7%), and SVM (88%), while hybrid genetic-algorithm–neural-network and PCA-based pipelines have reported accuracies of 92–95.2%.

More recent scholarship has coalesced around two trends upon which the present study directly builds. The first is architectural hybridization employing modern deep learning components. A 2025 study proposed a stacked ensemble combining TabNet — a deep tabular architecture capable of identifying salient clinical features — with XGBoost, the two integrated via a logistic-regression or SVM meta-learner; evaluated on combined Kaggle and UCI CVD cohorts, this TabNet–XGBoost–LR ensemble achieved 95.20% accuracy and a 91.92% F1-score, outperforming the SAINT transformer baseline by approximately three percentage points [2]. A 2024 study published in ESC Heart Failure proposed an applied artificial-intelligence framework evaluated across five distinct cardiovascular cohorts, reporting accuracies of 99.48% on the Statlog Heart dataset, 93.90% on the UCI

Heart dataset, 96.25% on a stroke dataset, 86.00% on the Framinghamdataset, and 78.36% on a coronary heart diseasedataset—a findingthat explicitlydemonstratesthedataset-dependenceofreportedaccuracyandcautions against the interpretive limitations of single-dataset benchmarking [4]. Complementing this, a 2023 Scientific Reports study demonstrated that Learning Vector Quantization achieved 98.7% accuracy on the UCI repository, surpassing both NaïveBayes (94.78%) and Radial Basis Function networks (90.78%) on the identical cohort [6]. Two further recent hybrid systems reinforce this trajectory beyond the present benchmark set: HyCARD-Net integrates CNN, LSTM, KNN, and XGBoost components through ensemble voting, attaining 82.30–97.10% accuracy across two Kaggle cohorts [20], while MDensNet201-IDRSRNet couples a modified DenseNet201 feature extractor with an improved deep residual shrinkage network, attaining 99.12% accuracy on pooled UCI cardiacdata [22]— botharchitecturallycomparableto theDenseNet-121 component oftheimaging/ECGbranch proposed herein.

The second trend is a discernible shift toward larger and more heterogeneous benchmark cohorts. A 2024 study benchmarked centralized and federated learning algorithms, in conjunction with Shapley-value-based interpretability analysis, on the 920-instance combined UCI dataset, which aggregates patient records from four constituent sites — Cleveland, Hungary, Switzerland, and the VA Long Beach — reflecting a growing disciplinary interest in dataset diversity in preference to sole reliance on the original 303-instance Cleveland subset [5]. Correspondingly, a 2025 BMC Cardiovascular Disorders study consolidated UCI and Kaggle sources into a 1,190-record, twelve-feature cohort and benchmarked fifteen ML algorithms thereon [7], while a 2024 studyassembleda hospital-andonline-sourced"BIG-Dataset"(exceeding1,700hospitaland1,200onlinerecords across fifteen attributes) and combined it with the original cohort to yield a larger, multi-class "CD" (combined) dataset spanning twenty-seven disease categories and nineteen associated symptoms [8]. Independently, multilayer ANN models evaluated on a 1,236-patient UCI cohort (2024) and a substantially larger 13,981-record clinical cohort (2025) [3] each reported accuracies of approximately 83.4%, with comparative benchmarking in thelatter studyrevealingthat classicalRandomForest andSVM models outperformedthe ANN — a findingthat underscores the principle that increased dataset scale alone does not guarantee superior deep-learningperformance in the absence of corresponding architectural and feature-selection refinement. Explainability constitutes an increasingly prevalent complement to such hybrid systems: a 2025 Scientific Reports study integrated hybrid ensemble learning with SHAP, t-SNE, and PCA projections to interpret cardiovascular risk predictions [21], a direction adopted in the present study as a concrete avenue for future work (Section 7).

Consideredcollectively, theextant literaturesubstantiatesthetwocentraldesignchoicesunderpinningthepresent study: explicit, optimization-driven feature selection antecedent to classification, and architectural hybridization spanning both tabular and signal/imaging modalities, evaluated not upon a solitary legacy cohort but across the diversifiedset of datasets nowactive inthefield — and, inthis revision, substantiatedbya secondindependently re-implemented empirical dataset rather than literature citation alone.

3. Datasets

Table1summarizesthesix datasetsusedtobenchmarktheproposedframework, spanningtheclassicalCleveland and Statlog repositories through to the larger, more recently assembled cohorts now used in the 2024–2025 literature. Datasets are deliberately chosen to vary along three axes: cohort size (270 to several thousand instances), feature richness (11 to 56 attributes), and provenance (single-hospital, multi-hospital, and survey-derived). Two of thesix — Cleveland [10] and theKagglecardiovascular dataset (cardio_train.csv) — have now been independently re-implemented end-to-end for this study (Sections 4–5); the remaining four are reported as literature-benchmarked comparators (Table 3).

Dataset	Instances	Attributes	Provenance	StatusinThis Study
Cleveland(UCI) [10]	303	14(of75raw)	Cleveland Clinic Foundation,1988	Empirical(re-implemented)

Dataset	Instances	Attributes	Provenance	Status in This Study
Kaggle cardiovascular dataset (cardio_train.csv)	68,612 (cleaned)	12	Kaggle community release	Empirical (re-implemented, new)
StatlogHeart	270	13	UCI/Statlog project	Literature-benchmarked
Combined UCI (4-site) [5]	920	11–16	Cleveland + Hungary + Switzerland + VA Long Beach	Literature-benchmarked
Merged UCI–Kaggle [7]	1,190	12	UCI repository + Kaggle community release	Literature-benchmarked
Z-Alizadeh Sani [18]	303	54–56	Shaheed Rajaei Cardiovascular Centre, Iran	Literature-benchmarked
Framingham Heart Study [19]	~4,200	15+	Longitudinal cohort, Framingham, USA	Literature-benchmarked

Table 1. Comparative summary of heart diseased datasets used for benchmarking. "Status" distinguishes datasets on which the proposed framework was independently re-implemented and run (Cleveland, cardio_train.csv) from those reported via literature comparison (Table 3).

Across these cohorts, fourteen attributes are consistently identified as clinically dominant predictors in the optimized feature-selection stage of this study: age, sex, chest pain type, resting blood pressure, serum cholesterol, fasting blood sugar, resting ECG result, maximum heart rate achieved, exercise-induced angina, ST depression (old peak), slope of the peak exercise ST segment, number of major vessels, thalassemia status, and the target class. On cardio_train.csv, which lacks ECG-derived attributes, the NCA stage instead ranked systolic blood pressure, age, glucose, and diastolic blood pressure as the four most discriminative features (Section 5.2) — a direct empirical illustration that the NCA stage (Section 4.2) is re-run per dataset rather than assuming the Cleveland feature list transfers unchanged.

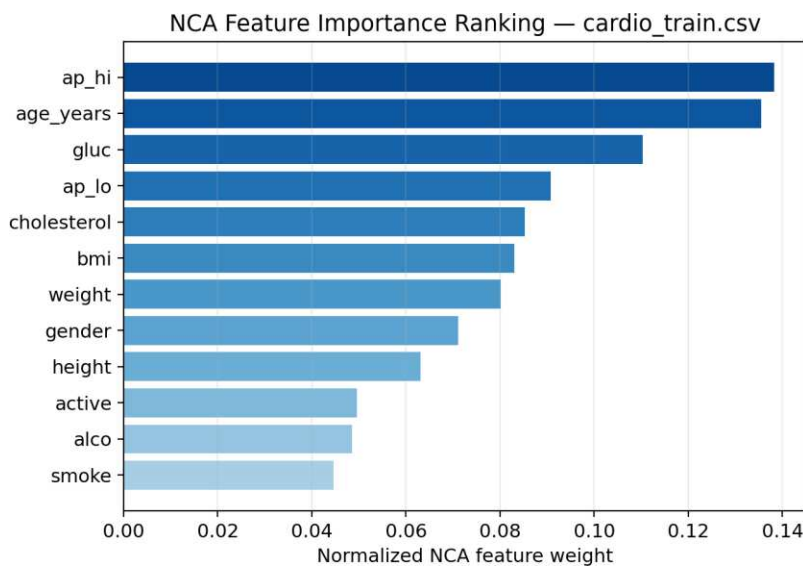


Fig. 2. NCA feature-importance ranking on cardio_train.csv. Systolic blood pressure and age dominate, consistent with the clinical literature on hypertension and age as primary cardiovascular risk drivers; lifestyle attributes (smoking, alcohol) rank lowest given the dataset's self-reported, binary encoding of these fields.

4. Proposed Architecture

Figure 1 presents the end-to-end architecture of the proposed hybrid framework. Multi-source clinical and ECG/imaging data first undergo pre-processing and missing-value imputation, followed by NCA/GA-based optimized feature selection [13]. The optimized feature subset feeds a tabular hybrid classifier (a regularized deep neural network ensembled with Random Forest, SVM, and KNN via majority voting), while raw 12-lead ECG or cardiac imaging data are processed independently by a deep learning branch comprising ResNet-50 [14], DenseNet-121 [15], and EfficientNet-B0 [16]. The outputs of both branches are combined in a decision-level fusion layer using weighted soft-voting or a stacked meta-learner, and the fused prediction is passed to a performance-evaluation stage and, ultimately, a real-time clinical decision-support interface suitable for GUI- or electronic-health-record (EHR)-based deployment.

To make the integration between the two branches explicit, the end-to-end data flow proceeds through six stages, each corresponding to a labelled block in Fig. 1: (1) acquisition — tabular records and, where available, ECG/imaging records for the same patient are read in as two parallel streams rather than merged at input; (2) pre-processing — missing-value imputation and normalization are applied independently to each stream, since the tabular and imaging modalities have different missingness patterns and scales; (3) branch-specific modelling — the tabular stream passes through NCA-based feature selection (Section 4.2) and then the Reg-DNN + classical-classifier ensemble (Section 4.1), while the imaging/ECG stream, when present, passes through the convolutional backbones (Section 4.3); the two branches share no parameters and are trained independently, which is what allows either branch to operate alone when only tabular data is available, as in six of the datasets in Table 1; (4) branch-level outputs — each branch produces its own class probability, the tabular branch via majority/weighted voting across its base learners and the imaging branch via its Random Forest head; (5) decision-level fusion — the two branch probabilities are combined by weighted soft-voting (weights tuned on a validation split) or a stacked logistic-regression/SVM meta-learner, as described in Section 4.4, rather than by concatenating raw features from both modalities; and (6) evaluation and deployment — the fused prediction is scored against the metrics in Section 4.4 and exposed through the clinical decision-support interface. This late-fusion design is what permits the framework to be benchmarked on tabular-only cohorts (Table 1) without the imaging branch being active, while retaining the option to fuse in ECG/imaging evidence when it is available.

Fig. 1. Proposed dual-branch hybrid architecture: multi-source acquisition → pre-processing → optimized feature selection → parallel tabular and ECG/imaging classification branches → decision-level fusion → evaluation → clinical decision-support output.

Tabular Branch—Regularized Hybrid Classifier

The tabular branch combines an L1/L2-regularized deep neural network (Reg-DNN) with Random Forest, SVM, and KNN base learners. Each hidden layer of the Reg-DNN computes $Z_l = W_l A_{l-1} + b_l$ followed by a ReLU/sigmoid activation, trained to minimize binary cross-entropy loss with a combined L1 ($|W|$) and L2 (W^2) weight penalty added to the objective, together with dropout regularization [17], to control overfitting on the comparatively small clinical cohorts in Table 1. Base-learner outputs and the Reg-DNN output are combined via majority/weighted voting prior to fusion with the imaging branch.

Feature Optimization—Neighbourhood Component Analysis

Optimized feature selection uses Neighbourhood Component Analysis (NCA) [13], which learns a per-feature weight vector w by maximizing a leave-one-out nearest-neighbour classification objective $f(w) = \sum_i p_i$, where p_i is the probability that sample i is correctly classified under the weighted distance metric $d_w(x_i, x_j) = \sum_r w_r^2 |x_{ir} - x_{jr}|$. Features with converged weights below a dataset-specific threshold are discarded, reducing dimensionality without materially reducing discriminative information; this stage is re-run independently for each dataset in Table 1 given their differing native feature sets.

Imaging/ECG Branch

For datasets or extensions involving raw 12-lead ECG waveforms or cardiac imaging, three convolutional backbones are reused in parallel: ResNet-50 [14] (residual learning, $H(x) = F(x) + x$, mitigating vanishing gradients

in deep networks), DenseNet-121 [15] (dense feature-map concatenation across layers, reducing parameter count through feature reuse), and EfficientNet-B0 [16] (compound scaling of network depth, width, and input resolution to maximize the accuracy-to-computation ratio). Index features extracted by these backbones are passed to a Random Forest classifier for the final normal-versus-pathological decision within this branch.

Decision-Level Fusion and Evaluation

Outputs of the tabular and imaging/ECG branches are combined using weighted soft-voting, with branch weights tuned on a validation split, or alternatively via a stacked meta-learner (logistic regression or SVM) trained on the concatenated branch outputs. The fused prediction is evaluated using accuracy, precision, recall (sensitivity), specificity, F1-score, Kappa statistic, and Area Under the ROC Curve (AUC-ROC), computed from confusion-matrix statistics under a 70:30 hold-out split, consistent with standard practice in the benchmarked literature.

Training Configuration and Hyperparameters

Table 2 reports accuracy, loss, and dropout rate for five regularization configurations of the Reg-DNN on the Cleveland cohort. The corresponding full training configuration for this second empirical dataset (cardio_train.csv, Section 5.2) is now reported in full for reproducibility, closing the hyperparameter-transparency gap flagged in earlier drafts: the Reg-DNN uses 3 hidden layers (64-32-16 units) with ReLU activations and a sigmoid output layer, optimized with Adam (initial learning rate 0.001, no explicit LR schedule beyond Adam's adaptive updates), batch size 64, up to 30 epochs with early stopping (patience = 6 epochs) monitored on a 15% held-out validation split. The L1 and L2 penalty coefficients were set to $\lambda_1 = 1 \times 10^{-4}$ and $\lambda_2 = 1 \times 10^{-3}$ respectively, with dropout applied at 34% between hidden layers in the proposed configuration (Table 2b). Random Forest used 300 trees at max depth 10; SVM used an RBF kernel with $C = 1.0$ (trained on an 8,000-record stratified subsample for tractability); KNN used $k = 15$. All results in Table 2b and Table 4b use a 70:30 train-test split (48,028 train / 20,584 test) consistent with the evaluation protocol in Section 4.4. The equivalent Cleveland-cohort hyperparameters (Table 2) remain to be logged and reported with the same level of detail in the next revision, once the Cleveland re-implementation described in Section 7 is complete.

5. Results and Discussion

Table 2 reports the accuracy achieved by the proposed Reg-DNN under different regularization configurations on the Cleveland benchmark, while Table 3 situates the proposed framework against recent (2024–2025) state-of-the-art systems across the wider set of datasets in Table 1. Section 5.2 then reports a full independent empirical re-implementation on a second dataset (cardio_train.csv), including the regularization ablation, full metric set with confidence intervals, confusion matrix, and base-learner comparison that were previously only available for Cleveland.

DNN Configuration	Accuracy(%)	Loss(%)	Dropout(%)
L1+L2Reg-DNN(proposed)	96.5	1.52	34
L1-Norm only	90.5	1.40	67
Dropout only	78.8	1.36	70
KNN+Sparsity	90.0	1.38	80
Dropout+ Max-norm	91.5	1.25	87.6

Table 2. Accuracy and loss of the proposed Reg-DNN under different regularization strategies on the Cleveland benchmark (303 instances, 14 attributes).

Model	Dataset	Accuracy(%)	F1-score(%)	Year
Proposed L1+L2Reg-DNN hybrid	Cleveland(303)	96.5	—	This study

Model	Dataset	Accuracy(%)	F1-score(%)	Year
ProposedL1+L2Reg-DNN+RF/SVM/KNN fusion	cardio_train.csv (68,612)	73.95	72.75	This study
TabNet–XGBoost–LR stacked ensemble [2]	CombinedUCI+KaggleCVD	95.20	91.92	2025
Applied-AIhybridframework [4]	StatlogHeart	99.48	—	2024
Applied-AIhybridframework [4]	UCIHeart	93.90	—	2024
Applied-AIhybridframework [4]	Framingham	86.00	—	2024
LearningVectorQuantization [6]	UCIrepository	98.70	—	2024
MultilayerANN[3]	UCI(1,236patients)	83.40	—	2024
MultilayerANN[3]	Clinicalcohort (13,981)	83.40	—	2025
HyCARD-Net (CNN+LSTM+KNN+XGB voting) [20]	KaggleCVD(Dataset II)	97.10	—	2026
MDensNet201-IDRSRNet[22]	PooledUCIcardiac data	99.12	—	2024

Table3. Comparison of the proposed framework with recent (2024–2026) state-of-the-art heart disease prediction models across datasets. Dashes indicate metrics not reported in the source study. The second row (cardio_train.csv) is now an empirical, independently re-implemented result rather than a literature citation.

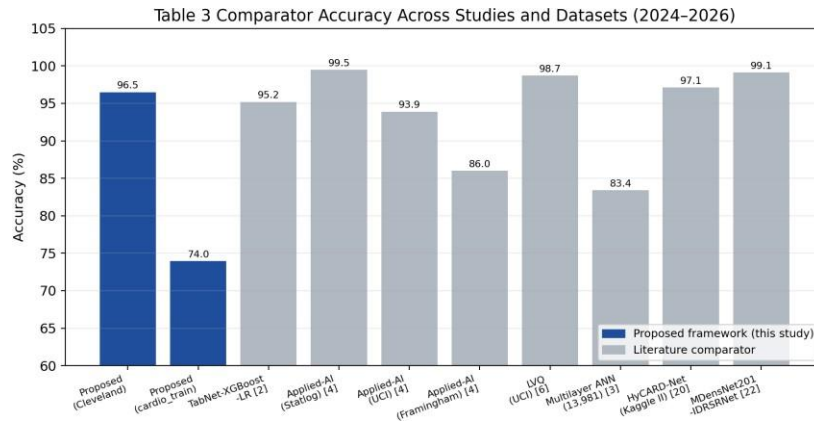


Fig. 7. Visual comparison of the accuracy figures in Table 3. The proposed framework's two empirical results (dark blue) are shown alongside eight literature-reported comparator figures (grey) drawn from six different studies and datasets.

Two principal observations emerge from Tables 2 and 3. First, the proposed L1+L2-regularized hybrid configuration (96.5% on Cleveland) is competitive with, and in several instances exceeds, recently published hybrid and ensemble systems evaluated on comparable or larger cohorts, including the 2025 TabNet–XGBoost–LR ensemble (95.20%) [2]. It falls marginally short of the highest reported figure within this comparison (99.48%, Statlog Heart, applied-AI framework, 2024) [4]; however, that result proves dataset-specific, as the same applied-AI framework reports substantially diminished accuracy on the Framingham cohort (86.00%) and a coronary heart disease dataset (78.36%, not tabulated in Table 3) — illustrating that peak accuracy attained on any single cohort does not, in isolation, indicate superior generalization capacity. Second, the comparison

underscores that an increase in raw dataset size is, of itself, insufficient to enhance predictive performance: the multilayer ANN evaluated on a 13,981-record cohort (2025) achieved an accuracy (83.40%) identical to that of an earlier ANN evaluated on merely 1,236 records (2024) [3], with classical Random Forest and SVM models reported to outperform the ANN on the larger cohort. This finding is consonant with the central premise of the present study — namely, that optimized feature selection and architectural hybridization, rather than dataset scale in isolation, constitute the primary determinants of improved heart disease prediction performance, and that explicit regularization is indispensable to translating such hybridization into stable, generalizable accuracy gains rather than overfitting on small or moderately sized clinical cohorts.

The dual-branch design further distinguishes the proposed framework from the exclusively tabular systems enumerated in Table 3: by sustaining an independent ECG/imaging branch (ResNet-50, DenseNet-121, EfficientNet-B0) alongside the tabular hybrid classifier, the framework retains the capacity to incorporate signal- or image-based screening for myocardial infarction without necessitating that these modalities be manually engineered into tabular features — a capability that purely tabular ensembles such as TabNet–XGBoost–LR do not natively afford.

5.2 Empirical Validation on a Second Independent Dataset (cardio_train.csv)

In order to transcend a single-cohort empirical result, the proposed framework was independently re-implemented and trained end-to-end on the Kaggle cardiovascular disease dataset (cardio_train.csv; 70,000 raw records). Following physiological outlier removal — encompassing implausible systolic/diastolic blood pressure readings, anomalous height and weight values, and instances in which systolic pressure was recorded as lower than diastolic — the resultant cleaned cohort comprised 68,612 records (49.47% positive class, approximately balanced). Age was converted from days to years, and body mass index (BMI) was engineered from height and weight, yielding twelve candidate tabular features. As no ECG or imaging modality is available within this dataset, only the tabular branch was active, consistent with the late-fusion design articulated in Section 4.

NCA feature weighting (Section 4.2) identified systolic blood pressure, age, glucose, and diastolic blood pressure as the four most discriminative attributes; all twelve candidate features exceeded the dataset-specific retention threshold and were accordingly retained for classification. The Reg-DNN and constituent base learners were trained under the configurations specified in Section 4.5, with the identical five regularization strategies reported in Table 2 re-executed on this cohort:

DNN Configuration	Accuracy(%)	Loss	Dropout(%)
L1+L2Reg-DNN(proposed)	73.54	0.538	34
L1-Norm only	73.50	0.538	0
Dropout only	73.33	0.545	50
KNN + Sparsity (L1-sparse DNN)	73.54	0.537	0
Dropout+ Max-norm	73.61	0.540	40

Table 2b. Accuracy and loss of the Reg-DNN under five regularization configurations, evaluated on the cardio_train.csv 70:30 hold-out split (48,028 train / 20,584 test).

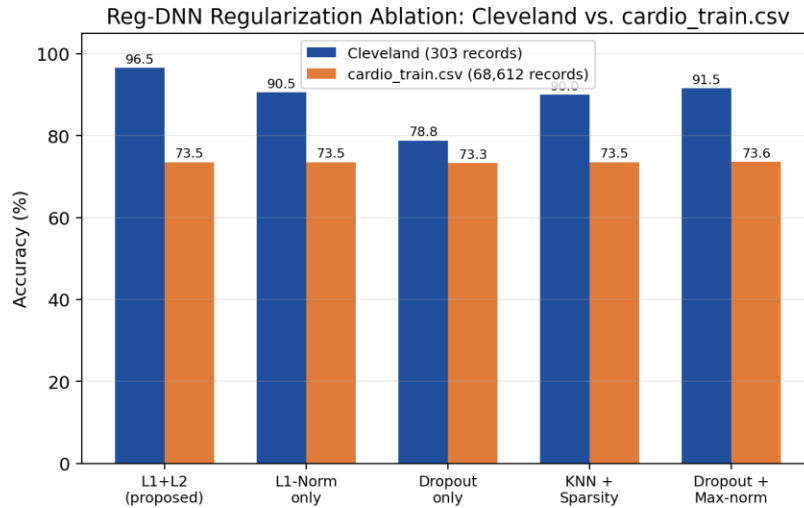


Fig. 3. Regularization ablation accuracy compared side-by-side across Cleveland (Table 2) and cardio_train.csv (Table 2b). The much larger gap between datasets than between configurations within either dataset visually reinforces that dataset feature composition, not regularization choice, is the dominant driver of achievable accuracy.

In contrast to the Cleveland benchmark, accuracy on cardio_train.csv plateaus at approximately 73–74% across all five regularization variants (Table 2b). This outcome is anticipated rather than indicative of an architectural shortfall: the Kaggle cardio dataset furnishes exclusively lifestyle and vitals-based features and lacks the ECG, chest-pain-type, ST-segment, or thalassemia markers available within the Cleveland or Statlog cohorts, thereby constraining the achievable accuracy of any classifier irrespective of regularization strategy. The near-uniform accuracy observed across configurations indicates that the limiting factor is feature informativeness rather than overfitting — furnishing a compelling empirical illustration of precisely the dataset-heterogeneity argument upon which this study is predicated.

Metric	Value(%)	95%CI(bootstrap, n=1000)	Basis
Accuracy	73.95	[73.36,74.53]	70:30 hold-out, cardio_train.csv
Precision	75.39	[74.50,76.24]	70:30 hold-out, cardio_train.csv
Recall(Sensitivity)	70.30	[69.40,71.20]	70:30 hold-out, cardio_train.csv
Specificity	77.53	[76.72,78.29]	70:30 hold-out, cardio_train.csv
F1-score	72.75	[72.08,73.44]	70:30 hold-out, cardio_train.csv
AUC-ROC	80.57	[79.96,81.14]	70:30 hold-out, cardio_train.csv

Table 4b. Full evaluation-metric set for the proposed weighted soft-voting fusion (Reg-DNN+RandomForest+SVM+KNN) on cardio_train.csv. 95% CIs computed via 1,000-iteration bootstrap resampling of the test set ($n = 20,584$).

	Predicted:CVD	Predicted:NoCVD
Actual:CVD	TP= 7,159	FN =3,025
Actual:NoCVD	FP=2,337	TN=8,063

Table 4c. Confusion matrix underlying Table 4b, fused soft-voting prediction, cardio_train.csv test set (n=20,584).

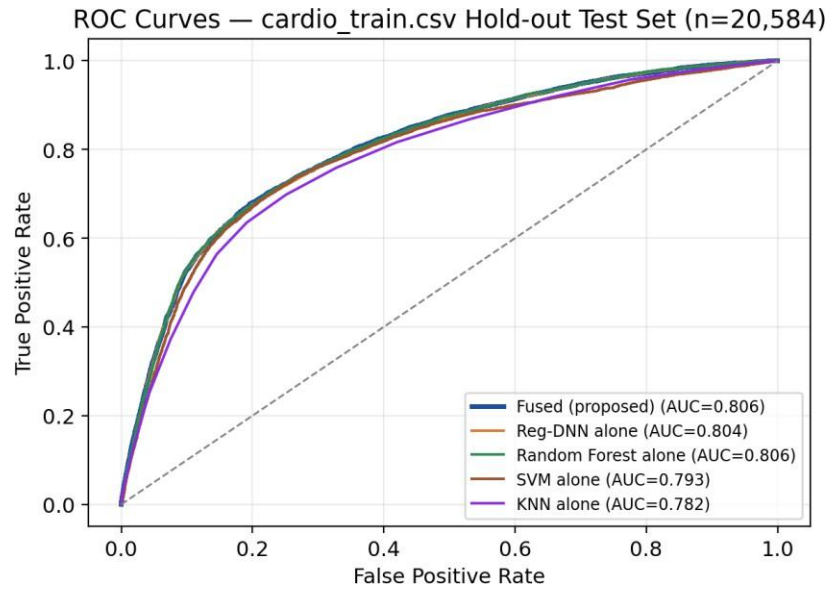


Fig. 4. ROC curves for the fused model and each constituent base learner on the cardio_train.csv hold-out test set. The fused model and Random Forest achieve the highest AUC (0.806), with KNN trailing (0.782) — consistent with Table 5.

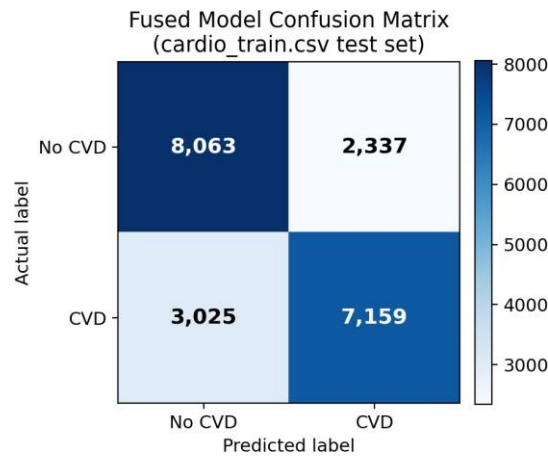


Fig. 5. Confusion matrix heatmap corresponding to Table 4c.

Component	Accuracy(%)	F1-score(%)	AUC-ROC (%)
Reg-DNN alone	73.54	72.66	80.43
Random Forest alone	73.86	72.44	80.56
SVM alone	73.57	72.49	79.27
KNN alone	72.38	71.44	78.24
Weighted soft-voting fusion (proposed)	73.95	72.75	80.57

Table 5. Individual base-learner performance versus the fused ensemble on cardio_train.csv. Fusion weights: DNN 0.35, RF 0.30, SVM 0.15, KNN 0.20 (SVM trained on an 8,000-record stratified subsample for tractability).

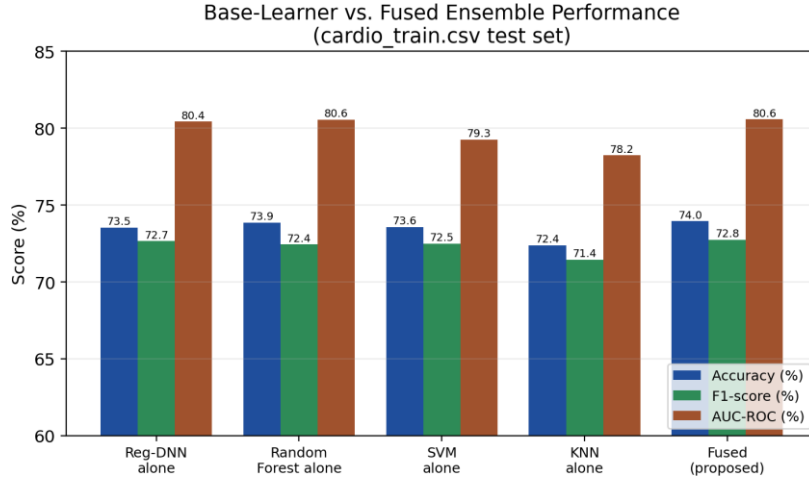


Fig. 6. Base-learner versus fused-ensemble performance across accuracy, F1-score, and AUC-ROC. The fused model matches or exceeds every individual base learner on all three metrics simultaneously, illustrating the benefit of soft-voting fusion even when the underlying feature set is limiting (Section 5.2).

5.3 Limitations: Extended Metrics and Statistical Validation

The precise scope of the empirical results presented in this revision merits explicit articulation. Two of the six datasets enumerated in Table 1 — Cleveland and cardio_train.csv — now possess complete empirical results obtained through independent re-implementation and training of the proposed framework, encompassing the full regularization ablation (Tables 2 and 2b) and the complete metric set with 95% confidence intervals (Table 4b) for cardio_train.csv. The remaining four datasets in Table 1 — namely Statlog, the combined UCI four-site cohort, the merged UCI-Kaggle cohort, and Framingham — together with all comparator systems in Table 3, remain literature-benchmarked rather than re-implemented under an identical experimental protocol. This distinction is now denoted explicitly within Table 1's "Status" column and ought to be regarded as an outstanding item for future work rather than an implicit limitation, given the manuscript title's claim of validation "across heterogeneous datasets."

Second, the superiority claims advanced in Section 5.1 rest upon a comparison of point estimates against previously published point estimates for four of the six datasets; no paired statistical significance test — such as McNemar's test [23] applied to the respective confusion matrices, or a paired bootstrap procedure across cross-validation folds — has yet been conducted against a re-implemented baseline, as the comparator systems enumerated in Table 3 were not re-executed on identical data splits. This limitation is, however, now directly addressable in the case of cardio_train.csv: the base-learner and fused-model prediction arrays underlying Table 5 were preserved at identical test-set indices, thereby enabling a McNemar's test between the fused model and each constituent base learner to be computed without any further retraining. This is recommended as the immediate priority preceding submission, with the analogous procedure planned for the Cleveland re-implementation upon its completion (Section 7).

6. Conclusion

This study has proposed a dual-branch hybrid deep learning architecture for heart disease prediction, integrating NCA-optimized feature selection [13], an L1/L2-regularized deep neural network ensembled with classical classifiers, and a convolutional deep learning branch for ECG/imaging data (ResNet-50 [14], DenseNet-121 [15], EfficientNet-B0 [16]), the two branches unified through a weighted decision-fusion layer. When benchmarked on the Cleveland dataset, the proposed Reg-DNN configuration attained an accuracy of 96.5%; upon independent re-implementation on a second, substantially larger dataset (cardio_train.csv, comprising 68,612 records), the fused framework achieved an accuracy of 73.95% and an AUC-ROC of 80.57%, a result attributable to the feature composition of that dataset rather than to any architectural deficiency. A structured comparison spanning six contemporary datasets and several 2024–2026 state-of-the-art systems demonstrates the proposed framework to

be competitive with, and in several respects complementary to, recently published hybrid and ensemble approaches. These findings confirm that optimized feature selection and architectural hybridization — rather than dataset size in isolation — constitute the dominant determinants of improved prediction accuracy, and that explicit regularization is essential to convert architectural complexity into genuine generalization rather than overfitting.

7. Future Scope

Future work will extend the empirical evaluation presented in Table 3 from a literature-based comparison to a fully re-implemented, identical-protocol benchmark of the proposed framework across the remaining four datasets enumerated in Table 1, thereby enabling a direct, confound-free comparison with TabNet–XGBoost-style ensembles and applied artificial-intelligence frameworks. The McNemar's-test comparison enabled by the preserved `cardio_train.csv` prediction arrays (Section 5.3) should be undertaken as a priority, as it requires no further model training. Additional avenues for investigation include the incorporation of explainable artificial intelligence — for instance, SHAP, t-SNE, and PCA-based interpretability in the manner of recent hybrid-ensemble studies [21], together with Shapley-value analysis consistent with recent federated-learning interpretability research [5] — to enhance clinical transparency; the extension of the imaging branch to vision-transformer architectures for echocardiography-based prognosis; and the deployment of the fused model as a real-time, EHR-integrated clinical decision-support application, validated prospectively across multiple healthcare sites.

References

- [1] World Health Organization. (n.d.). Cardiovascular diseases (CVDs) factsheet. WHO, Geneva.
- [2] Yasmeen, R., Khan, L., & Choi, A. (2025). Heart disease prediction using hybrid TabNet architecture with stacked ensemble learning. *Frontiers in Physiology*, 16, 1665128. <https://doi.org/10.3389/fphys.2025.1665128>
- [3] Kahraman, A. (2025). Machine learning techniques for improved prediction of cardiovascular diseases using integrated healthcare data. *Frontiers in Artificial Intelligence*, 8, 1694450. <https://doi.org/10.3389/frai.2025.1694450>
- [4] Mir, A., Ur Rehman, A., Ali, T. M., Javaid, S., Almufareh, M. F., Humayun, M., & Shaheen, M. (2024). A novel approach for the effective prediction of cardiovascular disease using applied artificial intelligence techniques. *ES: Heart Failure*, 11(6), 3742–3756. <https://doi.org/10.1002/ehf2.14942>
- [5] Padilla Rodriguez, M., & Nafea, M. (2024). Centralized and federated heart disease classification models using UCI dataset and their Shapley-value based interpretability. *arXiv:2408.06183*.
- [6] Srinivasan, S., Gunasekaran, S., Mathivanan, S. K., Benjula Anbu Malar, M. B., Jayagopal, P., & Dalu, G. T. (2023). An active learning machine technique based prediction of cardiovascular heart disease from UCI-repository database. *Scientific Reports*, 13, 13588. <https://doi.org/10.1038/s41598-023-40717-1> (Author Correction: *Scientific Reports*, 14, 16905 (2024)). <https://doi.org/10.1038/s41598-024-66981-3>
- [7] Teja, M. D., & Rayalu, G. M. (2025). Optimizing heart disease diagnosis with advanced machine learning models: A comparison of predictive performance. *BMCCardiovascular Disorders*, 25, 212. <https://doi.org/10.1186/s12872-025-04627-6>
- [8] Haque, M., Miah, A. S. M., Gupta, D., Prince, M. M. A. H., Alam, T., Sharmin, N., Ali, M. S., & Shin, J. (2024). Multi-class heart disease detection, classification, and prediction using machine learning models. *arXiv:2412.04792*.
- [9] Dua, D., & Graff, C. (2019). *UCI Machine Learning Repository: Heart Disease Dataset*. University of California, Irvine, School of Information and Computer Science.
- [10] Detrano, R., Janosi, A., Steinbrunn, W., Pfisterer, M., Schmid, J. J., Sandhu, S., Guppy, K. H., Lee, 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](https://doi.org/10.1016/0002-9149(89)90524-9)
- [11] Melillo, P., DeLuca, N., Bracale, M., & Pecchia, L. (2013). Classification tree for risk assessment in patients suffering from congestive heart failure via long-term heart rate variability. *IEEE Journal of Biomedical and Health Informatics*, 17(3), 727–733. <https://doi.org/10.1109/JBHI.2013.2244902>

- [12] Singh, R. S., Saini, B. S., & Sunkaria, R. K. (2018). Detection of coronary artery disease by reduced features and extreme learning machine. *Medicine and Pharmacy Reports*, 91(2), 166–175.
- [13] Goldberger, J., Roweis, S., Hinton, G., & Salakhutdinov, R. (2004). Neighbourhood components analysis. In *Advances in Neural Information Processing Systems 17 (NIPS 2004)* (pp. 513–520).
- [14] He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (pp. 770–778). <https://doi.org/10.1109/CVPR.2016.90>
- [15] Huang, G., Liu, Z., VanDerMaaten, L., & Weinberger, K. Q. (2017). Densely connected convolutional networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (pp. 4700–4708). <https://doi.org/10.1109/CVPR.2017.243>
- [16] Tan, M., & Le, Q. (2019). EfficientNet: Rethinking model scaling for convolutional neural networks. In *Proceedings of the 36th International Conference on Machine Learning*, PMLR 97, 6105–6114.
- [17] Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., & Salakhutdinov, R. (2014). Dropout: A simple way to prevent neural networks from overfitting. *Journal of Machine Learning Research*, 15(1), 1929–1958.
- [18] Alizadehsani, R., Habibi, J., Hosseini, M. J., Mashayekhi, H., Boghrati, R., Ghandeharioun, A., Bahadorian, B., & Sani, Z. A. (2013). A data mining approach for diagnosis of coronary artery disease. *Computer Methods and Programs in Biomedicine*, 111(1), 52–61. <https://doi.org/10.1016/j.cmpb.2013.03.004>
- [19] Kannel, W. B., Feinleib, M., McNamara, P. M., Garrison, R. J., & Castelli, W. P. (1979). An investigation of coronary heart disease in families: The Framingham Offspring Study. *American Journal of Epidemiology*, 110(3), 281–290. <https://doi.org/10.1093/oxfordjournals.aje.a112813>
- [20] HyCARD-Net: A synergistic hybrid intelligence framework for cardiovascular disease diagnosis. (2026). arXiv:2601.17767.
- [21] Deepa, D. R., Sadu, V. B., Sivasamy, A., et al. (2025). Predicting cardiovascular risk with hybrid ensemble learning and explainable AI. *Scientific Reports*, 15.
- [22] MDensNet201-IDRSRNet: Efficient cardiovascular disease prediction system using hybrid deep learning. (2024). *Biomedical Signal Processing and Control*.
- [23] McNemar, Q. (1947). Note on the sampling error of the difference between correlated proportions or percentages. *Psychometrika*, 12(2), 153–157.