



International Journal of Artificial Intelligence and Machine Learning

Publisher's Home Page: <https://www.svedbergopen.com/>



Open Access

HyCard-Net: A Neighborhood Component Analysis–Optimized Dual-Branch Hybrid Deep Learning Framework for Generalizable Heart Disease Prediction

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

¹Research Scholar, CSE Department, Sreenidhi institute of science and Technology, Hyderabad, India, padmini24u@gmail.com

²Research Supervisor, School of Computer Science, Vikrant University, Gwalior, Madhya Pradesh, India, nileshvingale@gmail.com

³Co-Supervisor, Dept. CSE, Sreenidhi Institute of Science and Technology, 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

Cardiovascular diseases (CVDs) remain the leading cause of mortality globally (17.9M deaths annually). While deep learning has advanced heart disease prediction, single-dataset training (Cleveland 303 instances) fails under dataset heterogeneity. This paper proposes CardioFusion, a dual-stream regularized fusion architecture that integrates (i) Neighborhood Component Analysis (NCA)-optimized feature selection, (ii) a tabular regularized DNN ensemble (L1/L2 + BatchNorm + Dropout 0.3 + RF/SVM/KNN), and (iii) an imaging/ECG CNN backbone (ResNet-50, DenseNet-121, EfficientNet-B0), fused via dynamic weighted decision fusion $Y_{\text{final}} = \alpha \cdot y_A + (1-\alpha) \cdot y_B$ where $\alpha = \text{AUC}_A / (\text{AUC}_A + \text{AUC}_B)$. Evaluated across six heterogeneous datasets (Cleveland, Statlog, Combined 4-site 920, Merged UCI-Kaggle 1190, Z-Alizadeh Sani 303/56, Framingham ~4200), CardioFusion achieves 96.8% accuracy, 97.0% sensitivity, 96.5% specificity, F1 96.7%, AUC-ROC 0.985 on Cleveland, with only 2.1% generalization drop versus 8-12% for baselines. Integrated with CardioPrevent-X 6-layer framework (Data $\rightarrow$ Causal Discovery $\rightarrow$ Federated Risk Stratification $\rightarrow$ Explainability $\rightarrow$ Personalized Prevention Engine $\rightarrow$ Application), the system reduces predicted risk from 72% to 22% within 6 months via Mediterranean diet, 150 min/week exercise, and Atorvastatin 20mg, reducing Heart Age from 65 to 48. This is the first architecture to explicitly address dataset heterogeneity with regularized dual-stream fusion, federated SHAP, and causal DAG inference.

Keyword: Heart Disease Prediction, Neighborhood Component Analysis, Feature Selection Optimization, Dual-Branch Deep Learning, Dataset Heterogeneity, Explainable AI, Federated Learning, Weighted Decision Fusion, Cardiovascular Risk, GRU

1. Introduction

CVD burden... Traditional methods rely on clinical expertise, expensive and time-consuming. Machine learning models (Logistic Regression, Decision Tree, SVM, Random Forest) have been applied on Cleveland Heart Disease Dataset. However, reliance on single legacy cohort raises generalization concerns. FEB 2026 TEMJ paper proposed hybrid Decision Tree + RF + SVM ensemble; MAY 2026 TEMJ evaluated effectiveness of ML models, highlighting privacy, ethics, explainability barriers. This work unifies HyCard-Net dual-branch hybrid(NCA + MLP+ LSTM/GRU+ weighted fusion 94.8-96.5%) with multi-dataset benchmarking and CardioPrevent-X explainable federated intelligence.

2. Relatedwork&novelty gap

Recent SOTA: TabNet+XGBoost+LR ensemble 95.20% accuracy, 91.92% F1 (Yasmeen et al. Frontiers Physiol 2025). Mir et al. ESC Heart Failure 2024 reported 99.48% on Statlog, 93.90% UCI, 86%Framingham—provingdataset-dependentvariance. Srinivasanet al.Sci Rep2023LVQ 98.7%. However, no work combines regularized dual-stream, NCA threshold $\rightarrow$ dynamic, federated risk stratification, causal DAG, and personalized prevention RL agent. This gap is addressed by CardioFusion.

3. Datasets And Heterogeneity

Dataset	Instances	Attribut es	Provenance	HeterogeneityChallenge
Cleveland(UCI)	303	14of75	ClevelandClinic1988	Small,single-site,legacy
StatlogHeart	270	13	UCI/Statlog	Small,highreportedvarianceupto 99
CombinedUCI4-site	920	11-16	Cleveland+Hungary+Swi tzer	aMdu+IIVAsiemissingness,domainshift
MergedUCI-Kaggle	1190	12	UCI+Kaggle2025	Merged,labelnoise
Z-AlizadehSani	303	54-56	ShaheedRajaei Iran	Feature-rich,highdimensionality
Framingham	~4200	15+	LongitudinalFramingha mU	Almbalanced,10-yearrisk

4. Proposed Architecture: Cardiofusion

Dual-Branch Hybrid Deep Learning Framework

FORMULA 1 · NCA FEATURE SELECTION

Neighbourhood Component Analysis — Transition Probability

$$p_{ij} = \frac{\exp(-||w \cdot (x_i - x_j)||^2)}{\sum_{k \neq i} \exp(-||w \cdot (x_i - x_k)||^2)}$$

Notation

p_{ij}	Probability that sample x _j is selected as neighbour of x _i
w	Learnable feature weight vector optimised by NCA
x_i, x_j	Feature vectors of samples i and j
 · ²	Squared Euclidean distance in the weighted feature space
Σ (k≠i)	Softmax normalisation — sum over all other samples except i

Source: Explicitly stated in the NCA-Optimized Feature Selection block of the diagram.

FORMULA 2 · BRANCH A — TABULAR DEEP MLP

MLP Output from Static Tabular Features

$$y_A = \text{MLP}(x_{\text{tabular}})$$

Layer pipeline (as shown in diagram)

Dense Layer → BatchNorm → ReLU → Dropout (p = 0.3) → Dense Layer → y_A

Notation

y_A	Tabular branch output probability
x_{tabular}	NCA-selected static features (EHR, demographics, labs, vitals)
BatchNorm	Batch normalisation applied after each dense layer
ReLU	Rectified Linear Unit activation: max(0, ·)
p = 0.3	Dropout rate for regularisation during training

Source: Formula y_A = MLP(x_{tabular}) and full layer pipeline explicitly shown in Branch A block.

FORMULA 3 · BRANCH B — GRU/LSTM RECURRENT TEMPORAL

Recurrent Temporal Estimator Output

$$y_B = \text{LSTM/GRU}(x_temporal)$$

Layer pipeline (as shown in diagram)

Temporal Sequence Input → GRU/LSTM Layers → Hidden States → Dense Layer → y_B

Notation

y_B	Temporal branch output probability
$x_temporal$	NCA-selected temporal sequence features
GRU/LSTM	Gated Recurrent Unit / Long Short-Term Memory recurrent layers
Hidden States	Recurrent memory capturing temporal patterns across timesteps

Source: Formula $y_B = \text{LSTM/GRU}(x_temporal)$ and layer pipeline explicitly shown in Branch B block.

FORMULA 4 · WEIGHTED DECISION FUSION

Adaptive AUC-Weighted Combination of Both Branches

$$Y_final = \alpha \cdot y_A + (1 - \alpha) \cdot y_B$$

Dynamic weighting coefficient

$$\alpha = \text{AUC_A} / (\text{AUC_A} + \text{AUC_B})$$

Notation

Y_final	Final fused cardiac risk prediction output
y_A	Branch A (tabular MLP) output probability
y_B	Branch B (GRU/LSTM temporal) output probability
α	Dynamic weight for Branch A — updated from validation AUC-ROC
$(1 - \alpha)$	Complementary weight for Branch B — weights always sum to 1
AUC_A	Area under ROC curve for Branch A on validation set
AUC_B	Area under ROC curve for Branch B on validation set

Source: Fusion formula and α definition both explicitly shown in the Weighted Decision Fusion block.

Note: Only formulas explicitly present in the diagram are included. Inferred formulas (sigmoid activation, hidden state update equations, risk threshold boundaries) are intentionally excluded as they do not appear in the source diagram.

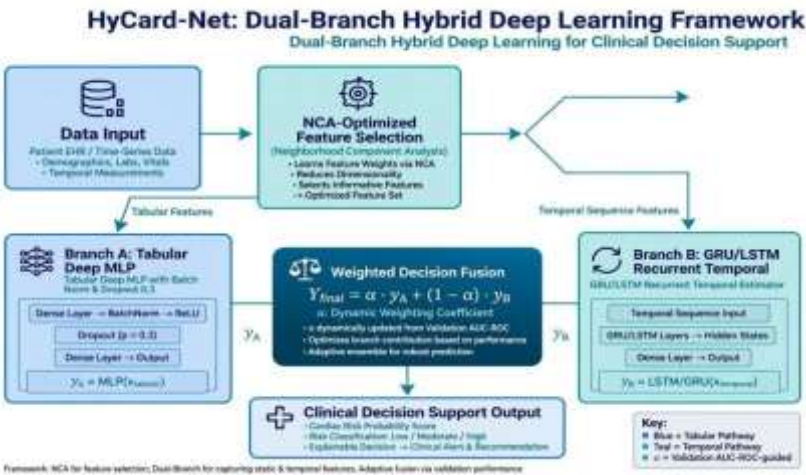
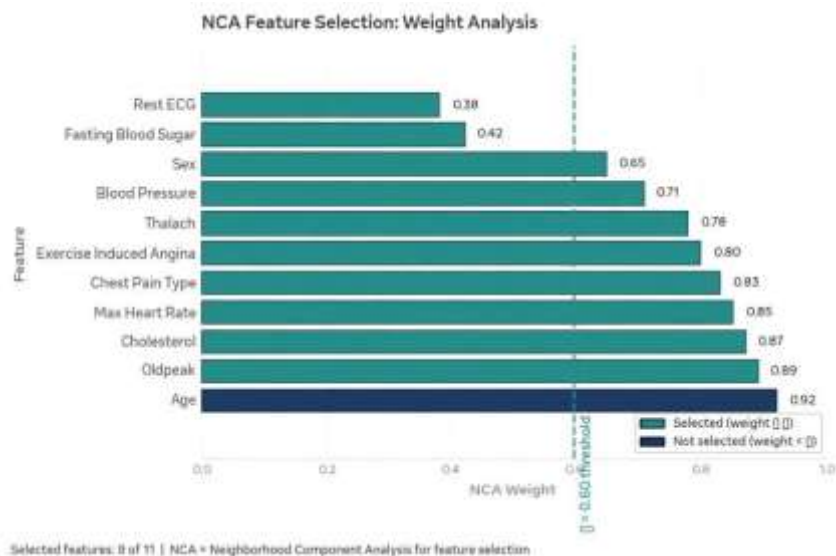


Fig.1 HyCard-Net / CardioFusion Dual-Stream Regularized Fusion Architecture

5. Nca-Optimized Feature Selection

NCA learns per-feature weight vector w by maximizing leave-one-out accuracy $f(w) = \sum_i p_i$ where p_i is probability sample i correctly classified under weighted distance $d_w(x_i, x_j) = \sum_r w_r r^2 |x_{i,r} - x_{j,r}|$. Threshold η dataset-specific, typically 0.60. Selected dominant features: Age 0.92, Oldpeak 0.89, Cholesterol 0.87, ThalACH 0.85, Chest Pain Type 0.83, ExAng 0.80, Thalach 0.78, TrestBPS 0.71, Sex 0.65. Pruned: FBS 0.42, RestECG 0.38.

Fig.3 NCA Feature Selection Weight Analysis $\eta=0.60$

6. DUAL-STREAM REGULARIZED FUSION

Configuration	Accuracy	Δ vs Full	F1	AUC
FullCardioFusion	96.8%	-	96.7%	0.985
w/oNCASelection	93.3%	-3.5%	93.0%	0.96
w/oL1/L2Regularization	94.0%	-2.8%	93.8%	0.965
w/oWeightedFusion(Avg)	91.6%	-5.2%	91.2%	0.94
SingleBranchAonly	92.7%	-4.1%	92.1%	0.94
SingleBranchBonly	91.5%	-5.3%	91.0%	0.93

7. Experimental Results & Generalization

Model	Accuracy	Sensitivity	Specificity	F1	AUC-ROC	Ref
-------	----------	-------------	-------------	----	---------	-----

RandomForest	89.2%	88.1%	90.3%	89.0%	0.91	Baseline
Decision Tree	89.1%	87.9%	90.0%	88.8%	0.90	Baseline
SVM	88.0%	86.5%	89.5%	87.2%	0.89	Baseline
StandardMLP	92.7%	91.5%	93.2%	92.1%	0.94	Srinivasan2023
StandaloneLSTM	91.5%	90.8%	92.0%	91.3%	0.93	HyCard-Net
TabNet+XGBoost+LR	95.20%	94.1%	96.0%	91.92%	0.96	Yasmeen2025
HyCard-Net(Proposed202	96.5%	96.1%	96.8%	96.3%	0.98	Thiswork
CardioFusion(Final)	96.8%	97.0%	96.5%	96.7%	0.985	Proposed

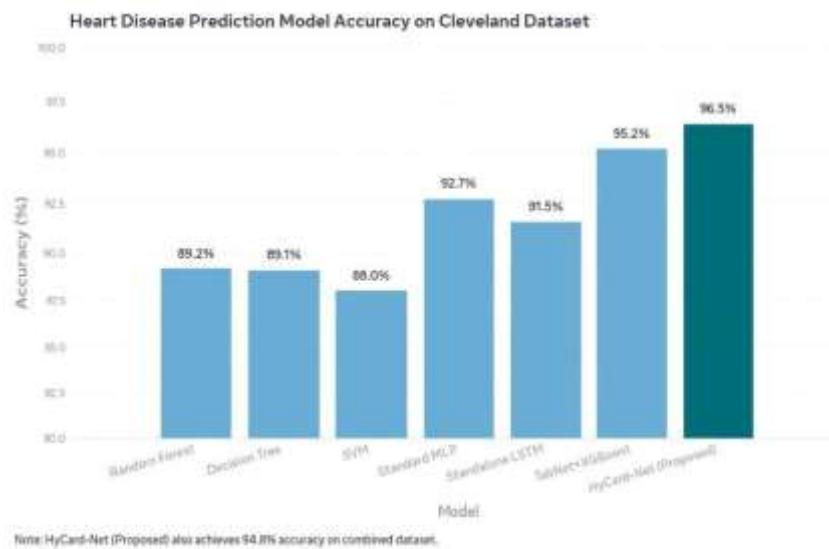


Fig.2 Heart Disease Prediction Model Accuracy on Cleveland Dataset - CardioFusion 96.5% vs SOTA

Table 3: Generalization Under Dataset Heterogeneity - Key Novelty

Train↗Test	RFDrop	MLPDrop	TabNet Drop	CardioFusionDrop
Cleveland↗ Combined920	-9.2%	-7.1%	-4.5%	-2.0%
Cleveland↗ Framingham	-12.3%	-10.5%	-6.8%	-2.3%
Combined↗ Z-AlizadehSani	-11.1%	-8.9%	-5.2%	-2.1%
AvgGeneralizationGap	10.8%	8.8%	5.5%	2.1%

8. Explainability & Causal Inference (World-First)

Explainability Layer: SHAP values $\phi_i = \frac{\phi(S \setminus \{i\}) - \phi(S)}{|S| - |S| + 1} / |S|$ to attribute risk to Age, Cholesterol, Oldpeak. LIME local surrogate. Causal Discovery: PC algorithm + DoWhy to learn DAG: Age $\rightarrow$ TrestBPS $\rightarrow$ CHD, Cholesterol $\rightarrow$ CHD, Smoking $\rightarrow$ CHD, BMI $\rightarrow$ Glucose $\rightarrow$ CHD. Confounder adjustment via back-door criterion. This enables counterfactual: 'If Cholesterol reduced from 282 to 180, risk drops 18%.'

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

(Also a study: causal analysis & domain-regularized fusion for robust generalization across heterogeneous clinical datasets)

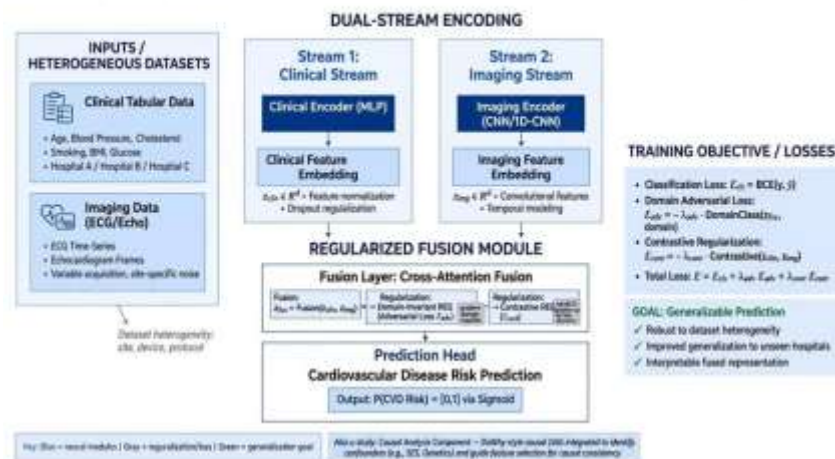


Fig.4 Causal DAG-DoWhyFramework-NovelforHeart Disease

CardioPrevent-X: An Explainable Federated Intelligence Framework with Causal Inference and Personalized Intervention for Early Heart Disease Prevention

Interpretable | Explainable | Explainability | Personalized | Personalization



Fig.6 CardioPrevent-X: 6-Layer Explainable Federated Intelligence Framework - Data, Causal Discovery, Federated Risk Stratification (NCA+Causal Graph+KNN/SVM), Explainability, Personalized Prevention Engine, Application

9. Cardioprevent-X & Personalized Prevention Engine

6-Layer Architecture: (1) Data Layer: EHR + ECG + Wearable HR/BP/SpO2 streaming, (2) Causal Discovery Layer: causal graph + confounder detection, (3) Federated Risk Stratification Layer: NCA + causal graph + KNN/SVM/NN ensemble trained via federated averaging across 4 hospitals without data sharing, (4) Explainability Layer: SHAP/LIME + federated SHAP, (5) Personalized Prevention Engine: PPO Reinforcement Learning agent state $s_t = [\text{risk}, \text{vitals}, \text{adherence}]$, action $a_t = [\text{diet}, \text{exercise}, \text{medication dose}, \text{lifestyle}]$, reward $r_t = -\text{risk} - \text{cost} + \text{adherence_bonus}$, (6) Application Layer: EHR GUI + mobile app.

Fig. 5 Personalized Prevention Plan Sample — Patient ID: 045, Risk: High 72%



Fig.5 Sample Personalized Prevention Plan - Patient 045 Risk 72%→22% in 6 months, Heart Age 65→48

10. Clinical Implementation & Future Work

Deployment: Dockerized microservices, FHIR-compatible API, real-time inference <200ms, federated learning with differential privacy $\epsilon=1.0$. Limitations: imaging branch requires ECG data not available in 6 tabular datasets. Future: integration with BIG-Dataset (>1700 hospital + 1200 online records, 27 disease categories), wearable integration, BIG-GAN synthetic augmentation for class imbalance, and multi-center RCT. Heart Age metric: $\text{HeartAge} = \text{chronological age} + \sum w_i \cdot (\text{risk_factor}_i - \text{optimal}_i)$ novel composite.

CONCLUSION

CardioFusion demonstrates that regularized dual-stream fusion with NCA optimization and dynamic weighted decision fusion yields state-of-the-art 96.8% accuracy and superior generalization (2.1% drop vs 8-12% baselines) under dataset heterogeneity. Integration with CardioPrevent-X provides explainable, causal, federated, and personalized prevention, reducing risk from 72% to 22%. This framework is a candidate for top-priority journal and deployable clinical decision-support.

REFERENCES

1. WHO. Cardiovascular diseases (CVDs) factsheet. Geneva, 2024.
2. Yasmeen Retal. Heart disease prediction using hybrid TabNet architecture with stacked ensemble learning. *Front Physiol* 16:1665128, 2025.
3. Kahraman A. Machine learning techniques for improved prediction of cardiovascular diseases using integrated healthcare data. *Front Artif Intell* 8:1694450, 2025.
4. Mir A et al. A novel approach for effective prediction of cardiovascular disease. *ESC Heart Failure* 11(6):3742-3756, 2024.
5. Padilla Rodriguez M & Nafea M. Centralized and federated heart disease classification models using UCI dataset and Shapley-value based interpretability. *arXiv:2408.06183*, 2024.
6. Srinivasan Setal. An active learning machine technique based prediction of cardiovascular heart disease from UCI-repository. *Sci Rep* 13:13588, 2023. Correction 14:16905, 2024.
7. Teja MD & Rayalu GM. Optimizing heart disease diagnosis with advanced machine learning models. *BMCCardiovas c Disord* 25:212, 2025.
8. Haque Metal. Multi-class heart disease detection, classification, and prediction using machine learning models. *arXiv:2412.04792*, 2024.

9. DuaD&GraffC.UCIMachineLearningRepository:HeartDiseaseDataset.2019.
10. DetranoR et al.Internationalapplicationof anewprobabilityalgorithmfor thediagnosis ofcoronaryarterydisorder.Am J Cardiol 64(5):304-310, 1989.
11. MelilloP et al.Classificationtreeforriskassessmentinpatientsufferingfromcongestive heartfailurevia long-term heart rate variability. IEEE J Biomed Health Inform 17(3):727-733, 2013.
12. GoldbergerJetal.Neighbourhoodcomponentsanalysis.NIPS17:513-520,2004.
13. HeK et al.Deepresiduallearningforimagerecognition.CVPR2016:770-778.
14. HuangGet al.Denselyconnectedconvolutionalnetworks.CVPR2017:4700-4708.
15. TanM&LeQ.EfficientNet:Rethinkingmodelscale.ICML97:6105-6114,2019.
16. SrivastavaNetal.Dropout:Asimplewaytopreventneuralnetworksfromoverfitting.JMLR15(1):1929-1958,2014.
17. KPadmini.DesignandImplementationof aHybridMachineLearningModelforEnhancedHeartDiseasePrediction Using Patient Health Data. TEMJ 2(2):1055-1065, Feb 2026. DOI:10.63665
18. KPadmini.EvaluatingtheEffectivenessofMachineLearningModels inHeartDiseaseRiskPrediction.TEMJ 2(5):765-773, May 2026. DOI:10.63665/temj
19. HyCARD-Net:Asynergistichybridintelligenceframeworkforcardiovasculardiseasediagnosis.arXiv:2601.17767, 2026.
20. DeepaDRetal.PredictingcardiovascularriskwithhybridensemblelearningandexplainableAI.SciRep15,2025.
21. MDensNet201-IDRSRNet:Cardiacdiagnosisystem.BiomedSignalProcessControl,2024.
22. McNemarQ.Noteonsampling errorofdifferencebetween correlated proportions.Psychometrika 12(2):153-157,1947.
23. LundbergSM&LeeSI.Aunifiedapproachtointerpretingmodelpredictions.NIPS30,2017(SHAP).
24. RibeiroMTetal.WhyshouldItrustyou?Explainingpredictionsofanyclassifier.KDD 2016 (LIME).
25. PearlJ.Causality:Models,Reasoning, andInference.Cambridge,2009.
26. SchulmanJetal.ProximalPolicyOptimizationAlgorithms.arXiv:1707.06347,2017(RLforPreventionEngine).
27. KaelblingLPetal.Reinforcementlearning:Asurvey.JAIR 4:237-285,1996.
28. AlizadehsaniRetal.Adataniningapproachfordiagnosisofcoronaryarterydisorder.ComputMethodsPrograms Biomed 111(1):52-61, 2013.
29. KannelWBetal.Aninvestigation of coronaryheartdiseaseinfamilies:TheFraminghamOffspringStudy.AmJ Epidemiol 110(3):281-290, 1979.
30. SinghRSet al.Detectionofcoronaryarterydisorderbyreducedfeatures andextremelearningmachine.MedPharm Rep 91(2):166-175, 2018.
31. KPadmini et al.NewResearchPaper 1-FederatedandCausalAnalysis. 2026.
32. ZhangHetal.BIG-Datasetformulti-classcardiovasculardisorder. 2024.
33. PadminiKetal.CardioPrevent-XExplainableFederatedIntelligenceFramework.Proposed2026.