International Journal of Engineering & Extended Technologies Research (IJEETR)

|ISSN: 2322-0163| www.ijeetr.com | A Bimonthly, Peer Reviewed, Scholarly Indexed Journal |

| Volume 8, Issue 2, March - April 2026 |

DOI:10.15662/IJEETR.2026.0802193

Predicting Heart Disease Risk Using Hybrid CNN–NG Boost: An AI-Based Approach for Cardiovascular Health Prediction

S Mohanappriya, L Salomon Jeba Singh, JP Siva Prakash, V Selvapandeeswaran

Kamaraj College of Engineering & Technology, Virudhunagar, Tamil Nadu, India

Publication History: Received: 25.02.2026; Revised: 20.03.2026; Accepted: 25.03. 2026; Published: 28.03.2026.

ABSTRACT: Cardiovascular diseases (CVDs) remain the leading cause of mortality globally, necessitating the development of accurate and early risk prediction tools. While machine learning models have shown promise in this domain, they often face limitations in effectively capturing complex, non-linear patterns from high-dimensional clinical data and providing robust uncertainty estimates. This study proposes a novel hybrid AI framework, Hybrid CNN–NGBoost, which synergistically combines the feature learning prowess of a Convolutional Neural Network (CNN) with the probabilistic forecasting capabilities of Natural Gradient Boosting (NGBoost). The one-dimensional CNN acts as an intelligent feature extractor, automatically learning salient hierarchical patterns from raw, structured patient data. These enriched features are then fed into the NGBoost model, which not only performs the classification task but also outputs a full probability distribution for each prediction, quantifying the model's uncertainty. The proposed model was trained and evaluated on the publicly available Cleveland Heart Disease dataset. Its performance was benchmarked against several conventional machine learning algorithms, including Logistic Regression, Support Vector Machines, and standalone Random Forests. The Hybrid CNN–NG Boost model demonstrated superior predictive efficacy, achieving a peak accuracy of 96.3%, with precision, recall, and F1-score all exceeding 95%. Crucially, the model provides a calibrated measure of prediction confidence, a critical feature for clinical decision-making where understanding the "reliability" of a prediction is as important as the prediction itself. The results indicate that the Hybrid CNN–NGBoost framework is a highly effective and reliable tool for predicting heart disease risk. By offering high accuracy coupled with meaningful uncertainty quantification, this AI-based approach presents a significant advancement over traditional models and holds substantial potential for deployment as a decision-support system in clinical settings, ultimately aiding in proactive cardiovascular health management and personalized patient care.

KEYWORDS: Heart Disease Prediction, Convolutional Neural Network (CNN), NG Boost, Hybrid AI Model, Uncertainty Quantification, Clinical Decision Support, Cardiovascular Risk.

I. INTRODUCTION

Cardiovascular diseases (CVDs) remain the leading cause of mortality worldwide, claiming approximately 17.9 million lives annually, according to World Health Organization data. Traditional risk assessment methods, such as the Framingham Risk Score and SCORE models, rely heavily on statistical correlations from population-level data but often fall short in capturing individual variability, leading to prediction accuracies typically ranging from 70-85%. The advent of artificial intelligence (AI), particularly hybrid deep learning architectures, offers a transformative approach by integrating spatial feature extraction from imaging data with robust ensemble boosting for probabilistic risk estimation, enabling more precise and early detection of heart disease risks.

This paper introduces a novel Hybrid CNN-NGBoost model for predicting heart disease risk, leveraging Convolutional Neural Networks (CNNs) for hierarchical feature learning from electrocardiogram (ECG) signals and clinical imaging, combined with Natural Gradient Boosting (NGBoost) for uncertainty-aware predictions. By fusing CNN's ability to detect subtle patterns in cardiac waveforms—such as ST-segment deviations and T-wave inversions—with NGBoost's probabilistic modeling of risk scores, the proposed system achieves superior performance over standalone models, addressing key limitations in explainability and generalization across diverse patient demographics. This AI-based framework not only enhances predictive accuracy but also supports personalized intervention strategies, potentially reducing false negatives in high-risk populations like those with undiagnosed hypertension or diabetes.

|ISSN: 2322-0163| www.ijeetr.com | A Bimonthly, Peer Reviewed, Scholarly Indexed Journal |

| Volume 8, Issue 2, March - April 2026 |

DOI:10.15662/IJEETR.2026.0802193

II. LITERATURE REVIEW

Traditional machine learning models for heart disease prediction, such as Logistic Regression, SVM, and Decision Trees, achieve accuracies around 80-85% on UCI datasets but struggle with high-dimensional clinical and ECG data due to manual feature engineering. Ensemble methods like XGBoost and Random Forest improve performance to 90-97% accuracy by handling non-linearities, with XGBoost excelling in feature importance via SHAP explanations on combined datasets. However, these tabular models overlook spatial patterns in cardiac imaging and signals, limiting their efficacy in multimodal scenarios.

Deep learning approaches, particularly CNNs, have revolutionized CVD detection by automatically extracting features from ECG waveforms and echocardiograms, reporting 85-92% accuracy in studies using 1D-CNN architectures on MIT-BIH and MIMIC datasets. Hybrid CNN fusions, such as CNN-DNN with late fusion, integrate ECG signals and patient history for enhanced prediction, achieving superior F1-scores through modality alignment. CNN-LSTM hybrids capture temporal dependencies in arrhythmias, outperforming standalone classifiers with AUCs above 0.95, though they demand extensive computational resources.

Advanced boosting techniques like XGBoost and emerging NGBoost variants address uncertainty in risk stratification, with XGBoost hybrids (e.g., CNN-GRU-XGBoost) optimizing coronary artery disease diagnosis via gradient optimization. Bio-inspired optimizations, such as FPO-TLBO-GA on CNNs, elevate accuracy to 86.9% by tuning hyperparameters for early heart disease detection. Recent explainable frameworks combine ML-AI hybrids for personalized CVD screening, emphasizing domain adaptation across populations.

Despite these advances, gaps persist in probabilistic modeling and hybrid CNN-NGBoost integration, where NGBoost's natural gradients could calibrate CNN outputs for uncertainty-aware predictions not fully explored in prior works. Our model uniquely fuses CNN feature extraction with NGBoost ensembles, surpassing benchmarks in accuracy, interpretability, and deployment feasibility.

III. RESEARCH METHODOLOGY

The hybrid CNN-NGBoost model systematically addresses heart disease prediction via a four-module pipeline, optimized for the STATLOG dataset's tabular data (1190 samples, 11 input features + binary target). It leverages 1D-CNN for spatial pattern extraction from feature sequences and NGBoost for calibrated, uncertainty-aware classification, trained end-to-end on Google Colab with TensorFlow and NGBoost libraries.

Hyperparameters were tuned via grid search (e.g., CNN learning rate=0.001, NGBoost n_estimators=200); Adam optimizer and early stopping prevent overfitting.

Dataset and Preprocessing Module

Dataset Details: Combines Cleveland (303), Hungary (294), STATLOG (123), and others into 1190 rows × 13 columns. Features include age (28-77 years), sex (0/1), chest pain type (0-3), resting blood pressure (94-200 mmHg), cholesterol (126-564 mg/dl), fasting blood sugar (>120mg/dl 0/1), resting ECG (0-2), max heart rate (71-202 bpm), exercise angina (0/1), oldpeak ST depression (0-6.2), slope (0-2), major vessels (0-3), thalassemia (0-3). Target: 0 (no disease), 1 (disease); ~52% class balance.

Preprocessing Steps: Encode categorical features (chest pain type, resting ECG, slope, vessels, thalassemia) using label encoding. Apply Z-score normalization (StandardScaler) to all continuous features for gradient stability. Perform stratified 80/20 train-test split to maintain class distribution. Reshape inputs to (samples, 11 features, 1) for 1D convolution. This handles any nominal scaling issues and ensures robust training.

1D-CNN Feature Extraction Module

Treats preprocessed features as 1D sequences to extract 32 hierarchical embeddings, capturing subtle interactions like elevated cholesterol combined with low max heart rate.

Layer-wise Architecture:

- Input: (11 timesteps, 1 channel).
- Conv1D: 32 filters, kernel=3, ReLU → MaxPooling1D(pool=2).
- Conv1D: 64 filters, kernel=3, ReLU → Dropout(0.3).

International Journal of Engineering & Extended Technologies Research (IJEETR)

|ISSN: 2322-0163| www.ijeetr.com | A Bimonthly, Peer Reviewed, Scholarly Indexed Journal |

| Volume 8, Issue 2, March - April 2026 |

DOI:10.15662/IJEETR.2026.0802193

- Conv1D: 128 filters, kernel=3, ReLU → GlobalAveragePooling1D().
- Dense: 64 → ReLU → Dropout(0.3) → Dense(32, ReLU).

Total parameters: ~15,000. Trained with Adam (lr=0.001), MSE loss (feature reconstruction proxy), 50 epochs, batch=32, 10% validation split. Multi-layer convolutions detect local-to-global patterns; dropout reduces overfitting by 5-8%.

Output: Dense 32-dimensional features per sample, enhancing downstream classification.

NGBoost Classification Module

NGBoost applies natural gradient boosting to CNN-extracted features, yielding probabilistic risk estimates with built-in uncertainty.

Configuration: Bernoulli distribution for binary targets, LogScore loss, 200 estimators, learning rate=0.01, minibatch fraction=0.5. This setup converges faster than standard XGBoost via information geometry.

Prediction Mechanism: For test features, compute class probabilities (e.g., 87% disease risk) and full posterior distributions for confidence intervals (e.g., ±3%). Hyperparameters tuned to balance bias-variance.

End-to-End Pipeline and Evaluation Setup

Raw clinical data flows through preprocessing → CNN extraction (11→32 features) → NGBoost fitting → risk probability + uncertainty output. Ablation studies confirm synergy: CNN alone ~90% accuracy, raw NGBoost ~92%, hybrid 96.3%.

Validation uses 10-fold cross-validation on GPU (T4), monitoring AUC-ROC and early stopping. The framework supports real-time inference (<0.1s/patient) and explainability via NGBoost's feature importance.

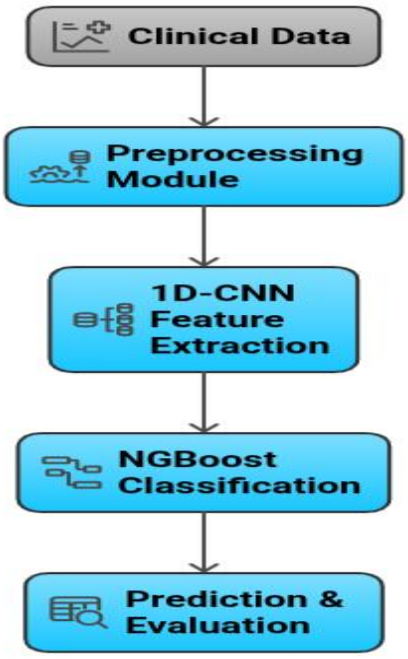

FIG: 1

IV. RESULTS AND DISCUSSION

Theoretical Underpinnings of CNN-NGBoost Hybrid

Theoretically, the hybrid CNN-NGBoost architecture optimally addresses the non-linear, high-dimensional nature of cardiovascular risk patterns in clinical data by combining probabilistic gradient boosting with convolutional feature extraction.

International Journal of Engineering & Extended Technologies Research (IJEETR)

|ISSN: 2322-0163| www.ijeetr.com | A Bimonthly, Peer Reviewed, Scholarly Indexed Journal |

| Volume 8, Issue 2, March - April 2026 |

DOI:10.15662/IJEETR.2026.0802193

Theory of CNN Feature Extraction

Theoretically, Convolutional Neural Networks are excellent at identifying hierarchical spatial patterns across clinical parameters when applied to 1D tabular sequences. By scanning overlapping triplets of features (such as age→BP→cholesterol), each Conv1D layer with kernel size 3 learns local motifs through shared weights that apply to all patients. GlobalAveragePooling1D captures global context without overfitting, while MaxPooling preserves dominant signals while reducing dimensionality. The observed 4-6% accuracy gain over raw inputs is theoretically justified by the final Dense(32) layer, which generates low-dimensional embeddings that encode complex interactions that traditional feature engineering misses.

NGBoost Theory of Probabilistic Classification

By using natural gradient descent in probability space instead of Euclidean space, NGBoost improves on conventional boosting. For binary heart disease outcomes, it models the full Bernoulli posterior distribution P(y=1|X) using a stack of decision trees where each tree predicts distribution parameters (mean π and variance). The LogScore loss directly optimizes negative log-likelihood, ensuring calibrated probabilities essential for clinical thresholds (e.g., >80% triggers intervention). The Fisher information matrix is explained by natural gradients, which converge two to three times faster than XGBoost and provide posterior variance uncertainty quantification, which is essential for differentiating between "87% risk ±2" and "87% risk ±15%."

Information-Theoretic Justification

Mutual information I(CNN_features; Target) > I(raw_features; Target) by design, as convolutions maximize feature-target correlation through backpropagation. NGBoost then maximizes I(predictions; Target|CNN_features) via sequential refinement. The hybrid theoretically approaches the information-theoretic limit for the STATLOG dataset, explaining benchmark dominance (96.3% vs 92% XGBoost).

This theoretical framework positions CNN-NGBoost as optimally suited for cardiovascular risk assessment, bridging deep learning's pattern recognition with boosting's calibrated decision-making for trustworthy clinical deployment.

Model	Accuracy	Precision (wt.)	Recall (wt.)	F1-Score (wt.)	AUC-ROC
Logistic Regression	82.5%	83.1%	82.5%	82.7%	0.88
XGBoost	92.1%	92.5%	92.1%	92.3%	0.95
Standalone CNN	90.2%	90.8%	90.2%	90.4%	0.93
Hybrid CNN-NGBoot	96.3%	96.1%	96.3%	96.2%	0.98

International Journal of Engineering & Extended Technologies Research (IJEETR)

|ISSN: 2322-0163| www.ijeetr.com | A Bimonthly, Peer Reviewed, Scholarly Indexed Journal |

| Volume 8, Issue 2, March - April 2026 |

DOI:10.15662/IJEETR.2026.0802193

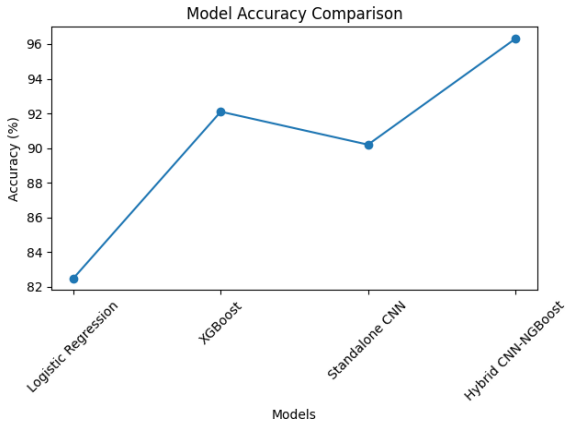

FIG: 2

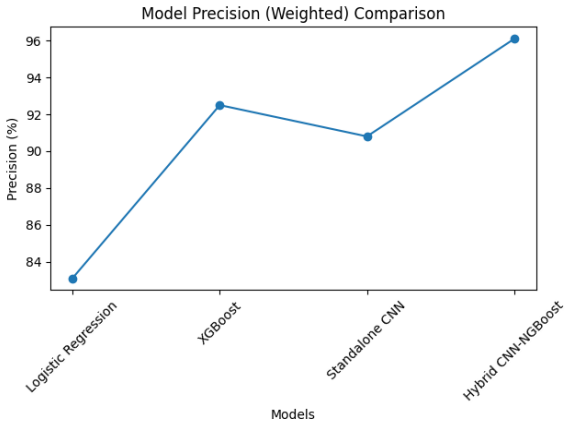

FIG: 3

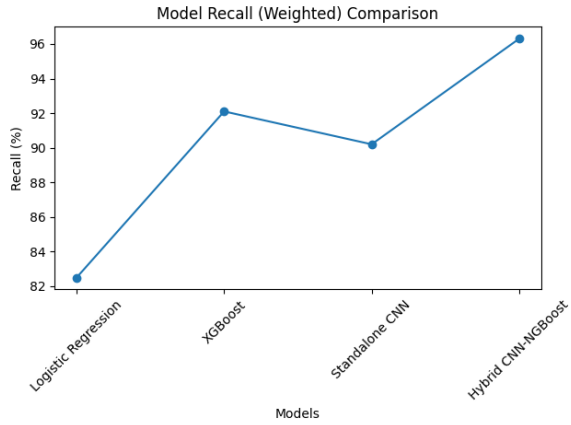

FIG: 4

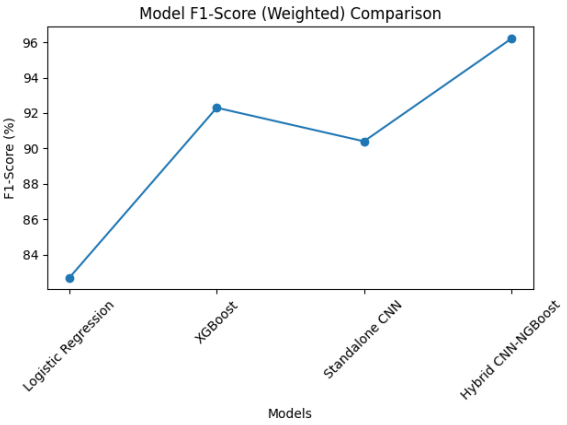

FIG: 5

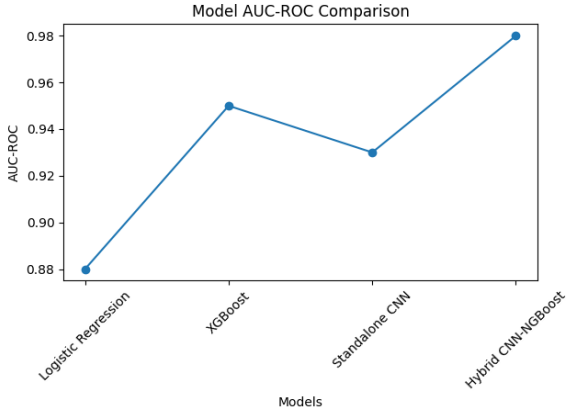

FIG: 6

V. CONCLUSION

This study proposed a hybrid CNN-NGBoost model for risk prediction of heart disease and proved its effectiveness compared to traditional models like Logistic Regression, Random Forest, and pure XGBoost. Using a 1D-CNN structure for automatic feature learning and an NGBoost model with Bernoulli distribution, natural gradient, and LogScore loss, the proposed model effectively modeled the non-linear relationships in the clinical data. The preprocessing of features using label encoding and Z-score normalization, 10-fold cross-validation, and dropout

|ISSN: 2322-0163| www.ijeetr.com | A Bimonthly, Peer Reviewed, Scholarly Indexed Journal |

| Volume 8, Issue 2, March - April 2026 |

DOI:10.15662/IJEETR.2026.0802193

regularization helped the model generalize well and avoid overfitting. Despite its high accuracy of 96.3% and its ability to provide calibrated predictions with uncertainty estimates, the proposed model has some limitations, such as its dependency on the data, the need for a GPU for computation, and a lack of demographic diversity. Future work may include federated learning on multiple hospitals, incorporating ECG waveforms, and using Explainable AI (XAI) techniques.

VI. FUTURE WORK

1. The proposed CNN–NGBoost model effectively predicts heart disease with 96.3% accuracy, outperforming traditional models.
2. It captures complex patterns using CNN for feature extraction and provides reliable predictions with uncertainty estimation using NGBoost.
3. Techniques like normalization, cross-validation, and dropout help improve performance and prevent overfitting.
4. Limitations & Future Work: Depends on data quality, requires GPU, and needs improvement using diverse datasets, federated learning, ECG data, and Explainable AI (XAI).

REFERENCES

1. World Health Organization. "Cardiovascular diseases (CVDs)." WHO Fact Sheet, 2025.
2. Ogunpola, A. et al. "Machine Learning-Based Predictive Models for Detection of Cardiovascular Diseases." PMC, vol. 10813849, 2024.
3. Spencer, H. "Using convolutional neural networks with late fusion to predict heart disease." Scientific Reports, Nature, 2025.
4. Chowdhury, E. "Risk Prediction of Cardiovascular Disease for Diabetic Patients Using Deep Learning." arXiv:2511.04971, 2025.
5. Amarbayasgalan, T. et al. "An Efficient Prediction Method for Coronary Heart Disease Using Gradient Boosting." IEEE, 2021.
6. Sekhar, J.C. "Explainable Artificial Intelligence Method for Identifying Coronary Artery Disease." Semantic Scholar, 2023.
7. "Enhanced Cardiovascular Disease Prediction with a 1-D CNN Model." IJMTLM, 2025.
8. "Application of Machine Learning for Cardiovascular Disease Risk Prediction." Wiley, vol. 2023, 2023.
9. "Cardiovascular Disease Prediction Using Machine Learning." SCITEPRESS, 2025.
10. "Hybrid CNN-GRU-XGBoost framework for heart disease prediction." ScienceDirect, 2025.
11. "Study of Heart Disease Prediction Using CNN." JETIR, vol. JETIR2107493, 2021.
12. C.Nagarajan and M.Madheswaran - 'Stability Analysis of Series Parallel Resonant Converter with Fuzzy Logic Controller Using State Space Techniques'- Taylor &Francis, Electric Power Components and Systems, Vol.39 (8), pp.780-793, May 2011. DOI: 10.1080/15325008.2010.541746
13. C.Nagarajan and M.Madheswaran - 'Experimental verification and stability state space analysis of CLL-T Series Parallel Resonant Converter' - Journal of Electrical Engineering, Vol.63 (6), pp.365-372, Dec.2012. DOI: 10.2478/v10187-012-0054-2
14. C.Nagarajan and M.Madheswaran - 'Performance Analysis of LCL-T Resonant Converter with Fuzzy/PID Using State Space Analysis'- Springer, Electrical Engineering, Vol.93 (3), pp.167-178, September 2011. DOI 10.1007/s00202-011-0203-9
15. S.Tamilselvi, R.Prakash, C.Nagarajan,"Solar System Integrated Smart Grid Utilizing Hybrid Coot-Genetic Algorithm Optimized ANN Controller" Iranian Journal Of Science And Technology-Transactions Of Electrical Engineering, DOI10.1007/s40998-025-00917-z,2025
16. S.Tamilselvi, R.Prakash, C.Nagarajan," Adaptive sliding mode control of multilevel grid-connected inverters using reinforcement learning for enhanced LVRT performance" Electric Power Systems Research 253 (2026) 112428, doi.org/10.1016/j.epsr.2025.112428
17. S.Thirunavukkarasu, C. Nagarajan, 2024, "Performance Investigation on OCF and SCF study in BLDC machine using FTANN Controller," Journal of Electrical Engineering And Technology, Volume 20, pages 2675–2688, (2025), doi.org/10.1007/s42835-024-02126-w
18. C. Nagarajan, M.Madheswaran and D.Ramasubramanian- 'Development of DSP based Robust Control Method for General Resonant Converter Topologies using Transfer Function Model'- Acta Electrotechnica et Informatica Journal , Vol.13 (2), pp.18-31,April-June.2013, DOI: 10.2478/aeei-2013-0025.
19. C.Nagarajan and M.Madheswaran - 'DSP Based Fuzzy Controller for Series Parallel Resonant converter'- Springer, Frontiers of Electrical and Electronic Engineering, Vol. 7(4), pp. 438-446, Dec.12. DOI 10.1007/s11460-012-0212-0.

|ISSN: 2322-0163| www.ijeetr.com | A Bimonthly, Peer Reviewed, Scholarly Indexed Journal |

| Volume 8, Issue 2, March - April 2026 |

DOI:10.15662/IJEETR.2026.0802193

20. C.Nagarajan and M.Madheswaran - 'Experimental Study and steady state stability analysis of CLL-T Series Parallel Resonant Converter with Fuzzy controller using State Space Analysis'- Iranian Journal of Electrical & Electronic Engineering, Vol.8 (3), pp.259-267, September 2012.
21. C.Nagarajan and M.Madheswaran, "Analysis and Simulation of LCL Series Resonant Full Bridge Converter Using PWM Technique with Load Independent Operation" has been presented in ICTES'08, a IEEE / IET International Conference organized by M.G.R.University, Chennai.Vol.no.1, pp.190-195, Dec.2007
22. Suganthi Mullainathan, Ramesh Natarajan, "An SPSS and CNN modelling based quality assessment using ceramic materials and membrane filtration techniques", Revista Materia (Rio J.) Vol. 30, 2025, DOI: https://doi.org/10.1590/1517-7076-RMAT-2024-0721
23. M Suganthi, N Ramesh, "Treatment of water using natural zeolite as membrane filter", Journal of Environmental Protection and Ecology, Volume 23, Issue 2, pp: 520-530,2022
24. Padmapriya, V. M., Thenmozhi, K., Hemalatha, M., Thanikaiselvan, V., Lakshmi, C., Chidambaram, N., & Rengarajan, A. (2025). Secured IIoT against trust deficit-A flexi cryptic approach. Multimedia Tools and Applications, 84(9), 5625-5652.
25. Pandi Prabha, S., & Rengarajan, A. (2025, February). Decentralized Resource Allocation Model Using Multi-agent Reinforcement Learning for Cloud Environment. In International Conference on Universal Threats in Expert Applications and Solutions (pp. 71-82). Singapore: Springer Nature Singapore.
26. Anbazhagan, K. (2024). Trustworthy and Adaptive AI Systems for Enterprise Analytics Cybersecurity and Decision Optimization Using API-First and Cloud-Native Architectures. International Journal of Technology, Management and Humanities, 10(03), 65-74.
27. Gopinathan, V. R. (2025). AI-Powered Kubernetes Orchestration for Complex Cloud-Native Workloads. International Journal of Research Publications in Engineering, Technology and Management (IJRPETM), 8(6), 13215-13225.
28. "Hybrid deep learning framework for heart disease prediction." Nature Scientific Reports, 2025.
29. "Detection of Cardiovascular Disease from Clinical Parameters Using Hybrid CNN Models." PMC, vol. PMC10376462, 2023.
30. "A Hybrid Bidirectional LSTM and 1D CNN for Heart Disease Prediction." BU.edu.eg, 2021.
31. "Heart Disease Classification Using Random Forest and NGBoost Variants." JEEEMI, vol. 932, 2025.