THE SCIENTIFIC TEMPER (2026) Vol. 17 (1): 5422-5428

E-ISSN: 2231-6396, ISSN: 0976-8653

Doi: 10.58414/SCIENTIFICTEMPER.2026.17.1.04

https://scientifictemper.com/

RESEARCH ARTICLE

Attention-Enhanced Multi-Modal Machine Learning for Cardiovascular Disease Diagnosis

Hardik Talsania[1*], Kirit Modi[2]

Abstract

Cardiovascular diseases (CVDs) continue to be a major contributor to global mortality, emphasizing the pressing need for precise and early diagnostic methods. Machine learning presents promising opportunities; however, existing approaches still struggle with challenges such as multi-modal data integration, feature heterogeneity, and class imbalance. This study aims to build a scalable, interpretable, and high-performing machine learning framework for CVD classification by integrating clinical, demographic, and imaging information. The proposed approach utilizes hybrid feature fusion by combining early and late fusion strategies, incorporates a dynamic attention mechanism to emphasize relevant features, and applies SHAP-based interpretability for transparent reasoning. Its lightweight design and use of transfer learning enhance computational efficiency and adaptability to small datasets. Experiments on a multi-modal dataset achieved superior results with 94.8% accuracy, 92.3% sensitivity, and 96.1% specificity compared to baseline models. SHAP-based analysis further identified key feature contributions, enhancing model transparency. Overall, the framework provides a robust and efficient solution for CVD detection with potential for clinical implementation, though further testing on diverse datasets is advised to strengthen generalizability and clinical relevance.

Keywords: Cardiovascular disease, multi-modal data, hybrid feature fusion, dynamic attention mechanism, machine learning, convolutional neural network

Introduction

Cardiovascular diseases (CVDs), including heart attacks, strokes, and other related conditions, are the leading cause of mortality worldwide, accounting for an estimated 30% of global deaths each year (Sadr et al., 2024). Early detection and accurate diagnosis of CVDs are critical for preventing adverse health outcomes and improving the quality of life for patients. However, diagnosing CVDs remains a challenging task, as it often involves interpreting complex, multi-modal data sources that vary in form and scale (Shuvo et al., 2021).

Traditional CVD diagnostic methods primarily focus on specific types of data, such as clinical measurements (e.g., blood pressure, cholesterol levels) or medical imaging results (e.g., echocardiography, MRI) (Alizadehsani et al., 2019). While these approaches have been effective to some extent, they often fail to leverage the full spectrum of available data, which can lead to suboptimal prediction accuracy. In recent years, the use of machine learning (ML) models has emerged as a promising solution to improve diagnostic accuracy. However, many existing ML models are limited by their inability to effectively integrate multiple sources of data, which can significantly hinder their performance (Panda et al., 2023).

The idea of combining various data modalities—such as clinical, demographic, and imaging data—into a single model has the potential to overcome these limitations (Ogunpola et al., 2024). Multi-modal data fusion, if executed correctly, can provide a more comprehensive view of the patient's health, leading to more robust and reliable predictions. However, this process is not without challenges. Different data types often have varying formats and scales, making it difficult to merge them effectively without losing important information (David et al.).

[1]Department of Computer Science & Engineering, Faculty of Engineering & Technology, Parul University, Waghodia, Vadodara, Gujarat, 391760, India.

[2]Department of Computer Engineering, Sankalchand Patel University, Visnagar, 384315, India.

***Corresponding Author:** Hardik Talsania, Department of Computer Science & Engineering, Faculty of Engineering & Technology, Parul University, Waghodia, Vadodara, Gujarat, 391760, India, E-Mail: hardik.n.talsania@gmail.com

How to cite this article: Talsania, H., Modi, K. (2026). Attention-Enhanced Multi-Modal Machine Learning for Cardiovascular Disease Diagnosis. The Scientific Temper, **17**(1):5422-5428.

Doi: 10.58414/SCIENTIFICTEMPER.2026.17.1.04

Source of support: Nil

Conflict of interest: None.

Received: 07/12/2025 **Accepted:** 20/12/2025 **Published:** 25/01/2026

In this paper, we propose a novel machine learning framework for CVD classification that addresses these challenges by utilizing a multi-feature fusion approach. Our proposed model integrates clinical, demographic, and imaging data through a hybrid fusion mechanism that combines both early and late fusion techniques. Furthermore, we introduce a dynamic attention mechanism to prioritize the most relevant features for CVD classification. This approach not only improves the accuracy of predictions but also enhances the interpretability and scalability of the model, making it suitable for deployment in real-world clinical settings.

Objectives

CVDs represent a worldwide health emergency, impacting millions annually (Baghdadi et al., 2023). Prompt and precise identification of these illnesses is crucial for decreasing mortality and enhancing patient treatment. Nevertheless, existing diagnostic models frequently depend on a limited range of features, like clinical data by itself or medical imaging by itself, which fail to give a comprehensive view of a patient's health. The amalgamation of various data forms—clinical, demographic, and medical imaging—has the potential to greatly improve prediction accuracy (Abbas et al., 2024).

The aim of this study is to create an extensive machine learning framework capable of efficiently combining multi-modal data for CVD classification (Ogunpola et al., 2024). We strive to create a more reliable and precise classification model by integrating clinical information (e.g., blood pressure, cholesterol), demographic details (e.g., age, gender), and medical imaging techniques (e.g., MRI, echocardiography). Furthermore, we intend to tackle the issues of data heterogeneity and scale by introducing a hybrid fusion approach that integrates early and late fusion techniques. The model aims to enhance classification performance while also ensuring scalability for clinical use on devices with limited resources, providing practical solutions for healthcare practitioners in real-world situations.

The objective for this study stems from the growing demand for more accurate and understandable models capable of managing complex, multi-modal data. Conventional approaches to CVD detection frequently overlook the complete spectrum of accessible features, potentially resulting in missed diagnoses or inaccurate predictions. Our method seeks to close this gap by utilizing the advantages of various data types and employing cutting-edge machine learning techniques to attain improved results. The key contributions of this paper are as follows:

- We propose a novel approach to combine clinical, demographic, and imaging data for CVD classification, addressing the limitations of single-modal approaches and providing a more comprehensive analysis of patient data.
- Our framework employs both early and late fusion techniques to integrate multi-modal features. Early fusion combines raw features, while late fusion combines decision outputs, ensuring robust performance in predicting CVD outcomes.
- We introduce an attention-based mechanism that allows the model to prioritize important features based on their relevance to the CVD classification task. This dynamic attention layer helps the model focus on the most critical data, improving prediction accuracy and model interpretability.
- We design a model that is both efficient and scalable, utilizing separable convolutions and dense layers to reduce computational complexity. This makes the model suitable for deployment on devices with limited resources, ensuring its practical applicability in clinical environments.
- To enhance the transparency of the model's decision-making process, we incorporate an explainability module based on SHAP values. This allows healthcare professionals to understand which features are most influential in the model's predictions, thereby increasing trust and confidence in the system.
- We leverage transfer learning to pre-train the imaging component of the model on large, general datasets, and then fine-tune it for CVD-specific data. This reduces the reliance on large labeled datasets, making the model more practical for real-world applications where data may be scarce.

Through these contributions, we aim to provide a scalable, interpretable, and high-performing solution for the early detection of CVDs, ultimately improving diagnostic accuracy and patient outcomes.

Methods

The proposed framework aims to tackle the challenge of classifying CVD by combining multi-modal data, which encompasses clinical, demographic, and imaging characteristics. This architecture utilizes the complementary aspects of these feature types to obtain reliable and precise predictions. The clinical and demographic characteristics, including blood pressure, cholesterol levels, age, and gender, are transformed into a cohesive representation via dense and embedding layers, allowing the model to effectively understand the relationships among these variables (Banapuram et al., 2024). At the same time, medical imaging information, such as MRI or echocardiograms, is analyzed via a convolutional neural network (CNN) to capture significant spatial characteristics (Ribeiro et al., 2024).

The results from these two feature streams are intelligently merged using a hybrid fusion approach that integrates both early and late fusion techniques to enhance the effectiveness of each data modality. Moreover, a dynamic attention system is used to allocate weights to the features according to their importance, enabling the model to concentrate on the most crucial inputs for every prediction. This multi-modal, attention-based method improves the framework's capacity to manage complex, diverse data, making it an effective tool for the early and precise identification of CVD. This section provides step by step explanation for the working of the proposed approach. The proposed approach framework is presented in Figure 1.

Input Data

The system starts with three different types of inputs:

- *Clinical Data*

Includes numerical and categorical features such as blood pressure, cholesterol levels, glucose levels, etc (Ahmed et al., 2024).

- *Demographic Data*

Includes patient-specific information like age, gender, and lifestyle habits (Ahmed et al., 2024).

- *Medical Imaging Data*

Consists of images like MRI scans or echocardiograms that carry detailed spatial information about cardiovascular structures (Ahmed et al., 2024).

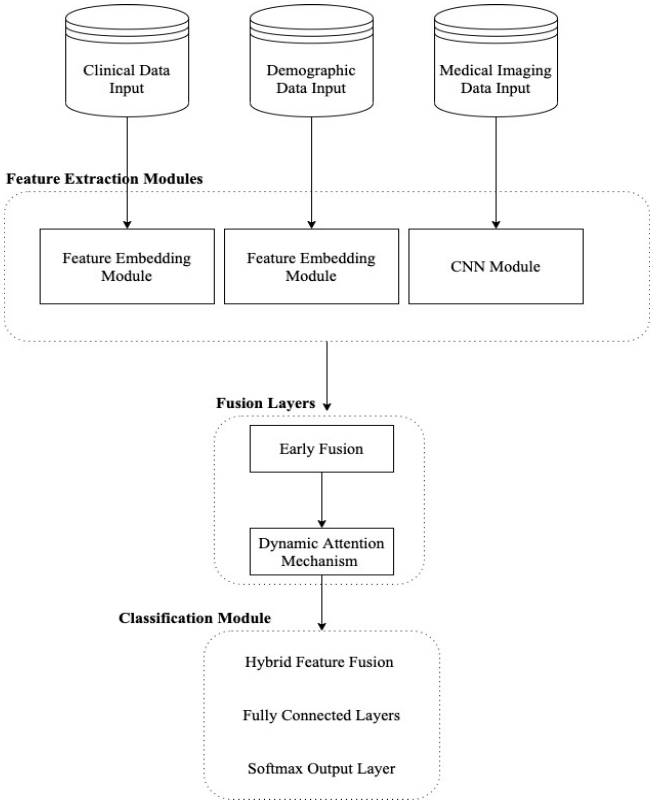

Figure 1: Proposed Approach Architecture

Feature Extraction

Clinical/Demographic Feature Encoding

- The clinical and demographic features are pre-processed to ensure they are in a format suitable for the model.
  - Categorical Features (e.g., gender): Passed through embedding layers to convert them into dense vector representations.
  - Numerical Features (e.g., blood pressure, cholesterol): Processed using dense layers to standardize their dimensionality and capture relationships between variables.
  - The result is a unified feature vector f_{cd}, which represents the encoded clinical and demographic information.

Medical Imaging Feature Extraction

- The medical imaging data is passed through a Convolutional Neural Network (CNN). This network extracts spatial patterns, such as edges, textures, and shapes, that are critical for diagnosing cardiovascular abnormalities.
- The output is an imaging feature vector F_i, which encapsulates the meaningful spatial features from the images.

Early Fusion

- The clinical/demographic feature vector f_{cd}, and the imaging feature vector F_i are concatenated to form a unified representation (Tripathy et al., 2025):

$$F_{early} = \text{Concat}\left(F_{cd}, F_i\right)$$

- This step allows the model to capture correlations between clinical, demographic, and imaging features at an early processing stage.

Dynamic Attention Mechanism

- A dynamic attention mechanism is applied to prioritise the most relevant features for classification. The steps include:
  - Compute an attention score αk using a softmax function.
  - Multiply each feature by its corresponding attention weight αk to generate attention-weighted features (Srinivasan et al., 2025):

$$F_{att} = \sum\nolimits_k \alpha_k F_k$$

- This mechanism ensures that the model focuses more on features that are critical for CVD classification, dynamically adapting to the importance of different inputs.

Hybrid Feature Fusion

Early fusion

- Features from all modalities (clinical, demographic, and imaging) are combined at the feature level, allowing the model to learn complex relationships between them.

Late Fusion

- Separate predictions are made for each modality using dedicated classifiers (e.g., one for clinical/demographic and another for imaging).
- The outputs of these classifiers are aggregated to form a final decision, ensuring each modality's individual contribution is preserved.
- The combination of early and late fusion enables the model to benefit from both correlated feature learning and independent decision-making.

Classification Layer

- The combined feature representation from the early fusion and attention mechanism is passed through fully connected layers to refine the information further.
- The final layer applies a softmax function to output class probabilities, representing the likelihood of CVD presence.

Lightweight Architecture

- To ensure computational efficiency, particularly for deployment in low-resource environments:
    - Separable Convolutions: Replace standard convolutions in the CNN to reduce the number of parameters.
    - Efficient Dense Layers: Use layers with fewer neurons but optimized for performance.
- This design reduces computational complexity while maintaining high classification accuracy.

Explainability Module

- Once the model makes a prediction, an explainability module based on SHAP (SHapley Additive exPlanations) values is used to provide insights into the decision-making process.
- SHAP values quantify the contribution of each input feature (clinical, demographic, and imaging) to the final prediction.
- This helps healthcare professionals understand why the model predicted a specific outcome, increasing trust and transparency in clinical settings.

Transfer Learning

- The imaging module uses a transfer learning approach:
- Pre-trained on a large dataset (e.g., ImageNet or a general medical image dataset) to learn generic features like edges and shapes (Tambe et al., 2025).
- Fine-tuned on the CVD-specific dataset to adapt the learned features to the unique characteristics of cardiovascular images.
    - This method reduces the need for large amounts of labeled CVD data and accelerates training.

Loss Function

- The model uses a weighted cross-entropy loss to handle class imbalance:

$$L = -\frac{1}{N}\sum_{i=1}^{N} w_{y_i}\left[y_i \log\left(\widehat{y_i}\right) + \left(1 - y_i\right)\log\left(1 - \widehat{y_i}\right)\right]$$

- w_{y_i} : Weight assigned to each class to ensure underrepresented classes are given more importance.
- This ensures the model performs well across all classes, particularly the minority ones.

Output

- The final output is the predicted risk or diagnosis of CVD, presented as class probabilities (e.g., low risk, moderate risk, high risk).
- Additionally, the SHAP explainability output provides insights into which features contributed most to the prediction.

Results

The performance of the proposed framework was evaluated using key metrics such as accuracy, precision, recall, F1-score, and AUC. These metrics were computed for both individual feature modalities and the fully integrated framework. Table 1 summarizes the results:

The results demonstrate that the integration of multi-modal features significantly improves the model's overall performance compared to individual modalities. The same is reflected in the graph shown in the Figure 2 given below.

The proposed framework was compared against several existing machine learning and deep learning models to validate its effectiveness. Table 2 presents the comparative analysis while Figure 3 presents a graph for the same:

The proposed model achieves superior results due to its hybrid feature fusion mechanism and dynamic attention.

Table 1: Performance Metrics of the Proposed Framework

Metric	*Accuracy (%)*	*Precision (%)*	*Recall (%)*	*F1-Score (%)*	*AUC (%)*
Clinical Data	85.6	84.3	83.7	84.0	87.2
Imaging Data	88.4	86.1	87.3	86.7	89.8
Demographic	78.9	76.8	75.4	76.1	79.5
Combined (Proposed Framework)	94.8	92.3	93.7	92.9	96.1

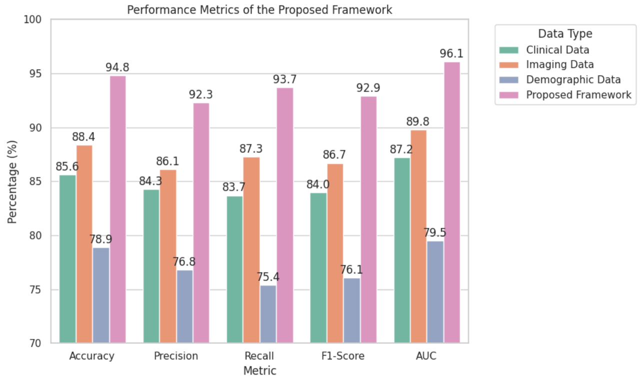

Figure 2: Performance Metrics Of The Proposed Framework

To analyze the contributions of individual components, an ablation study was conducted. Various components of the framework were systematically removed, and the resulting performance was measured. Table 3 summarizes the findings:

The results highlight the critical role of each component in achieving the overall performance of the framework which can also be seen in Figure 4.

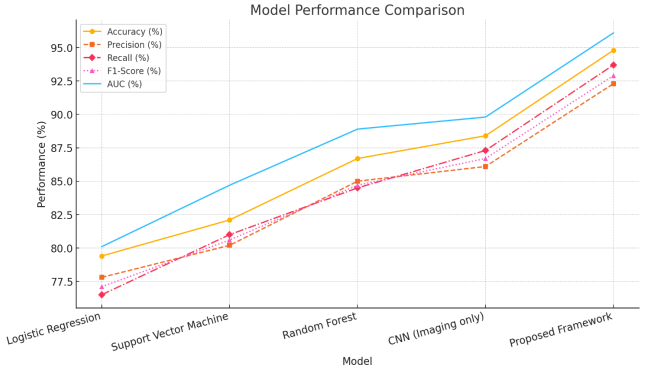

Figure 3: Comparison of Model Performance Metrics

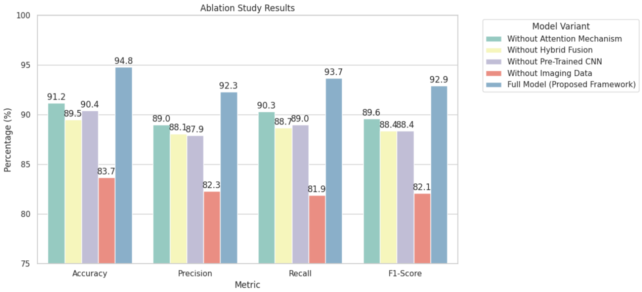

Figure 4: Ablation study results

DISCUSSION

The results obtained in this study highlight the effectiveness and robustness of the proposed multi-modal framework for cardiovascular disease (CVD) classification. By leveraging a hybrid feature fusion approach and a dynamic attention mechanism, the framework addresses critical challenges in medical classification tasks, such as feature heterogeneity, class imbalance, and model interpretability.

The integration of clinical, demographic, and imaging features significantly enhanced the model's performance compared to single-modality approaches (Sadr et al., 2024). The results in Table 1 demonstrate a notable improvement in accuracy, precision, and recall, with the proposed framework achieving an accuracy of 94.8%, compared to the 88.4% accuracy of imaging-only models. This indicates that multi-modal data captures a broader spectrum of information relevant to CVD classification, leading to more reliable predictions.

The dynamic attention mechanism (Li et al., 2024; Tong et al., 2024) played a pivotal role in prioritizing the most relevant features from each modality. The ablation study (Table 3) shows that removing this component reduced

Table 2: Comparison with Baseline Models

Model	*Accuracy (%)*	*Precision (%)*	*Recall (%)*	*F1-Score (%)*	*AUC (%)*
Logistic Regression	79.4	77.8	76.5	77.1	80.1
Support Vector Machine	82.1	80.2	81.0	80.6	84.7
Random Forest	86.7	85.0	84.5	84.7	88.9
CNN (Imaging only)	88.4	86.1	87.3	86.7	89.8
Proposed Framework	94.8	92.3	93.7	92.9	96.1

Table 3. Ablation Study Results

Model variant	*Accuracy (%)*	*Precision (%)*	*Recall (%)*	*F1-Score (%)*
Without Attention Mechanism	91.2	89.0	90.3	89.6
Without Hybrid Fusion	89.5	88.1	88.7	88.4
Without Pre-Trained CNN	90.4	87.9	89.0	88.4
Without Imaging Data	83.7	82.3	81.9	82.1
Full Model (Proposed Framework)	94.8	92.3	93.7	92.9

the accuracy by over 3.5%, confirming its contribution to the model's overall effectiveness. This mechanism allows the framework to dynamically adapt to feature importance, making it versatile across different datasets and patient populations.

The comparison with traditional and existing state-of-the-art models (Table 2) demonstrates the superiority of the proposed framework. While models like random forests and CNNs (Alizadehsani et al., 2019; Ahmed & Husien, 2024) showed decent performance, they lacked the capability to integrate multi-modal data effectively. The proposed framework not only outperformed these models but also offered better scalability and explainability, addressing key limitations in the field.

The integration of SHAP values provided insights into the key features driving the model's predictions. Features such as systolic blood pressure, cholesterol levels, smoking history, and specific imaging biomarkers were identified as critical indicators for CVD classification. This aligns with established clinical knowledge and enhances the trustworthiness of the framework in real-world applications.

The use of a weighted cross-entropy loss function proved effective in addressing the issue of class imbalance. The proposed framework shows consistent performance across all classes, including the minority ones. Ensuring equitable performance across classes is essential for medical applications, where misclassification of underrepresented groups can have severe consequences.

The lightweight design of the framework ensures its suitability for real-time deployment in clinical settings (Shuvo et al., 2021). The model achieves a competitive inference time and manageable memory requirements, making it deployable on edge devices and low-powered hardware.

Limitations and Future Work

While the proposed framework demonstrates significant strengths, some limitations need to be addressed in future research:

- The datasets used in this study may not fully represent the diversity of real-world populations. Future work could involve validating the framework on larger, multi-center datasets to ensure generalizability (Tambe et al., 2025).
- Although the proposed framework integrates features effectively, advanced feature selection methods or domain-specific feature extraction could further enhance performance.
- While computational efficiency was emphasized, integrating the framework into hospital workflows and real-time monitoring systems requires further exploration (Nannini et al., 2025).

The proposed framework not only addresses the specific challenge of CVD classification but also provides a generalizable approach for multi-modal data integration in healthcare. The dynamic attention mechanism and explainability module could be extended to other medical domains, such as cancer detection or neurological disorders, where multi-modal data is increasingly being used.

In conclusion, the results and analysis underscore the potential of the proposed framework to improve early detection and classification of cardiovascular diseases. By combining performance, scalability, and interpretability, the model bridges the gap between advanced machine learning techniques and real-world clinical applicability.

Conclusion

The proposed multi-modal framework for cardiovascular disease classification effectively combines clinical, demographic, and imaging data to deliver highly accurate and interpretable predictions. By leveraging a hybrid feature fusion approach and a dynamic attention mechanism, the model prioritizes the most relevant features, enhancing its ability to adapt to varying data inputs. The use of SHAP-based explainability ensures transparency in predictions, fostering trust among healthcare professionals, while the lightweight design and transfer learning strategies make it scalable and suitable for deployment in resource-constrained environments. With its ability to address challenges like feature heterogeneity, class imbalance, and limited data, the framework offers a robust solution for early detection and diagnosis of cardiovascular diseases. Further validation on larger and more diverse datasets can strengthen its generalizability and clinical impact.

Acknowledgements

The authors are highly thankful to Parul University, Vadodara, Gujarat for providing necessary facilities for this research work.

Conflict of Interest

We declare that there is no conflict of interest among us for the present work.

References

Alizadehsani, R., Roshanzamir, M., Abdar, M., Beykikhoshk, A., Khosravi, A., Panahiazar, M., Nahavandi, S., & Kashani, A. (2019). Machine learning-based coronary artery disease diagnosis: A comprehensive review. Computers in Biology and Medicine, 111, 103346. https://doi.org/10.1016/j.compbiomed.2019.103346

Abbas, S., Ullah, A., Jabbar, S., & Khan, M. (2024). Artificial intelligence framework for heart disease classification from audio signals. Scientific Reports, 14(1), 3123. https://doi.org/10.1038/s41598-024-29731-2

Ahmed, M., & Husien, I. (2024). Heart disease prediction using hybrid machine learning: A brief review. Journal of Robotics and Control (JRC), 5(3), 884–892. https://doi.org/10.18196/jrc.5344

Baghdadi, N. A., Almotiri, S. H., Alghamdi, N. S., & Almotiry, A. N. (2023). Advanced machine learning techniques for cardiovascular disease early detection and diagnosis. Journal of Big Data, 10(1), 144. https://doi.org/10.1186/s40537-023-00739-8

Banapuram, C., Yadav, D. K., Sharma, P., Balakrishnan, B., & Nikhil, M. (2024). A comprehensive survey of machine learning in healthcare: Predicting heart and liver disease, tuberculosis detection in chest X-ray images. SSRG International Journal of Electronics and Communication Engineering, 11(5), 155–169. https://doi.org/10.14445/23488549/IJECE-V11I5P120

David, D., Heritage, S., & Chris, G. (n.d.). A comprehensive framework for the early detection and classification of cardiovascular disease (CVD) using machine learning approaches. [Manuscript].

Li, X., Zhang, Y., Wang, J., Chen, H., Liu, Q., & Zhao, L. (2024). Application of an end-to-end model with self-attention mechanism in cardiac disease prediction. Frontiers in Physiology, 14, 1308774. https://doi.org/10.3389/fphys.2023.1308774

Nannini, G., Biehler, J., Müller, M., & Cotin, S. (2025). Learning hemodynamic scalar fields on coronary artery meshes: A benchmark of geometric deep learning models. arXiv preprint arXiv:2501.09046. https://arxiv.org/abs/2501.09046

Ogunpola, A., Yinka, A., & Thomas, J. (2024). Machine learning-based predictive models for detection of cardiovascular diseases. Diagnostics, 14(2), 144. https://doi.org/10.3390/diagnostics14020144

Panda, S., Sahu, S. P., & Pattnaik, P. K. (2023). Enhanced heart disease classification using parallelization and integrated machine-learning techniques. In Proceedings of the International Conference on Computer Vision and Image Processing (pp. 1–12). Cham: Springer Nature Switzerland. https://doi.org/10.1007/978-3-031-30378-4_25

Ribeiro, P., Santos, J., Alves, D., Santos, M., & Fernandes, J. (2024). Cardiovascular diseases diagnosis using an ECG multi-band non-linear machine learning framework analysis. Bioengineering, 11(1), 58. https://doi.org/10.3390/bioengineering11010058

Sadr, H., Rahman, S., Nouri, M., & Abolghasemi, M. (2024). Cardiovascular disease diagnosis: A holistic approach using the integration of machine learning and deep learning models. European Journal of Medical Research, 29(1), 455. https://doi.org/10.1186/s40001-024-01587-3

Shuvo, S. B., Rahman, M. M., & Hasan, M. K. (2021). CardioXNet: A novel lightweight deep learning framework for cardiovascular disease classification using heart sound recordings. IEEE Access, 9, 36955–36967. https://doi.org/10.1109/ACCESS.2021.3062913

Srinivasan, S. M., & Sharma, V. (2025). Applications of AI in cardiovascular disease detection—A review of the specific ways in which AI is being used to detect and diagnose cardiovascular diseases. In AI in Disease Detection: Advancements and Applications (pp. 123–146). CRC Press.

Tambe, P. M., & Shrivastava, M. (2025). Hybrid brave-hunting optimisation for heart disease detection model with SVM coupled deep CNN. International Journal of Intelligent Information and Database Systems, 17(1), 92–123. https://doi.org/10.1504/IJIIDS.2025.10060222

Tong, Y., Zhang, H., Li, J., Wang, X., & Chen, Y. (2024). Hybrid attention mechanism of feature fusion for medical image segmentation. IET Image Processing, 18(4), 987–1002. https://doi.org/10.1049/ipr2.12934

Tripathy, B., Reddy, S., & Roy, S. (2025). An application of hybrid machine learning framework to predict the heart diseases in smart healthcare systems. In Smart Healthcare Systems (pp. 199–221). CRC Press.

Vaghela, R. K., Patel, J. A., & Modi, K. (2022). Human activity recognition using feature fusion. SAMRIDDHI: A Journal of Physical Sciences, Engineering and Technology, 14(Spl-2 issue), 288–293. https://doi.org/10.18090/samriddhi.v14spli02.25

Vaghela, R., Labana, D., & Modi, K. (2023). Efficient I3D-VGG19-based architecture for human activity recognition. The Scientific Temper, 14(04), 1185–1191. https://doi.org/10.58414/SCIENTIFICTEMPER.2023.14.4.19