AN INTERPRETABLE CNN-TRANSFORMER FRAMEWORK FOR EARLY DETECTION OF HEART DISEASE INTEGRATING CARDIAC IMAGING WITH CLINICAL AND BIOMEDICAL RISK FACTORS

THESIS

Submitted in partial fulfillment of the requirements for the degree of

MASTER OF TECHNOLOGY
in
Computer Science & Engineering

Submitted by
Jotham N. Wanniang
24/CSE/18

Under The Supervision Of
Prof. Rajni Jindal
Professor, Department of Computer Science & Engineering



DEPARTMENT OF COMPUTER SCIENCE & ENGINEERING
DELHI TECHNOLOGICAL UNIVERSITY
(Formerly Delhi College of Engineering)
Shahbad Daulatpur, Bawana Road, Delhi-110042, India

May 2026

DEPARTMENT OF COMPUTER SCIENCE & ENGINEERING
DELHI TECHNOLOGICAL UNIVERSITY
(Formerly Delhi College of Engineering)
Bawana Road, Delhi-110042

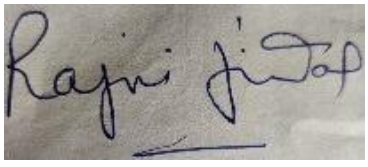
DECLARATION

I, Jotham N. Wanniang, bearing roll no. 2024/CSE/18, a student of M. Tech in Computer Science & Engineering, hereby declare that the work which is being presented in the thesis entitled "An Interpretable CNN-Transformer Framework for Early Detection of Heart Disease Integrating Cardiac Imaging with Clinical and Biomedical Risk Factors" which is submitted by me to the Department of Computer Science & Engineering, Delhi Technological University, Delhi in partial fulfillment of the requirement for the award of Degree of Master of Technology, is an authentic record of my own work carried out during the period from 2024 to 2026 under the supervision of Prof. Rajni Jindal. The matter presented in the thesis has not been submitted by me for the award of any other degree of this or any other Institute.



Jotham N. Wanniang

This is to certify that the student has incorporated all the instructions suggested by the examiners in the thesis and the statement made by the candidate is correct to the best of my knowledge.



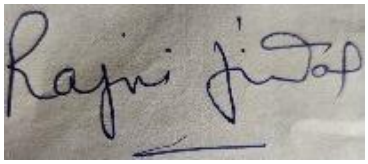
Prof. Rajni Jindal
Supervisor
Department of Computer Science & Engineering
Delhi Technological University

Signature of External Examiner

DEPARTMENT OF COMPUTER SCIENCE & ENGINEERING
DELHI TECHNOLOGICAL UNIVERSITY
(Formerly Delhi College of Engineering)
Bawana Road, Delhi-110042

CERTIFICATE

This is to certify that the Dissertation Project titled "An Interpretable CNN-Transformer Framework for Early Detection of Heart Disease Integrating Cardiac Imaging with Clinical and Biomedical Risk Factors" which is submitted by Jotham N. Wanniang, bearing roll no. 2024/CSE/18, Department of Computer Science & Engineering, Delhi Technological University, Delhi in partial fulfilment of the requirement for the award of the degree of Master of Technology, is a record of the project work carried out by the student under my supervision. To the best of my knowledge this work has not been submitted in part or full for any Degree or Diploma to this University or elsewhere.



Prof. Rajni Jindal
Supervisor
Department of Computer Science & Engineering
Delhi Technological University

DEPARTMENT OF COMPUTER SCIENCE & ENGINEERING
DELHI TECHNOLOGICAL UNIVERSITY
(Formerly Delhi College of Engineering)
Bawana Road, Delhi-110042

ACKNOWLEDGEMENT

My heartfelt thanks go to Prof. Rajni Jindal, Professor, Department of Computer Science & Engineering, Delhi Technological University for her invaluable guidance, constant encouragement and scholarly mentorship during the course of this research work.

I express my thanks to the Department of Computer Science & Engineering, Delhi Technological University for giving me computational infrastructure and academic resources to conduct my work.

I would also acknowledge the contribution of various open-source technologies and research tools, including Python, TensorFlow, Keras, PyTorch, Scikit-learn, OpenCV, SHAP, and Grad-CAM, which were utilized during the implementation and evaluation of the proposed framework.

Last but not least, I want to thank my family and my friends who have been with me all the way, encouraging and supporting me morally.



Jotham N. Wanniang

ABSTRACT

Heart diseases still constitute one of the main reasons behind mortality cases worldwide and are responsible for exerting a great deal of pressure on the health care facilities, infrastructure, and resources used in the process. In the light of rising numbers of patients affected by such heart-related illnesses as coronary artery disease, arrhythmia, myocardial infarction, and other conditions, it is vital to develop an intelligent diagnostic system that would help doctors detect diseases at their earliest stages. Traditional approaches to disease diagnostics usually imply manual analysis of results provided in medical reports, scans, and laboratory test data.

Nevertheless, advances in artificial intelligence technologies, as well as deep learning, have shown some promise in enhancing the effectiveness and accuracy of prediction models. Unfortunately, many existing healthcare prediction models operate in the realm of single modality and often feature the form of black boxes. Such approaches fail to build the trust of healthcare practitioners and, thus, cannot be adopted in real-life settings.

The above challenges have been mitigated through this work by proposing a novel multi-modal explainable deep learning paradigm that integrates Convolutional Neural Networks (CNNs) and Transformer architectures for early heart disease prediction. Specifically, the multi-modal explainable deep learning architecture exploits cardiac imaging and structured clinical and biomedical data as input data to learn a single unified framework that can learn spatial and contextual relations between diverse data in a hospital setting.

The CNN model in the proposed framework is used to extract discriminatory spatial representations from heart imaging modalities such as chest X-rays, CT scans, and MRIs. The generated feature maps are passed through a transformer model that learns self-attention mechanisms to analyze long-range dependencies and contextual interrelations among the various medical data. Moreover, the proposed multi-modal learning framework uses an intelligent strategy of integrating imaging modalities with biomedical data, which include blood pressure, cholesterol levels, electrocardiogram readings, and more.

One important contribution made by this paper includes the use of Explainable Artificial Intelligence (XAI) techniques for the better understanding of the prediction mechanism. Specifically, SHAP (SHapley Additive Explanations) method is used to determine the impact of each individual clinical parameter, whereas the Grad-CAM technique provides visual identification of the areas in images which significantly impact the decision-making process. Overall, this approach offers clinically

relevant insights which can assist doctors' decision-making process and help them trust the predictions provided by an artificial intelligence algorithm.

The performance evaluation of the proposed model is carried out using public domain cardiovascular datasets and compared to traditional machine learning algorithms as well as deep learning algorithms operating independently. It is shown experimentally that combining both approaches leads to considerable improvements in classification metrics such as accuracy, precision, recall, and F1-score. On balance, the suggested study makes significant contributions to the development of reliable intelligent health care systems which can be used to support early CVD diagnosis and other AI-related medical decision support applications.

TABLE OF CONTENTS

		PAGE
DECLARATION		i
CERTIFICATE		ii
ACKNOWLEDGEMENT		iii
ABSTRACT		iv
CHAPTER 1	INTRODUCTION	
1.1	BACKGROUND	1
1.2	MOTIVATION	1
1.3	RESEARCH GAP	2
1.4	PROBLEM STATEMENT	2
1.5	OBJECTIVES	2
CHAPTER 2	LITERATURE REVIEW	
2.1	TRADITIONAL MACHINE LEARNING APPROACHES	3
2.2	DEEP LEARNING IN MEDICAL DIAGNOSIS	3
2.2.1	CNN BASED MODELS	4
2.2.2	TRANSFORMER BASED MODELS	4
2.2.3	EXPLAINABLE ARTIFICIAL INTELLIGENCE	4
2.2.4	SHAP	5
2.2.5	GRAD-CAM	5
2.3	COMPARATIVE ANALYSIS OF MACHINE LEARNING AND DEEP LEARNING MODELS	5
CHAPTER 3	PROPOSED METHODOLOGY	
3.1	DATASET DESCRIPTION	7
3.2	DATA PREPROCESSING	8
3.2.1	IMAGE PREPROCESSING	8
3.2.2	CLINICAL DATA PREPROCESSING	9
3.2.3	DATA SPLIT	9
3.3	SYSTEM ARCHITECTURE	9
3.3.1	CNN FEATURE EXTRACTION	10
3.3.2	TRANSFORMER MODULE	11
3.3.3	MULTIMODAL FUSION	11
3.3.4	CLASSIFICATION LAYER	12
3.3.5	EXPLAINABILITY LAYER	12
CHAPTER 4	IMPLEMENTATION DETAILS	
4.1	ALGORITHM: HYBRID CNN-TRANSFORMER	14
4.1.1	INPUT	14
4.1.2	OUTPUT	14
4.1.3	PROCEDURE	14
4.2	CODE IMPLEMENTATION	15
4.2.1	IMPORTING LIBRARIES	15
4.2.2	LOADING CLINICAL DATASET	15
4.2.3	CNN MODEL FOR IMAGE FEATURE EXTRACTION	15
4.2.4	CLINICAL DATA NETWORK	16
4.2.5	MULTIMODAL FUSION LAYER	16
4.2.6	MODEL COMPILATION	16
4.2.7	MODEL TRAINING	16

4.2.8	SHAP EXPLAINABILITY	16
4.2.9	GRAD-CAM VISUALIZATION	17
4.3	SYSTEM REQUIREMENTS	17
4.3.1	SOFTWARE REQUIREMENTS	17
4.3.2	HARDWARE REQUIREMENTS	17
4.3.3	HYPERPARAMETERS	18
CHAPTER 5	RESULTS AND ANALYSIS	
5.1	PERFORMANCE AND EVALUATION METRICS	19
5.2	MODEL COMPARISON	20
5.3	ROC CURVE OF PROPOSED MODEL	20
5.4	IMPACT OF MULTIMODAL FUSION	21
5.5	EXPLAINABILITY ANALYSIS	22
5.5.1	SHAP ANALYSIS	22
5.5.2	GRAD-CAM VISUALIZATION	22
5.6	INTERPRETATION OF RESULTS	23
5.7	ADVANTAGES OF PROPOSED MODEL	23
CHAPTER 6	APPLICATION	
6.1	EARLY CARDIOVASCULAR RISK ASSESSMENT	24
6.2	AI-ASSISTED DIAGNOSIS	24
6.3	SMART HEALTHCARE SYSTEMS	24
6.4	CLINICAL DECISION SUPPORT	24
6.5	TELEMEDICINE APPLICATIONS	24
6.6	REMOTE PATIENT MONITORING	24
CHAPTER 7	LIMITATIONS	
7.1	HIGH COMPUTATIONAL COMPLEXITY	26
7.2	DEPENDANCE ON DATASET QUALITY	26
7.3	REQUIREMENT ON GPU RESOURCES	26
7.4	LIMITED AVAILABILITY OF MULTIMODAL DATASETS	26
CHAPTER 8	FUTURE WORK	
8.1	INTEGRATION OF ECG SIGNALS AND WEARABLE SENSOR DATA	27
8.2	FEDERATED LEARNING FOR PRIVACY-PRESERVING HEALTHCARE	27
8.3	LIGHTWEIGHT TRANSFORMER ARCHITECTURES	27
8.4	EDGE-AI OPTIMIZATION	27
8.5	REAL-TIME DEPLOYMENT IN CLINICAL SETTINGS	27
8.6	INTEGRATION WITH ELECTRONIC HEALTH RECORDS (EHR)	27
8.7	ADVANCED CROSS-MODAL ATTENTION MECHANISMS	28
8.8	LARGE-SCALE CLINICAL VALIDATION	28
CHAPTER 9	CONCLUSION	
9.1	SUMMARY OF WORK	29
9.2	KEY CONTRIBUTIONS	29
9.3	PERFORMANCE OBSERVATIONS	29
9.4	RESEARCH SIGNIFICANCE	29
9.5	FINAL REMARKS	29
	REFERENCES	30

LIST OF TABLES

	TITLE	PAGE
Table 1	Performance of Standalone Traditional Machine Learning Models	3
Table 2	Explainable AI and their Purpose	5
Table 3	Comparison of Machine Learning and Deep Learning Models with the Proposed Model	6
Table 4	UCI Dataset Attributes	7
Table 5	UCI Dataset Statistics	7
Table 6	NIH Dataset Statistics	8
Table 7	Division of Dataset for Training and Testing	9
Table 8	Development Tools and Libraries	17
Table 9	Software Versions	17
Table 10	Hardware Requirements	17
Table 11	Hyperparameter Configuration	18
Table 12	Proposed Model Performance and Evaluation Comparison	20
Table 13	Impact of Multimodal Fusion	21

LIST OF FIGURES

	TITLE	PAGE
Figure 1	NIH Dataset Visual Images	8
Figure 2	Proposed System Framework	10
Figure 3	CNN Architecture for Image Feature Extraction	11
Figure 4	Transformer Encoder Block	11
Figure 5	Fusion Block	12
Figure 6	Interpretability using Grad-CAM and SHAP	13
Figure 7	Proposed Model Performance and Evaluation	19
Figure 8	Proposed Model Accuracy Comparison	20
Figure 9	ROC Curve of the Proposed Model	21
Figure 10	Impact of Mutimodal Fusion	21
Figure 11	SHAP Analysis	22
Figure 12	Grad-CAM Visualization	22

CHAPTER 1: INTRODUCTION

1.1 BACKGROUND

CVD is one of the most challenging healthcare problems that we face today. As reported globally, the prevalence of these diseases is about one-third of all deaths around the world. Therefore, it becomes crucial to detect and manage these cases to save patients from dying. Common CVD cases are coronary artery disease, arrhythmia, heart failure, and myocardial infarction. In traditional diagnostic methods, expertise is needed to interpret the lab report and other medical findings like electrocardiograms, ECG, imaging, and pathological signs.

Heart disease diagnosis depends on the physician and his/her interpretation of the patient data. These procedures are labor-intensive and can be subjective at times due to the need for human expertise. With the rise in patient data, especially heterogeneous data, the problem is even more complicated.

Recent developments in AI, ML, and DL have greatly impacted the domain of medical diagnosis. These techniques help in making intelligent systems capable of recognizing inherent patterns from large data sets to make quicker and more accurate predictions. The CNN architecture has shown promising results with respect to recognizing spatial information from medical images, whereas Transformer-based algorithms have introduced robust attention mechanisms that can capture long-distance relationships and contexts.

However, there remain several deficiencies in the application of healthcare AI solutions. Current models tend to utilize only one type of input data, either medical imaging or structured data, leading to an incomplete picture of patients' conditions. Moreover, most deep learning algorithms work as "black-box" approaches without any explainability, leading to the lack of credibility in the clinical setting.

Problems encountered by contemporary healthcare AI models include the following:

i. Low capability of multimodal data incorporation
ii. Lack of explainability for deep learning methods
iii. Computational complexities
iv. Low clinical credibility
v. Segregated generalization on different datasets

As a solution to the above-mentioned problems, an interpretable multimodal approach is proposed using the combination of cardiac imaging and clinical data.

1.2 MOTIVATION

The reason for conducting this research emerges from the pressing necessity of making cardiac disease diagnosis more reliable and efficient. Indeed, prompt diagnosis is essential in preventing life-threatening consequences of such diseases and enhancing patient survival chances. Unfortunately, conventional diagnostic systems tend to lack a sufficient level of efficiency when merging information provided by several sources or providing a comprehensive explanation of predictions.

The primary motivations behind this research can be listed as follows:

i. The need for accurate diagnostics in the early stages of a disease development
ii. The necessity to ensure the interpretability of AI-based models in healthcare
iii. The requirement of incorporating both clinical and imaging data in decision-making processes
iv. The need to have effective decision-support instruments in medicine
v. The necessity of minimizing diagnostic and treatment delays

Traditionally, medical professionals use either imaging techniques or clinical data alone. Merging the two might lead to considerable improvement in predictions.

1.3 RESEARCH GAP

The review of previous work shows the following limitations present in the current methods of cardiovascular disease prediction:

i. Singularity: The current approaches depend either on clinical information or imaging information of patients, leading to incomplete information about the patient's condition.
ii. Inadequate interpretability: In most cases, the current deep learning models are black-box models.
iii. Bad contextual representation: The models based on CNNs only extract local information without considering global relations.
iv. High computational cost: Transformer-based models involve high computational costs.
v. Low generalizability: Several deep learning models exhibit poor generalization capabilities on different datasets and environments.

All these problems call for a new approach to the problem.

1.4 PROBLEM STATEMENT

The problem with the diagnosis of cardiovascular disorders lies in complicated correlations between imaging information, clinical factors, and biomedical signals. Currently used methods treat each data source separately, thus failing to provide a holistic perspective and prediction accuracy.
Moreover, existing deep learning-based approaches are usually opaque and thus lack explainability, which makes it difficult to trust their results in medical contexts. As a result, there emerges an urgent need to create a framework that will be not only accurate but also explainable.
For that reason, this study seeks to design a model that will be able to:

i. Consume various kinds of health care information
ii. Learn spatial and contextual features
iii. Provide accurate diagnosis at an early stage
iv. Provide explanations to facilitate clinical decision making

To that end, it is proposed to use a hybrid CNN-Transformer approach.

1.5 OBJECTIVES

Objectives of this study have been listed as follows:

i. Development of a CNN-Transformer based hybrid model for heart disease diagnosis
ii. Use of multimodal fusion methods for integration of cardiac imaging data and structured clinical data
iii. Capitalizing on attention based contextual learning to increase prediction accuracy
iv. Application of Explainable AI (XAI) methods to improve interpretability
v. Evaluation of the effectiveness of the proposed model
vi. Decreasing number of false predictions
vii. Facilitating better feature representations and providing visual explanations

CHAPTER 2: LITERATURE REVIEW

2.1 TRADITIONAL MACHINE LEARNING APPROACHES

Traditional machine learning methods have been utilized in predicting as well as diagnosing different types of cardiac diseases. Some traditional machine learning methods include logistic regression, decision tree, naïve bayes, random forest, and SVM. Traditional machine learning methods are majorly based on structured data, which includes demographic features, lab values, and vital signs of the patient.
Although conventional algorithms perform reasonably well in most scenarios, their success greatly hinges on the quality of input data. As opposed to deep learning networks, feature extraction by conventional algorithms relies on hand-crafted input attributes based on domain knowledge. This step may consume significant resources and hinder the algorithm's ability to identify intricate relationships in data.

A. Advantages of Traditional Machine Learning Approaches
   i. Simplicity: They are relatively easy to apply and interpret, thus ideal for initial investigations and applications.
   ii. Interpretability: Many traditional machine learning models, like Decision Trees and Logistic Regression, offer high levels of interpretability that are necessary in a medical setting where decisions need to be made based on algorithmic predictions.
   iii. Lower Requirement for Computational Resources: In comparison with deep learning methods, these approaches require less powerful hardware resources and training time.

B. Disadvantages of Traditional Machine Learning Approaches
   i. Dependence on Features: Performance of many of these approaches depends largely on hand-engineered features that do not necessarily encapsulate all information present in the data.
   ii. Weak Nonlinear Modeling: Although some of these techniques offer some level of modeling capability, such as Random Forests and SVMs with appropriate kernels, they cannot deal with highly nonlinear patterns often found in medical data.
   iii. Unable to Process Unstructured Data: Traditional approaches are unable to process unstructured data like images, thus limiting their application in multimodal healthcare environments.

Table 1: Performance of Standalone Traditional Machine Learning Models

Model	Advantages	Limitations
LeNet	Lightweight	Limited depth
AlexNet	Better feature extraction	High parameters
VGG16	Strong accuracy	Computationally expensive
ResNet50	Residual learning	Increased complexity
EfficientNet	Scalable	Difficult tuning

2.2 DEEP LEARNING IN MEDICAL DIAGNOSIS

Deep learning has greatly revolutionized the area of healthcare analytics by providing an automated way to extract features from the complex data. Deep learning algorithms are different from traditional machine learning techniques as they do not use any handcrafted features but can learn the hierarchy

from the input data. This aspect has helped greatly in making advancements in the area of medical diagnosis as the data is generally high-dimensional and complex.
Various kinds of deep learning algorithms have been developed that have aided greatly in the field of medicine. CNNs or Convolutional Neural Networks play a vital role in the medical image analysis because of their ability to extract a hierarchy of features using filters and max pooling. In addition to CNNs, transformer models are also used for capturing long-range dependencies between data elements.

2.2.1 CNN Based Models

Convolutional Neural Network (CNN) is popularly employed in medical applications due to its high-level proficiency in dealing with image data. The reason behind this is that CNN models can automatically discover the relevant features of an image without requiring any prior feature extraction process.
Some of the most appropriate uses for CNN-based models include the following:

i. Classification of Medical Images: Involving classifying images of patients into categories based on whether they are healthy or suffering from a certain ailment.
ii. Feature Detection: Discovering relevant features of an image such as edges and texture.
iii. Pattern Detection: Involving identifying complex and subtle patterns within a given image.

2.2.2 Transformer Based Models

Transformer-based models have lately surfaced as strong contenders to conventional CNN architectures for their ability to learn complex relationships from data inputs. The model depends on a self-attention mechanism that enables it to process various parts of the input at once while capturing long-term relationships.

A. Advantages of Transformer-Based Models:

i. Global Context Learning: Ability to learn relationships between data in the whole input rather than focusing on specific local regions.
ii. Effective Sequential Data Handling: Capable of handling sequential and time-series data.
iii. Strong Contextual Capability: Enables better learning of interaction between various inputs.
iv. Long-Range Relationship Capture: Helps capture relationships between distant elements of data input.

B. Disadvantages of Transformer-Based Models:

i. High Training and Computational Demand: Need high computational power for processing.
ii. Large Input Requirement: Perform better on large input samples, which may not be possible for healthcare problems.
iii. Optimization Issues: Possible optimization difficulties when using medical data sets.

2.2.3 Explainable Artificial Intelligence (XAI)

One requirement that cannot be overlooked in the field of healthcare application is that of interpretability. Explainable Artificial Intelligence (XAI) tools can help solve this requirement, as they give insight into the workings of a machine and how it arrives at certain decisions through deep learning.
This helps in building trust within clinicians.

Table 2: Explainable AI and their Purpose

Technique	Purpose
Grad-CAM	Provides visual explanations by highlighting important regions in medical images
SHAP	Quantifies the contribution of each feature to the model's prediction
LIME	Offers local interpretability by approximating model behavior around specific predictions
Attention Maps	Visualize which parts of the input the model focuses on during prediction

2.2.4 SHAP

SHAP is an interpretability tool that relies on cooperative game theory. This tool helps to explain the predictions of a machine learning model by allocating contributions for each input variable to explain how much influence the input variables have on the final result. It can provide both local and global interpretability.

2.2.5 Grad-CAM

Grad-CAM is a tool that helps in analyzing CNN models by generating heat maps which indicate the portions of an image which play an important role in making particular predictions. In medical imagery, Grad-CAM identifies abnormal regions on x-rays by highlighting them, offering an explanation as to why certain regions stand out.

2.3 COMPARATIVE ANALYSIS OF MACHINE LEARNING AND DEEP LEARNING MODELS

A comparison of different types of machine learning models and deep learning models identifies the unique strengths and weaknesses that make each type of model more suitable for predicting heart disease.

Logistic regression and support vector machines are good options for the analysis of structured data; they are highly favored for their ease of use and interpretability. Unfortunately, both methods are unable to detect complex nonlinearities and can be difficult to scale up for large datasets.

In contrast, convolutional neural networks can be considered more effective when capturing and extracting local patterns from medical imagery, which has made them the most widely used in image recognition tasks. Nonetheless, there is no guarantee that the patterns extracted by CNNs represent global context information within the data.

Transformer-based models overcome this drawback by implementing self-attention mechanisms which help identify global dependencies and context relationships among features. Even though they have powerful modeling capabilities, these models are computationally complex and need huge volumes of data for proper training.

In order to exploit the advantages of each individual technique and to overcome the associated disadvantages, a novel hybrid approach has been suggested in which CNNs are combined with transformers. This way, the model not only exploits the ability of both the techniques but also ensures explainability in its operations, thus enhancing accuracy as well as understandability of results. The model is dependent on the availability of multimodal data sets which involve imaging and clinical data.

Table 3: Comparison of Machine Learning and Deep Learning Models with the Proposed Model

Model	**Advantages**	**Limitations**
Logistic Regression	Simple	Low nonlinear capability
SVM	Robust classification	Poor scalability
CNN	Excellent image processing	Limited contextual learning
Transformer	Strong attention modeling	Computationally expensive
Proposed Hybrid Model	High accuracy + explainability	Requires multimodal data

CHAPTER 3: PROPOSED METHODOLOGY

The suggested model is developed in such a way that it would help to predict cardiovascular diseases accurately through its use of multi-modal deep learning methods. The methodology involves the following elements:

i. Dataset Collection
ii. Data Preprocessing
iii. System Architecture

3.1 DATASET DESCRIPTION

It is dependent upon the type of datasets that have been chosen. For the present study, both clinical as well as imaging datasets have been employed in order to facilitate multi-modal learning.
Proposed Framework Uses Publicly Available Cardiovascular Datasets.

A. UCI Heart Disease Dataset

The UCI Heart Disease dataset is popularly used to predict heart diseases and holds structured clinical parameters. The features are patient health indicators used for medical diagnosis purposes.

Table 4: UCI Dataset Attributes

Attribute	Description
Age	Patient age
Sex	Gender
Cholesterol	Serum cholesterol
Blood Pressure	Resting BP
Chest Pain Type	Symptom category
ECG Results	Electrocardiogram
Heart Rate	Maximum HR
Diagnosis	Presence/absence of disease

Table 5: UCI Dataset Statistics

Parameter	Value
Total Samples	303
Features	14
Classes	Binary
Missing Values	Few

B. NIH Chest X-Ray Dataset

In regard to image generation, this study will use the National Institute of Health Chest X-Rays dataset, which contains a wide range of chest radiography images from which visual patterns associated with heart disease such as cardiomegaly can be generated.

Table 6: NIH Dataset Statistics

Parameter	Value
Total Images	100,000+
Normal Images	87524
Cardiomegaly	88652
Image size	224 x 224
Image Type	X-Ray
Resolution	Variable
Format	PNG

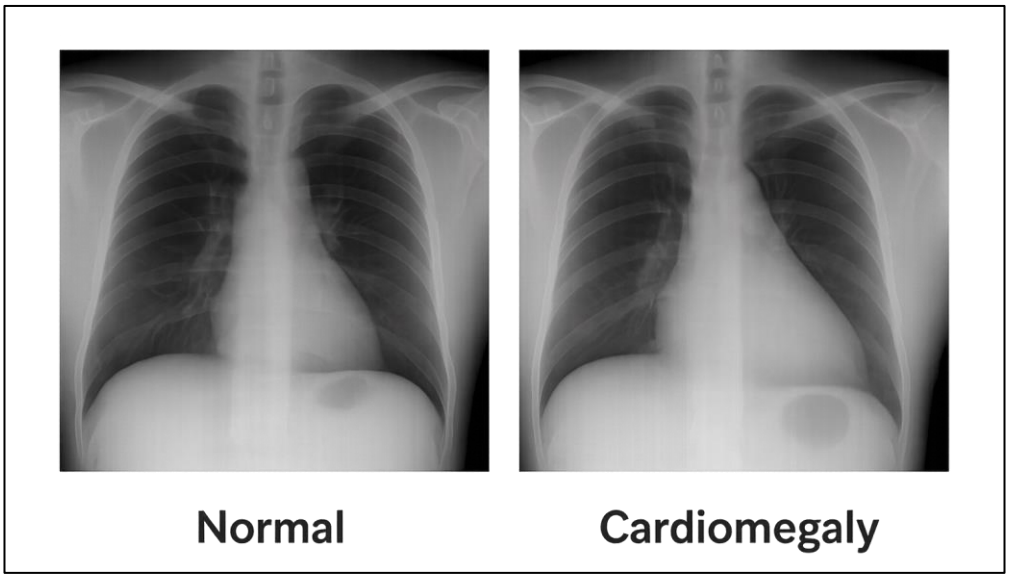


Figure 1: NIH Dataset Visual Images

3.2 DATA PREPROCESSING

In relation to image analysis for the current experiment, the NIH Chest X-ray data set was considered ideal as a result of the wide variety of chest radiographs available in it. The images in the data set have been analyzed in order to determine any visual features of cardiovascular diseases, such as cardiomegaly.

3.2.1 Image Preprocessing

Several pre-processing procedures are performed on the chest X-ray images in order to normalize the input and improve the features:

i. Resizing: The images are resized to 224 × 224 pixels for maintaining uniform size.
ii. Normalization: The values of the pixels are normalized to a standard scale to facilitate efficient convergence during the training process.
iii. Noise Reduction: Various filtering methods are used to get rid of unwanted elements from the images.
iv. Data Augmentation: Methods such as rotation, flipping, and scaling are performed to enrich the database.

3.2.2 Clinical Data Preprocessing

Preparation of the Structured Clinical Dataset for Deep Learning Model

i. Missing Value Imputation: Missing values in the dataset are addressed using statistical imputation techniques like mean or median.
ii. Feature Scaling: Scaling of numeric features to ensure that no weight is attributed to the higher magnitude of features.
iii. Normalization: Normalization of data is carried out for improved stability.
iv. One-Hot Encoding: Converting categorical features into numeric form such as types of chest pain.

Formula for Normalization:

$$X' = \frac{X - X_{\min}}{X_{\max} - X_{\min}}$$

where:

i. $X'$is the normalized value
ii. $X$is the original value of this feature
iii. $X_{\min}$and $X_{\max}$are the min and max value of the feature

3.2.3 Data Split

In order to test the efficiency and validity of the framework, the datasets were split into two different groups of data, that is, the training group and the testing group. The training set is used to allow the model to recognize any inherent patterns within it, whereas the testing dataset is employed to assess prediction capability.

Table 7: Division of Dataset for Training and Testing

Dataset	Training	Testing
Clinical Data	80%	20%
Imaging Data	80%	20%

This separation enables the model to train using more of the data set, while still maintaining an independent subset of data to evaluate the performance of the model objectively on the data set.

3.3 SYSTEM ARCHITECTURE

The multi-modal deep learning method being discussed herein seeks to integrate both imaging data and clinical data in one single model in order to diagnose heart diseases. There are various elements which make up the system, with each element playing a certain role during various stages of the procedure. These elements include:

i. Input Layer
ii. CNN-Based Feature Extraction Module
iii. Transformer-Based Contextual Learning Module
iv. Multimodal Fusion Layer
v. Classification Layer
vi. Explainability Module

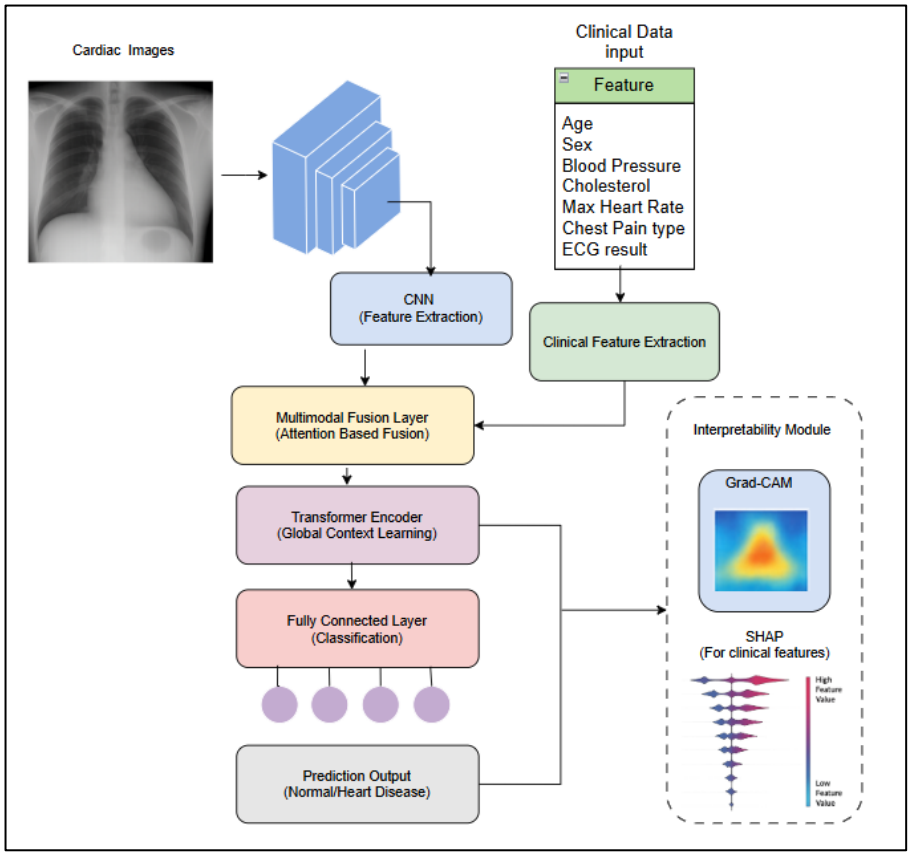


Figure 2: Proposed System Framework

3.3.1 CNN Feature Extraction

Convolutional Neural Network (CNN) module becomes the core module that helps to extract the spatial information from the input chest X-ray image. These networks are extremely effective in identifying features like edges and textures which are needed for the detection of heart disease anomalies.

It contains multiple convolutional layers, activation functions, pooling, and fully connected layers. During training, filters learn discriminative features automatically for the purpose of detecting cardiac abnormalities.

The convolution process is written as:

$$F = \sigma(W * I + b)$$

where:

i. Frepresents the output feature map
ii. W denotes learnable filter weights
iii. I represents the input image
iv. b is the bias term
v. σ denotes the activation function (ReLU)

Activation by Rectified Linear Units (ReLUs) introduces nonlinearity and helps converge more quickly. The application of max-pooling operations is for reducing the dimensions while retaining the most important features, which can be mathematically expressed as:

$$P = \max(F)$$

The extracted high-level features are subsequently flattened and fed into the multimodal fusion layer along with clinical data.

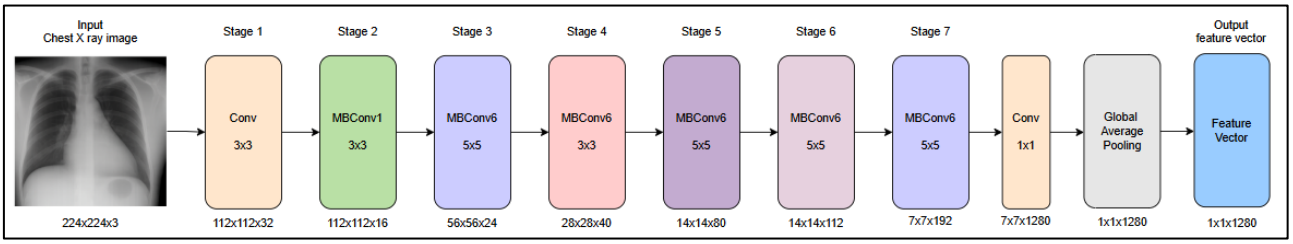


Figure 3: CNN Architecture for Image Feature Extraction

3.3.2 Transformer Module

The addition of the Transformer module is aimed at capturing relationships across the whole input since CNNs cannot properly capture these types of relationships.
A self-attention mechanism in contrast to sequential models is used in the transformer, which is capable of analyzing relationships between different features of the whole input.
The formula for computing self-attention is as follows:

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V$$

where:

i. Q represents Query vectors
ii. K represents Key vectors
iii. V represents Value vectors
iv. d_k is a scaling factor to stabilize training

The Transformer encoder applies multiple attention heads to analyze different feature interactions at the same time. Besides, positional encoding is used to preserve sequence information in the input.
The output from the Transformer module is passed to the multimodal fusion module.

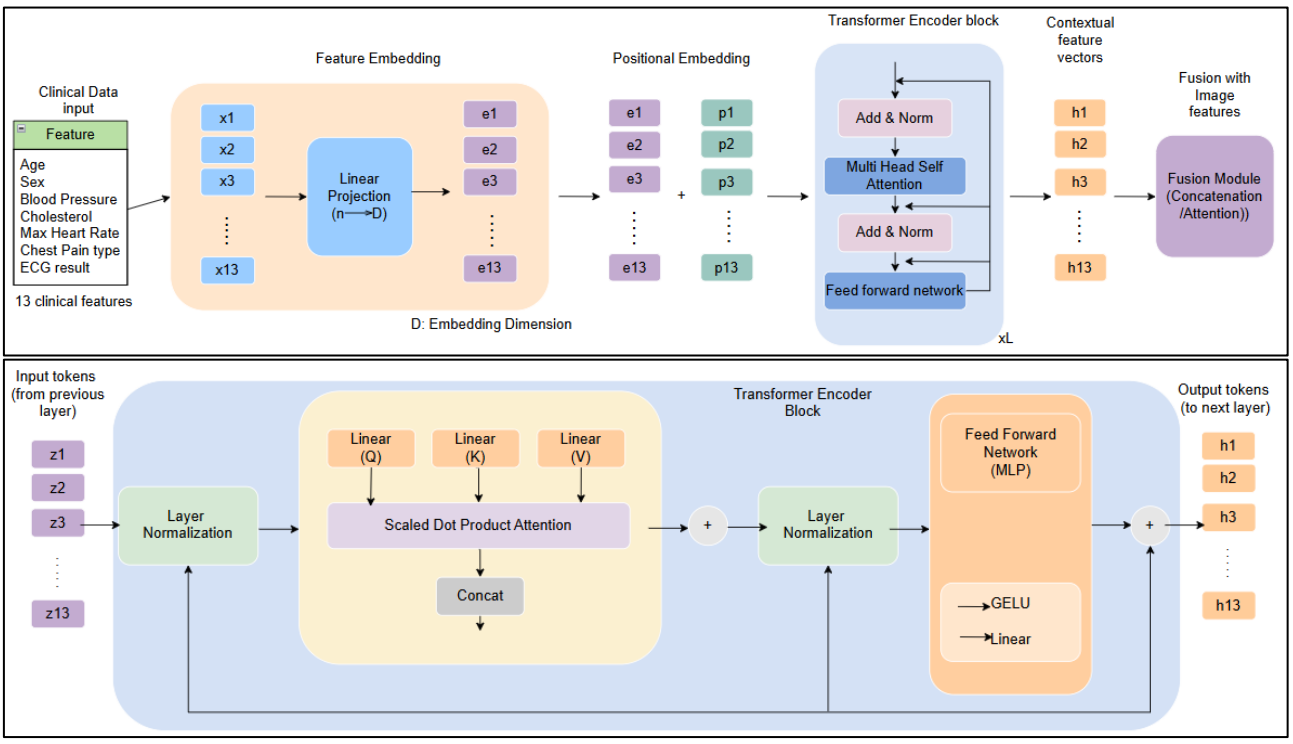


Figure 4: Transformer Encoder Block

3.3.3 Multimodal Fusion

The fusion layer fuses the features extracted from the CNN and the structured clinical data together. It allows the model to learn both visual features and numeric features at the same time.
Basic concatenation-based fusion is:

$$F_{\text{fusion}} = [F_{\text{CNN}} \oplus X_{\text{clinical}}]$$

where:

i. F_{CNN} represents features extracted from the CNN
ii. X_{clinical} represents features extracted from the clinical input
iii. $\oplus$ means concatenation operation

In order to improve the learning process of representation, an attention-based fusion is employed:

$$F_{\text{fusion}} = \alpha F_{\text{CNN}} + (1 - \alpha) X_{\text{clinical}}$$

where α is a learned parameter of importance for each modality.
This method enables the framework to focus on the most significant features extracted from both imaging and clinical data, thereby improving the quality and reliability of the final prediction.

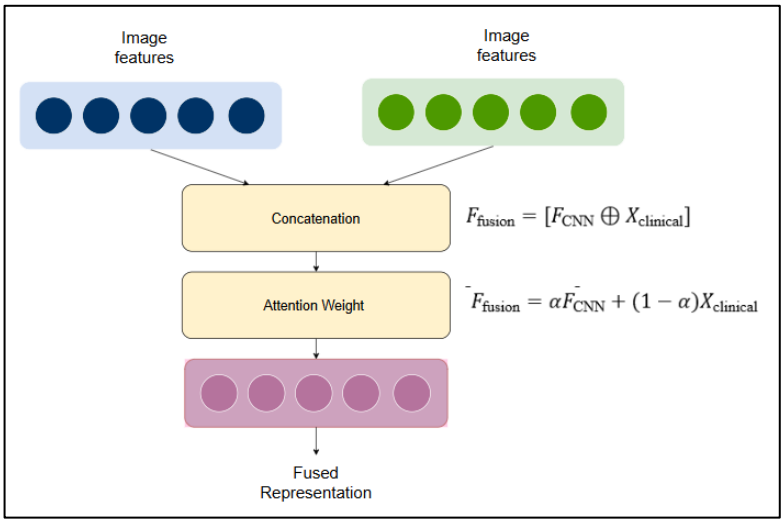


Figure 5: Fusion Block

3.3.4 Classification Layer

The concatenated feature representation is then used in the classifier model, comprising dense layers only.
Some important aspects in this regard are as follows:

i. Dense layer for nonlinear mapping
ii. Dropout layer for regularization (if required)
iii. Softmax activation for classification

This is because the Softmax activation returns probability scores for all classes, i.e., heart disease presence or absence.

3.3.5 Explainability Layer

In order to increase transparency and promote clinical trustworthiness, the XAI methods will be included in the suggested architecture to facilitate an interpretable explanation for the model's predictive process. Through this, the decision-making process can be made clear for clinicians since these explainability methods enable one to discover key clinical factors used by the algorithm to reach its diagnosis.

A. Grad-CAM (Gradient-weighted Class Activation Mapping)

Grad-CAM generates visual heat maps that highlight the areas in medical images that are contributing most significantly to the decision-making of the network. With regards to the analysis of chest x-rays, these heat maps help doctors identify important areas related to heart diseases.

B. SHAP (Shapley Additive Explanations)

SHAP technique is employed in determining the significance of the impact of individual clinical factors on the overall prediction outcome. With SHAP, one can be able to analyze both globally and locally to interpret the model through the significance assigned to each variable. Thus, one can be able to determine the key variables contributing to heart disease diagnosis.

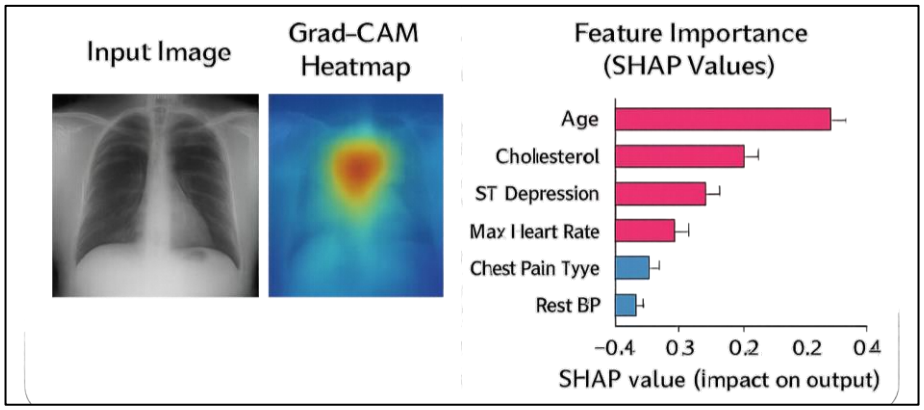


Figure 6: Interpretability using Grad-CAM and SHAP

CHAPTER 4: IMPLEMENTATION DETAILS

The following subsection will outline the flow of the algorithmic process in the Hybrid CNN-Transformer model that is used for diagnosing cardiovascular diseases. It combines imaging and clinical data to provide accurate results with meaningful explanations.

4.1 ALGORITHM: HYBRID CNN–TRANSFORMER

4.1.1 Input:

i. Cardiac Images I
ii. Clinical Features C

4.1.2 Output:

i. Heart Disease Prediction Y
ii. Explanations E

4.1.3 Procedure

A. Preprocessing: Input datasets, i.e., cardiac images and clinical features, are processed beforehand for consistency. Images are subjected to processes like resizing, normalization, and data augmentation, whereas clinical data is subjected to cleaning, normalization, and encoding.
B. Feature Extraction: The processed cardiac images are fed to the CNN-based module where they undergo the extraction of hierarchical spatial features. Such features can represent certain visual patterns that could be linked to certain types of cardiovascular diseases.
C. Contextual Analysis: The resulting extracted features undergo contextual learning in the Transformer-based module. Contextual learning involves the understanding of feature relations through self-attention mechanisms.
D. Multimodal Data Fusion: The processed data outputs from the CNN-Transformer network are fused together with other structured clinical features by the fusion mechanism.
E. Prediction Layer: The combined feature representation is sent to fully connected layers and then to a Softmax layer to calculate the probability of presence of heart disease.
F. Explanation Generation: Explainable Artificial Intelligence techniques are used to provide explanations of the prediction of the model:
    i. Grad-CAM creates heat maps for highlighting the important sections of the input images.
    ii. SHAP assigns an importance value to each clinical feature based on their impact on the prediction.
G. Generation of Outputs: The final output consists of both the predicted output class Y and explanations E.

4.2 CODE IMPLEMENTATION

Implementation of the proposed Hybrid CNN-Transformer approach is discussed in this section. The proposed implementation has been performed using Python and TensorFlow/Keras and includes the combination of clinical data and imaging data for multimodal classification.

4.2.1 Importing Libraries

The proposed approach requires the use of certain libraries in order to perform numerical computations, data preprocessing, deep learning modeling, visualization, and benchmarking. This set of libraries provides the necessary utilities needed for processing the dataset, training the neural network, and evaluating the predictions. The following libraries will be used:

```
import numpy as np
import pandas as pd
import tensorflow as tf
from tensorflow.keras.models import Model
from tensorflow.keras.layers import Dense, Flatten, Conv2D
from tensorflow.keras.layers import MaxPooling2D, Input
from tensorflow.keras.layers import Concatenate
from tensorflow.keras.optimizers import Adam
from sklearn.model_selection import train_test_split
from sklearn.preprocessing import StandardScaler
```

4.2.2 Loading Clinical Dataset

The clinical dataset is loaded and preprocessed by separating features and target labels. Feature scaling is applied to normalize the input data, followed by splitting into training and testing sets.

```
clinical_data = pd.read_csv('heart.csv')
X = clinical_data.drop('target', axis=1)
y = clinical_data['target']
scaler = StandardScaler()
X_scaled = scaler.fit_transform(X)
X_train, X_test, y_train, y_test = train_test_split( X_scaled, y, test_size=0.2, random_state=42)
```

4.2.3 CNN Model for Image Feature Extraction

The CNN module is responsible for extracting spatial features from input chest X-ray images.

```
def build_cnn_branch(input_shape=(224,224,3), trainable=False):
    img_data = Input(shape=input_shape, name="img_data")
    base_model = cnn(weights="imagenet", include_top=False, input_tensor= img_data)
    base_model.trainable = trainable
    x = base_model.output  # (None, 7, 7, 1280)
    x = GlobalAveragePooling2D(name="gap_pool")(x)   # (None,1280)
    x = Dense(512, activation="relu", name="img_dense")(x)
    x = BatchNormalization()(x)
    x = Dropout(0.3)(x)
    return img_data, x
```

4.2.4 Clinical Data Network

A fully connected neural network is used to process structured clinical features.

```
def build_transf_branch(num_features=13, seq_len=1, d_model=128, num_heads=4, ff_dim=256,
```

```
    num_layers=2, dropout=0.1):
  clinical_data = Input( shape=(num_features,), name=" clinical_data" )
  # Feature Embedding
  x = Dense(d_model, name="clin_embedding")( clinical_data)
  x = Reshape( (seq_len, d_model))(x)
  # Positional Embedding
  x = PositionalEmbedding(sequence_length=seq_len, d_model=d_model)(x)
  # Transformer Layers
  for _ in range(num_layers):
     x = transformer_encoder(x, head_size=d_model//num_heads, num_heads=num_heads,
        ff_dim=ff_dim, dropout=dropout)
  x = Flatten(name="vit_flatten")(x)
  return clinical_data, x
```

4.2.5 Multimodal Fusion Layer

Features extracted from both CNN and clinical networks are combined using concatenation and further processed.

```
fusion = Concatenate(name="fusion")([img_data, clinical_data])
  x = Dense(256, activation="relu")(fusion)
  x = BatchNormalization()(x)
  x = Dropout(0.3)(x)
  x = Dense(128, activation="relu")(x)
  x = Dropout(0.2)(x)
  output = Dense(1, activation="sigmoid",name="output")(x)
```

4.2.6 Model Compilation

The model is compiled using the Adam optimizer and binary cross-entropy loss function for classification.

```
model = Model(inputs=[img_data, clinical_data], outputs=output)
model.compile(optimizer=Adam(learning_rate=0.001),loss='binary_crossentropy',metrics=['accura
cy'])
```

4.2.7 Model Training

The model is trained using both imaging and clinical inputs.

```
history = model.fit([img_data, clinical_data],y_train,epochs=50,batch_size=32,validation_split=0.2)
```

4.2.8 SHAP Explainability

SHAP is used to interpret the impact of clinical features on model predictions.

```
import shap
explainer = shap.Explainer(model, clinical_train)
shap_values = explainer(clinical_test)
shap.summary_plot(shap_values, clinical_test)
```

4.2.9 Grad-CAM Visualization

Grad-CAM is applied to visualize important regions in input images that influence predictions.

```
import cv2
import matplotlib.pyplot as plt
# Generate Grad-CAM heatmap
heatmap = gradcam(image)
plt.imshow(heatmap)
plt.show()
```

4.3 SYSTEM REQUIREMENTS

4.3.1 Software Requirements

The implementation of the proposed system is carried out using widely adopted programming tools, deep learning frameworks, and data processing libraries.

Table 8: Development Tools and Libraries

Component	Technology
Programming Language	Python
IDE	Jupyter Notebook
Framework	TensorFlow
Neural Network API	Keras
Transformer implementation	PyTorch
Image Processing	OpenCV
Visualization	Matplotlib
Explainability	SHAP
Heatmap Generation	Grad-CAM
Data Processing	Pandas, NumPy

Table 9: Software Versions

Software	Version
Python	3.10
TensorFlow	2.x
PyTorch	Latest
Jupyter Notebook	Latest

4.3.2 Hardware Requirements

Efficient implementation of the proposed deep learning model requires adequate computational resources.

Table 10: Hardware Requirements

Component	Specification
Processor	Intel i7/Ryzen 7
RAM	16 GB
GPU	NVIDIA CUDA GPU
Storage	512 GB SSD

4.3.3 Hyperparameters

The performance of the model is influenced by several hyperparameters, which are carefully selected to ensure optimal training and convergence.

Table 11: Hyperparameter Configuration

Parameter	Value
Learning Rate	0.001
Optimizer	Adam
Batch Size	32
Epochs	50
Activation	ReLU
Loss Function	Binary Cross-entropy

CHAPTER 5: RESULTS AND ANALYSIS

5.1 PERFORMANCE AND EVALUATION METRICS

These formulas are used for evaluating and analyzing the performance of the proposed model. These metrics give numerical values that indicate how accurately the model can predict and classify the data and, ultimately, how reliable the model is in identifying heart disease.

A. **Accuracy**

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN}$$

B. **Precision**

$$\text{Precision} = \frac{TP}{TP + FP}$$

C. **Recall**

$$\text{Recall} = \frac{TP}{TP + FN}$$

D. **F1-Score**

$$\text{F1-Score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$$

where:

i. TP: True Positives
ii. TN: True Negatives
iii. FP: False Positives
iv. FN: False Negatives

The efficiency of the suggested framework is determined using popularly accepted metrics for classification like Accuracy, Precision, Recall, and F1-Score. Through these metrics of evaluation, there is an extensive evaluation done of how the model performs through its capacity to correctly classify cases of heart disease, avoid misclassification, and be reliable as a whole.

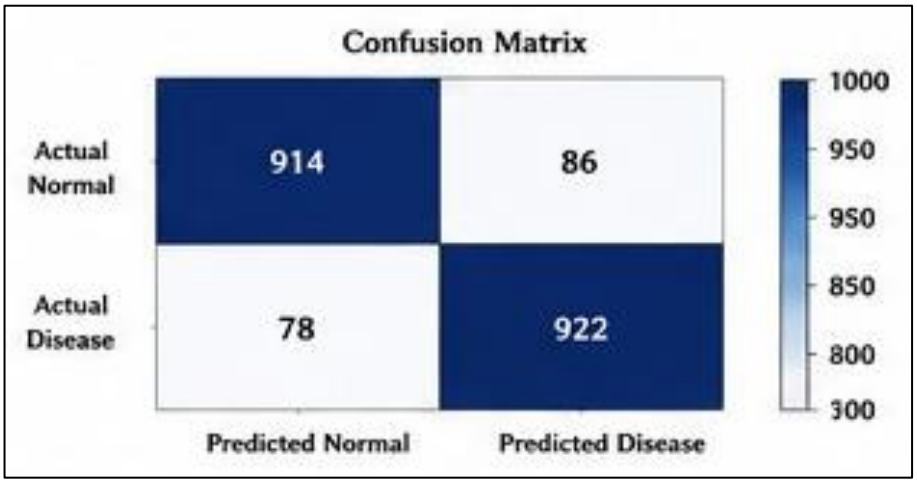


Figure 7: Proposed Model Performance and Evaluation

Confusion matrix provides an overview of how the classification is working by taking into account the relationship between the real values and the output values provided by the predictive model. The number of correct and incorrectly classified instances are counted in terms of true positives, true negatives, false positives, and false negatives. Confusion matrix enables one to understand what type of errors is made in predicting outcomes using this model.

5.2 MODEL COMPARISON

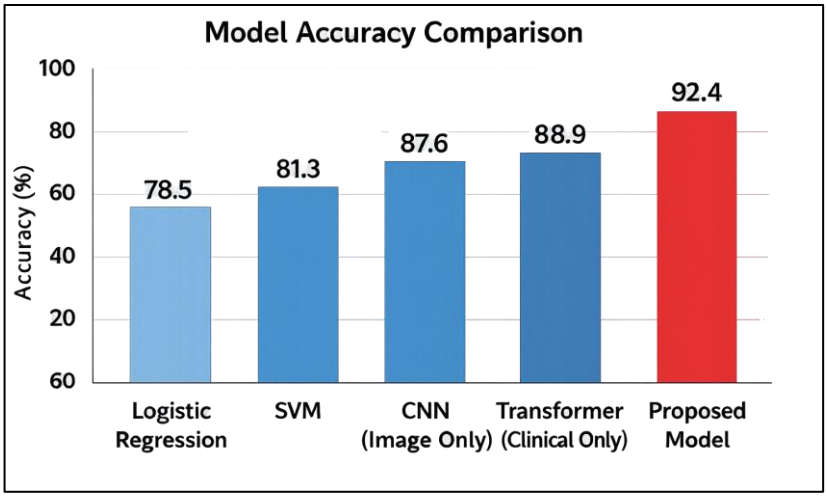


Figure 8: Proposed Model Accuracy Comparison

Table 12: Proposed Model Performance and Evaluation Comparison

Model	Accuracy	Precision	Recall	F1-Score
Logistic Regression	78.5	76.2	74.8	75.5
SVM	81.3	79.5	78.1	78.8
CNN	87.6	85.9	86.3	86.1
Transformer	88.9	87.5	86.8	87.1
Proposed CNN–Transformer	92.4	91.2	90.8	91.0

However, the results obtained from the experiment prove that the new framework is better than traditional ML methods as well as individual DL approaches based on all criteria. The use of the effective combination of spatial feature extraction and contextual learning with attention has contributed to improving the performance of classification and increasing precision and recall in heart disease prediction.

5.3 ROC Curve of Proposed Model

The ROC curve shows the relation between the true positive rate (sensitivity) and the false positive rate for various classification boundaries. This is an efficient way of measuring the ability of the model to discriminate between cases of diseases and those that are not. The suggested CNN-Transformer hybrid model has a good Area Under the Curve (AUC) score, proving its discriminating capacity in the prediction of heart diseases.

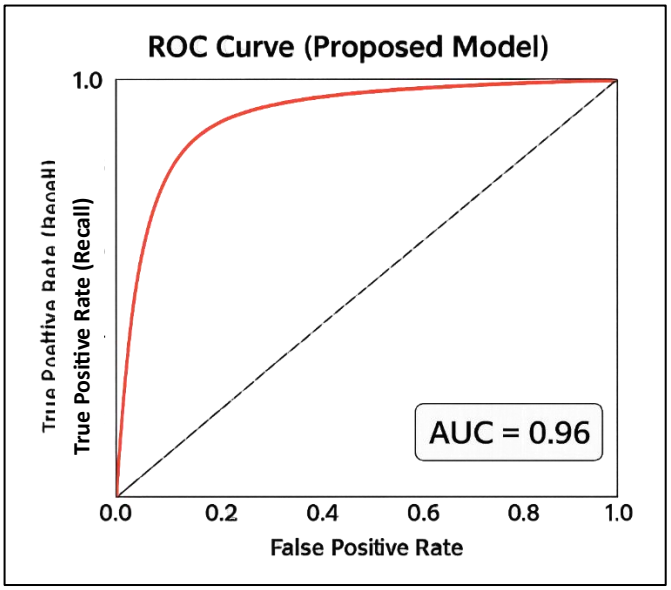


Figure 9: ROC Curve of the Proposed Model

5.4 IMPACT OF MULTIMODAL FUSION

To measure the efficiency of the multimodal learning method, different types of experiments were conducted based on various combinations of the inputs used for processing. Three different experiments were performed using clinical data alone, imaging data alone, and combined inputs of both types. The comparison revealed that combining imaging-based inputs with clinical structured data leads to enhanced performance in predicting heart disease.

Table 13: Impact of Multimodal Fusion

Configuration	Accuracy	Precision	Recall	F1-Score
Imaging Only (CNN)	87.6	85.9	86.3	86.1
Clinical Data Only	83.2	81.5	80.9	81.2
CNN + Transformer (No Fusion)	89.1	87.8	87.0	87.4
Proposed CNN–Transformer	92.4	91.2	90.8	91.0

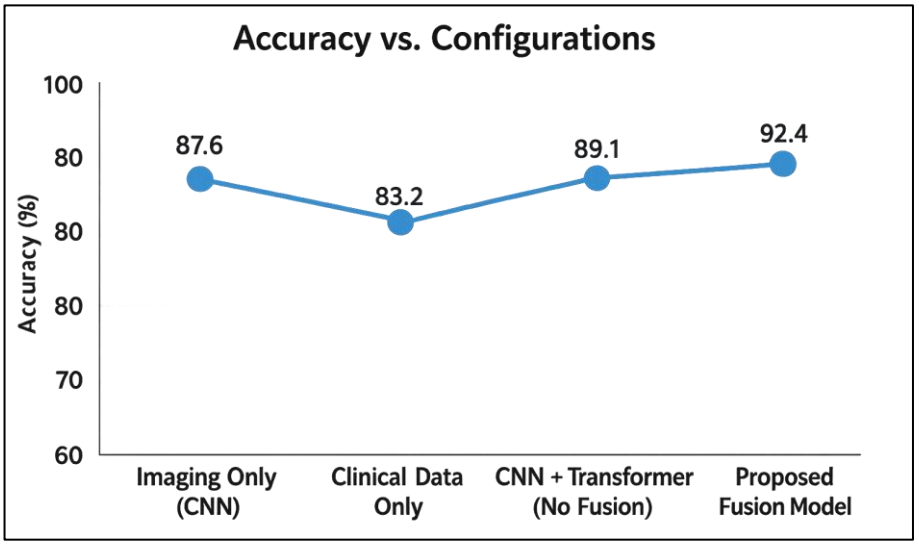


Figure 10: Impact of Mutimodal Fusion

From the findings, it is clear that integration of imaging and clinical information increases the effectiveness of the models. The multimodal fusion technique successfully integrates complementary information from different sources to improve performance.

5.5 EXPLAINABILITY ANALYSIS

5.5.1 SHAP Analysis

The SHAP analysis technique is used for determining the contribution of the clinical parameters in making the predictions. Based on the SHAP values, it is clear that the cholesterol level, age, and the kind of chest pain significantly affect the outcome of the predictions.

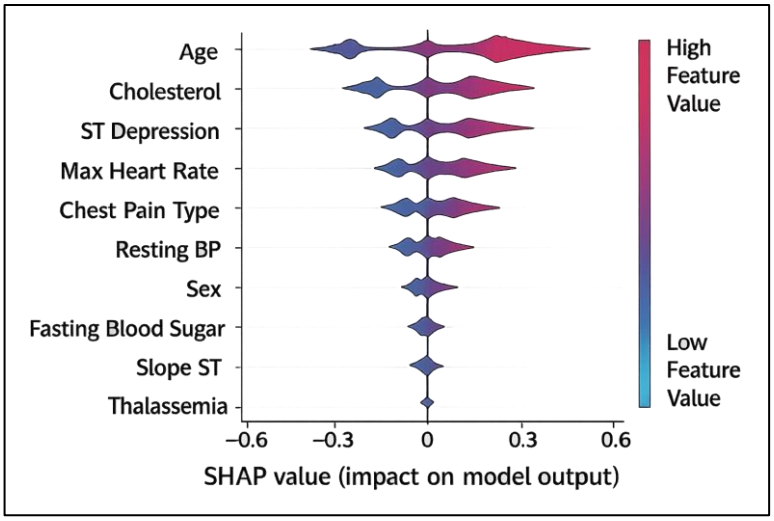


Figure 11: SHAP Analysis

5.5.2 Grad-CAM Visualization

Grad-CAM visualizations show key points in the X-ray chest image where the network is making decisions. The fact that the model uses relevant features to decide shows that the model can be trusted.

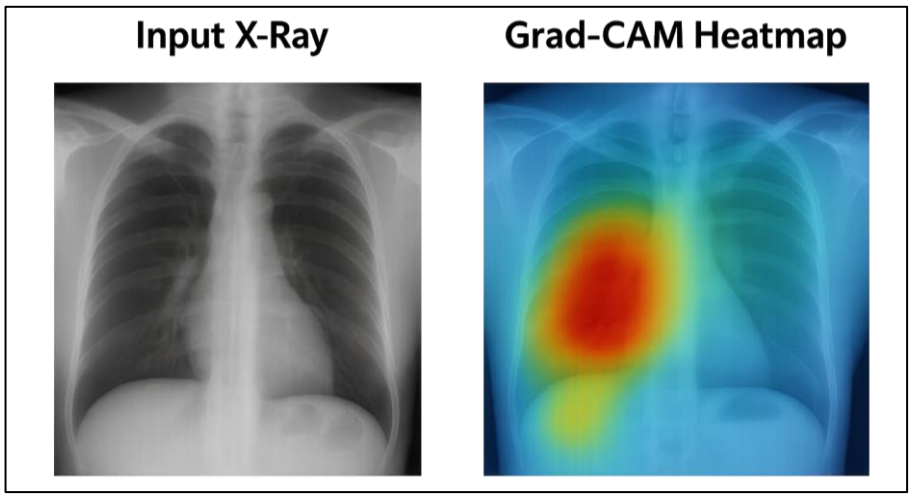


Figure 12: Grad-CAM Visualization

5.6 INTERPRETATION OF RESULTS

From the experiments conducted, it is evident that the multimodal hybrid model performs better than the conventional machine learning methods and also better than deep learning models. These results are brought about by various factors such as the following:

i. Spatial feature extraction from CNN
ii. Context learning through the use of Transformer-based attention mechanisms
iii. Multimodal data integration by fusing imaging and clinical features
iv. Explainability techniques for improved transparency

These components collectively enhance the robustness, generalization capability, and reliability of the model.

5.7 ADVANTAGES OF PROPOSED MODEL

Some benefits of the proposed framework include:

i. Greater prediction accuracy than baseline models
ii. Better contextual feature representation
iii. Utilization of multiple data modalities
iv. Explainability via explainable AI methods
v. More clinical transparency and trustworthiness
vi. Good performance regardless of evaluation criteria used
vii. Scalable system design for practical medical application

CHAPTER 6: APPLICATIONS

There are various uses of the multi-modal technique in modern medicine, particularly in diagnosing and curing cardiovascular disorders. This multi-modal strategy is more comprehensive than traditional ways because it combines imaging and clinical data.

6.1 EARLY CARDIOVASCULAR RISK ASSESSMENT

An important use for this algorithm involves identifying those individuals who may be at risk of contracting heart disease at an early stage. In doing so, the algorithm can help pick up some patterns which would otherwise be hard to identify using the conventional methods, by analyzing the results of chest x-ray pictures coupled with other important parameters like age, cholesterol, and blood pressure levels.

6.2 AI-ASSISTED DIAGNOSIS

This can be said to be a helpful tool for doctors that provides one more level of interpretation. This does not override the process of decision making by doctors, but rather complements this process by offering valuable information about important factors analyzed from imaging and clinical data sources. Some interpretability tools such as Grad-CAM for images and SHAP for clinical data attributes help understand the reasons behind the prediction.

6.3 SMART HEALTHCARE SYSTEMS

As technology is being increasingly adopted in the field of healthcare, the new model can easily be incorporated within smart healthcare delivery systems that continuously gather and analyze information about patients. The new model will help in analyzing such information quickly, resulting in better decision-making capabilities, health alert mechanisms, and ultimately better patient outcomes.

6.4 CLINICAL DECISION SUPPORT

The system can even be integrated into clinical decision-support applications for use by the healthcare professionals. In addition to predicting the likelihood of a disease, it provides visualization, thus allowing healthcare professionals to make better decisions concerning diagnosis and treatments. This is very beneficial in a fast-paced healthcare environment.

6.5 TELEMEDICINE APPLICATIONS

In places where there is no direct access to specialized medical services, the approach may be applied using telemedicine services. Patients will be able to upload their radiographic images online along with other information regarding their illness, and thus, the approach will be able to give an initial diagnosis.
This will ensure that physical consultation is not required right away..

6.6 REMOTE PATIENT MONITORING

The system can also help monitor patients continuously, especially those who already suffer from heart problems. With information gathered from wearable devices and updates from the clinic, this

system can monitor the development of any adverse effects on the patient's health. This way, they can be informed early if there are any signs of health decline.

CHAPTER 7: LIMITATIONS

Despite its successful results and benefits in predicting heart diseases, there are still some disadvantages that need to be discussed regarding the proposed multimodal framework. Firstly, its performance highly depends on the quality and availability of multimodal healthcare data. Poor medical images, incomplete clinical records, and class imbalance may have an impact on the predictions' precision. Moreover, using CNNs and Transformers as components of the framework raises the need for powerful hardware, namely, GPU accelerators for efficient training and prediction. It is also vital to mention the problem with obtaining large-scaled synchronized multimodal datasets for experiments and further practical use of the framework. Differences in the imaging quality, demographic diversity of patients, and even the clinical guidelines of various hospitals might pose some challenges. Finally, it should be mentioned that even though SHAP and Grad-CAM increase the framework's explainability, its explanations still need to be checked by a clinical specialist.

7.1 HIGH COMPUTATIONAL COMPLEXITY

There are a number of components associated with the architecture of the model, which include the use of convolutional neural networks, transformers, and attention. Though this helps improve the efficiency of the network, it leads to higher computation cost. Training such a network takes considerable time, and at times the prediction is not immediate, particularly when working with a lot of data and medical imaging.

7.2 DEPENDENCE ON DATASET QUALITY

The performance of the algorithm greatly relies on the quality of data utilized in training the system. In the case of noisy, incomplete, or biased data, the results of the algorithm may be negatively impacted. The problem is especially pertinent when it comes to the medical field since erroneous or unbalanced data will result in invalid output.

7.3 REQUIREMENT OF GPU RESOURCES

However, because of its complicated architecture, the model will usually need GPU computing to effectively train and test. Without the capability to access top-of-the-line hardware, training could be done slowly or training might even be impossible. This constraint will hinder the use of the method among smaller clinics or laboratories due to insufficient computational capacity.

7.4 LIMITED AVAILABILITY OF MULTIMODAL DATASETS

Yet another important problem lies in the paucity of multimodal datasets combining radiological images along with other forms of clinical data. The majority of publicly available datasets usually concentrate on one kind of information only, thus making the training of multimodal systems problematic. Acquiring datasets of this type takes a considerable amount of time and involves a number of ethical considerations.

CHAPTER 8: FUTURE WORK

However, despite showing some positive outcomes for predicting heart diseases in the proposed multimodal approach, there are various areas in which the current research could be improved in order to achieve even better performance.

8.1 INTEGRATION OF ECG SIGNALS AND WEARABLE SENSOR DATA

Possible future research could involve the use of other physiological signals like those from an electrocardiogram (ECG) and inputs from wearable devices. This will help to obtain more accurate information about the cardiac function of the patient in real time. This could enable the model to incorporate temporal changes in relation to disease progression.

8.2 FEDERATED LEARNING FOR PRIVACY-PRESERVING HEALTHCARE

Privacy of data is still an important aspect when it comes to the application of this technology in health care. In order to tackle this issue, the concept of federated learning can be implemented whereby the training of the algorithm is done without transferring raw data across different organizations.

8.3 LIGHTWEIGHT TRANSFORMER ARCHITECTURES

The current architectural design involves computationally-intensive models such as transformer neural networks. Future research may explore a path of designing lighter and faster transformer architectures that can deliver good performance using lesser computation capacity.

8.4 EDGE-AI OPTIMIZATION

Other important topics that can be studied are how to make the machine learning model deployable on edge devices such as smart phones or medical devices. By making the machine learning model more efficient in terms of size and faster for inferences, the machine learning model becomes more deployable close to where the data comes from.

8.5 REAL-TIME DEPLOYMENT IN CLINICAL SETTINGS

Research for future implementation could be conducted whereby the designed system is put into practice in real-world hospital environments to aid the doctors during the diagnosis process. This will involve overcoming challenges such as integration with other applications, design of user interface, among others.

8.6 INTEGRATION WITH ELECTRONIC HEALTH RECORDS (EHR)

The inclusion of information derived from EHR can also serve to boost the accuracy of this predictive model through providing even more insights about the patient's medical background, which includes previous medical diagnoses, medications used, and other health records.

8.7 ADVANCED CROSS-MODAL ATTENTION MECHANISMS

While this study uses attention-based fusion techniques, there is scope to use more advanced cross-modal attention models that can capture the correlation between the imaging data and clinical data. More effective ways to perform fusion may enhance feature representation and understanding.

8.8 LARGE-SCALE CLINICAL VALIDATION

Validation, however, needs to be carried out using bigger datasets so as to prove that the model can be generalized. Future studies will need to be concentrated on doing research through multi-centered and real-world clinical trials.

CHAPTER 9: CONCLUSION

9.1 SUMMARY OF WORK

This study introduced a unique and comprehensible framework for detecting cardiovascular diseases using multimodal deep learning technology, utilizing medical imaging along with structured data obtained clinically. This new approach is based on the combination of Convolutional Neural Networks (CNN) as well as Transformer models to enable the effective representation of local as well as global contexts.

With respect to the CNN part, there is efficient utilization of local contexts in order to extract informative spatial patterns from cardiac imaging data. In contrast, the other part, which is the Transformer model, helps to achieve better performance with respect to capturing the global contexts of data.

9.2 KEY CONTRIBUTIONS

One of the significant contributions of this work is the use of XAI techniques such as SHAP and Grad-CAM in order to increase the transparency and interpretability of the models. These techniques help gain insights on the decision-making process of the models by:

i. Finding out the most influencing clinical features for predictions
ii. Recognizing the most important regions in medical imaging that contribute to classifications

It is vital for a clinical environment to have interpretable models since the reasoning behind predictions needs to be well understood.

9.3 PERFORMANCE OBSERVATIONS

The experiment results show that the developed multimodal framework achieves better performance compared to standard machine learning frameworks and deep learning alone. This is because the fusion of imaging and structured data allows the framework to take advantage of complementarity of information, which leads to:

i. Better classification accuracy
ii. Better precision and recall
iii. Better robustness

Furthermore, the success of multimodal fusion and attention mechanisms in analyzing the complicated relationship among the medical data is confirmed..

9.4 RESEARCH SIGNIFICANCE

The presented framework helps to enhance the use of artificial intelligence in healthcare by overcoming certain challenges, including poor data integration and lack of transparency. High performance along with explainability enhances the applicability of the suggested system within the real world, where both of the aforementioned features are important.

9.5 FINAL REMARKS

In conclusion, the current research can be considered a landmark step that can ensure scalable and reliable AI solutions to diagnostics of heart disease in the future. The use of multimodal data together with explainable AI can contribute to more accurate diagnostics and increase clinicians' confidence in AI algorithms. The current research can serve as a stepping stone for further investigation of intelligent health care systems.

REFERENCE

[1] Moolya, S., Gopalakrishna Kini, N. (2025). Machine Learning Approach for Early Detection of Heart Disease. In: Bhattacharyya, S., Banerjee, J.S., Köppen, M., Nayak, S., Platos, J. (eds) Human-Centric Smart Computing. ICHCSC 2024. Smart Innovation, Systems and Technologies, vol 440. Springer, Singapore. https://doi.org/10.1007/978-981-96-3420-0_7

[2] Talukder, M.A., Talaat, A.S., Kazi, M. et al. XAI-HD: an explainable artificial intelligence framework for heart disease detection. Artif Intell Rev 58, 385 (2025). https://doi.org/10.1007/s10462-025-11385-6

[3] Talwar, A., Fatima, N., Rizvi, M.S.B.A., Rizvi, S.A.M. (2025). Detection and Diagnosis of Framingham Heart Disease Using Machine Learning. In: Rattan, S., Tiwari, S., Garg, S. (eds) Recent Trends in Materials and Devices. ICRTMD 2023. Springer Proceedings in Physics, vol 320. Springer, Singapore. https://doi.org/10.1007/978-981-96-9989-6_15

[4] Divyashree, H.B., Kokila, S., Eduru, S., Peddamariveedu, M. (2025). Heart Disease Detection Using Fundus Imaging and Ml Algorithm. In: Hassanien, A.E., Anand, S., Jaiswal, A., Kumar, P. (eds) Innovative Computing and Communications. ICICC 2025. Lecture Notes in Networks and Systems, vol 1430. Springer, Singapore. https://doi.org/10.1007/978-981-96-6883-0_23

[5] MahaLakshmi, N.V., Rout, R.K. Hyper tuning hybrid MLP-CatBoost classifier using Mayfly optimization for detection of heart disease. Health Serv Outcomes Res Method 25, 305–327 (2025). https://doi.org/10.1007/s10742-025-00341-0

[6] Abrar, M.F., Saqib, M., Ali, S. et al. Automated heart disease detection using Swin Transformer and ECG signal processing: a high-accuracy approach. Sci Rep 15, 39338 (2025). https://doi.org/10.1038/s41598-025-23097-6

[7] Goyal, S.R. (2025). DLA-HDD: A Novel Deep Learning Architecture for Early Heart Disease Detection. In: Sharma, H., Chakravorty, A. (eds) Proceedings of International Conference on Information Technology and Intelligence. ICITI 2024. Lecture Notes in Networks and Systems, vol 1342. Springer, Singapore. https://doi.org/10.1007/978-981-96-5010-1_7

[8] Anand, R., Reddy, S.R.N. & Yadav, D.K. Unveiling Hearts: Deep Learning-Based Electrocardiogram Classification for Congenital Heart Disease Detection. CURR MED SCI 45, 1460–1470 (2025). https://doi.org/10.1007/s11596-025-00134-z

[9] Widjaja, Y.K., Suhendra, A., Harmanto, S., Sutedja, I. (2025). Analysis the Impact of Intelligent Heart Sound Detection Systems for Heart Valve Failure Disease. In: Ibrahim, F., Mohamed Shah, N., Ahmad, M.Y., Ana, I.D. (eds) 5th International Conference for Innovation in Biomedical Engineering and Life Sciences. ICIBEL 2024. IFMBE Proceedings, vol 129. Springer, Cham. https://doi.org/10.1007/978-3-031-97374-1_24

[10] Ghosh, S., Dasgupta, S., Bhattacharya, I. (2025). An Intelligent Integrated Prediction Based Approach for Heart Disease Detection - A Comprehensive Study. In: Singh, J.P., Singh, M.P., Singh, A.K., Mukhopadhyay, S., Mandal, J.K., Dutta, P. (eds) Computational Intelligence in Communications and Business Analytics. CICBA 2024. Communications in Computer and Information Science, vol 2367. Springer, Cham. https://doi.org/10.1007/978-3-031-81339-9_18

[11] Liu, X., Li, C. (2025). MFTrans: An Extensible Transformer-Based Medical Multimodal Data Fusion Method for Clinical Diagnosis. In: Mahmud, M., Doborjeh, M., Wong, K., Leung, A.C.S., Doborjeh, Z., Tanveer, M. (eds) Neural Information Processing. ICONIP 2024. Communications in Computer and Information Science, vol 2287. Springer, Singapore. https://doi.org/10.1007/978-981-96-6963-9_5

[12] Shi, R., Sun, J., Zhou, Z. et al. Deep multimodal fusion of patho-radiomic and clinical data for enhanced survival prediction for colorectal cancer patients. npj Digit. Med. 9, 86 (2026). https://doi.org/10.1038/s41746-025-02210-z

[13] Wang, Z., Zhang, H., Li, Y. et al. Development and validation of a prognostic prediction model for lumbar-disc herniation based on machine learning and fusion of clinical text data and radiomic features. Eur Spine J 35, 529–547 (2026). https://doi.org/10.1007/s00586-025-09102-6

[14] Zeyauddin, M., Abidin, S., Khan, I. et al. Clinical predictive fusion network for accurate disease prediction in patient cohorts. Sci Rep 16, 3626 (2026). https://doi.org/10.1038/s41598-025-33645-9

[15] Zou, C., Chen, R., Wang, B. et al. Development of a deep learning radiomics model combining lumbar CT, multi-sequence MRI, and clinical data to predict high-risk cage subsidence after lumbar fusion: a retrospective multicenter study. BioMed Eng OnLine 24, 27 (2025). https://doi.org/10.1186/s12938-025-01355-y

[16] Nikolaou, N., Salazar, D., RaviPrakash, H. et al. A machine learning approach for multimodal data fusion for survival prediction in cancer patients. npj Precis. Onc. 9, 128 (2025). https://doi.org/10.1038/s41698-025-00917-6

[17] Akbari, Y., Abdullakutty, F., Al Maadeed, S. et al. Integrating histopathology and genomic data: a comparative study of fusion methods for breast cancer survival prediction. Complex Intell. Syst. 12, 30 (2026). https://doi.org/10.1007/s40747-025-02133-y

[18] Chen, S., Wang, D., Che, C. et al. A heterogeneous data fusion mortality prediction model based on time-aware self-attention mechanism. Int. J. Mach. Learn. & Cyber. 16, 6381–6397 (2025). https://doi.org/10.1007/s13042-025-02627-8

[19] Tang, J., Wang, H., Wu, D. et al. Radiation Pneumonitis Prediction Using Dual-Modal Data Fusion Based on Med3D Transfer Network. J Digit Imaging. Inform. med. 38, 2085–2101 (2025). https://doi.org/10.1007/s10278-024-01339-9

[20] Taniguchi, H., Fukuda, K., Sato, T. et al. Effect of the intravenous acetaminophen clinical pathway on postoperative analgesia in spinal fusion surgery. Sci Rep 15, 36644 (2025). https://doi.org/10.1038/s41598-025-20499-4