A

Project Report

on

CARDIAC DISEASE PREDICTION USING MACHINE LEARNING

submitted as partial fulfillment for the award of`

BACHELOR OF TECHNOLOGY DEGREE

SESSION 2024-25

In

Computer Science and Engineering

By

Manav Rohilla (2100290100090)

Sukriti Rai (2100290100169)

Parkhi Gupta (2100290100112)

Under the Supervision of

Ms. Nishu Gupta

KIET Group of Institutions, Ghaziabad

Affiliated to

Dr. A.P.J. Abdul Kalam Technical University, Lucknow

(Formerly UPTU)

May,2025

DECLARATION

We hereby declare that this submission is our own work and that, to the best of our knowledge and belief, it contains no material previously published or written by another person nor material which to a substantial extent has been accepted for the award of any other degree or diploma of the university or other institute of higher learning, except where due acknowledgment has been made in the text.

Signature:
Name: Manav Rohilla
Roll No.: 2100290100090

Signature:
Name: Parkhi Gupta
Roll No.: 2100290100112

Signature:
Name: Sukriti Rai
Roll No.: 2100290100169

CERTIFICATE

This is to certify that Project Report entitled "**Cardiac Disease Prediction Using Machine Learning**" Project Group No.: **33** which is submitted by **Manav Rohilla, Sukriti Rai, and Parkhi Gupta** in partial fulfillment of the requirement for the award of degree B. Tech. in the Department of Computer Science and Engineering of Dr. A.P.J. Abdul Kalam Technical University, Lucknow is a record of the candidates' own work carried out by them under my supervision. The matter embodied in this report is original and has not been submitted for the award of any other degree.

Supervisor:
Ms. Nishu Gupta

Dean CSE
Dr. Vineet Sharma

DATE: May 2025

ACKNOWLEDGEMENT

We are extremely grateful for the opportunity to present this report on our B.Tech Final Year Project, titled Cardiac Disease Prediction Using Machine Learning. This project has been an enriching experience, allowing us to explore the intersection of healthcare and artificial intelligence. The primary objective of our research was to develop an efficient, automated, and accurate system for predicting cardiac diseases using advanced texture feature extraction techniques such as Gray-Level Co-occurrence Matrix (GLCM), Histogram of Oriented Gradients (HOG), and Local Binary Patterns (LBP), integrated with machine learning models like SVM, Random Forest, Decision Tree and XGBoost. Our study aimed to provide a non-invasive and scalable diagnostic tool, contributing to early detection and better patient care.

First and foremost, we extend our heartfelt gratitude to our Supervisor of the Department of Computer Science and Engineering, KIET, Ghaziabad, for her continuous support, insightful guidance, and encouragement throughout the project. Her sincerity, rigor, and perseverance have been a constant source of inspiration, and her invaluable contributions have played a crucial role in shaping our work. It was only through her conscious efforts that this project successfully took off and reached completion.

We also take this opportunity to express our sincere appreciation to Dr. Vineet Sharma, Dean of Computer Science and Engineering, KIET, Ghaziabad, for his unwavering support and assistance during the development of our project. His encouragement and valuable insights have greatly contributed to our research and its successful execution.

Furthermore, we extend our gratitude to all faculty members of the Department of Computer Science and Engineering for their kind assistance and cooperation throughout the project. Their expertise and valuable inputs have been instrumental in refining our approach and methodology. We would especially like to acknowledge the contribution of faculty members, industry professionals, and external mentors who provided guidance and shared their knowledge, helping us enhance the quality and impact of our work.

Finally, we would like to acknowledge the dedication and efforts of our entire team. Every team member's commitment, collaboration, and hard work have been essential in bringing this project to fruition. Without their perseverance and teamwork, this research would not have been possible.

We sincerely appreciate all those who have contributed to this project in any capacity, directly or indirectly, and have helped us in making it a success.

DATE: May 2025

Signature:
Name: Manav Rohilla
Roll No.: 2100290100090

Signature:
Name: Parkhi Gupta
Roll No.: 2100290100112

Signature:
Name: Sukriti Rai
Roll No.: 2100290100169

ABSTRACT

Congenital heart diseases (CHDs) continue to be a major global health concern, accounting for a significant number of deaths worldwide. Early diagnosis and timely intervention are crucial to improving patient outcomes. This study introduces an innovative approach to cardiac disease prediction by integrating advanced texture analysis techniques with state-of-the-art machine learning algorithms. The proposed methodology extracts intricate structural patterns from heart shape data using three prominent feature extraction techniques: Gray-Level Co-occurrence Matrix (GLCM), Histogram of Oriented Gradients (HOG), and Local Binary Patterns (LBP). These methods are known for their effectiveness in capturing complex textural characteristics from medical images, providing a comprehensive representation of cardiac structures.

The extracted features are subsequently used to train various machine learning models, including SVM, Random Forest, Decision Tree and XGBoost. To ensure a rigorous and unbiased evaluation, the dataset is partitioned into distinct training and testing sets, with performance metrics such as accuracy, precision, recall, and F1-score employed for assessment. Our findings reveal that while individual feature extraction techniques contribute valuable insights, their combination significantly enhances predictive accuracy. Among the evaluated classifiers, XGBoost and Random Forest demonstrate superior performance, showcasing their ability to model complex feature spaces effectively and improve classification reliability.

This study highlights the transformative potential of combining texture analysis with machine learning for early-stage cardiac disease detection. By integrating advanced feature extraction techniques with ensemble learning strategies, our method achieves superior diagnostic accuracy, offering a valuable tool for proactive cardiac healthcare. The results emphasize the importance of AI-driven frameworks in enabling timely medical decisions and improving patient outcomes. Future research may explore deep learning-based feature extraction and larger, diverse datasets to enhance model generalizability and clinical applicability, contributing to the advancement of precision medicine. This work paves the way for the development of non-invasive, highly accurate cardiac disease prediction systems that can significantly enhance preventive healthcare strategies.

TABLE OF CONTENTS

List of Figures

List of Abbreviations

CHD	Congenital Heart Diseases
GLCM	Gray Level Co-occurrence Matrix
LBP	Local Binary Patterns
HOG	Histogram of Oriented Gradients
SVM	Support Vector Machine
CT	Computed Tomography
ASD	Atrial Septal Defect
AVSD	Atrioventricular Septal Defect
VSD	Ventricular Septal Defect
CA	Coarctation
TOF	Tetralogy of Fallot
PDA	Patent Ductus Arteriosus
PuA	Pulmonary Atresia
TGA	Transposition of the Great Arteries
PAS	Pulmonary Artery Sling
AAH	Aortic Arch Hypoplasia
DORV	Double Outlet Right Ventricle
CAT	Common Arterial Trunk
DAA	Double Aortic Arch
APVC	Anomalous Pulmonary Venous Drainage
IAA	Interrupted Aortic Arch
DSVC	Double Superior Vena Cava

CHAPTER 1: INTRODUCTION

Heart disease has become one of the biggest health concerns worldwide, causing millions of deaths every year. With the rise in congenital heart diseases (CHDs), early detection and timely treatment are more important than ever. Traditional diagnostic methods like electrocardiograms (ECG), echocardiography, and angiography are highly effective, but they require specialized expertise, take time, and may not always catch the disease in its early stages. This creates a strong need for automated and accurate diagnostic systems that can assist doctors in making faster, more precise decisions about heart health.

With advancements in artificial intelligence (AI) and machine learning (ML), healthcare is undergoing a technological transformation. These computational methods allow us to analyze vast amounts of medical data, identify hidden patterns, and improve the accuracy of disease prediction. AI-powered systems can work alongside traditional medical techniques to provide smarter, faster, and more reliable diagnoses, leading to better patient outcomes.

In this study, we introduce a new approach to predicting heart disease by analyzing heart shape and texture patterns in medical images. Instead of relying solely on standard tests, we extract key structural details using three feature texture techniques:

- **Gray-Level Co-occurrence Matrix (GLCM):** Analyzes the texture of the heart by measuring relationships between pixel intensities, capturing details like contrast, smoothness, and uniformity.

- **Histogram of Oriented Gradients (HOG):** Focuses on detecting shapes and edges in heart images, making it useful for identifying important structural patterns.

- **Local Binary Patterns (LBP):** Highlights fine textural differences by encoding the relationships between pixels, helping to distinguish between normal and diseased heart tissues.

By combining these techniques, we extract crucial features from medical images, which are then fed into machine learning models for classification. We experiment with SVM, Decision Tree, Random Forest, and XGBoost, evaluating their performance in predicting heart disease. Our results show that XGBoost and Random Forest outperform the others, demonstrating their ability to handle complex data and improve prediction accuracy.

The goal of this research is to develop an AI-powered, non-invasive, and scalable tool that can help doctors detect heart disease early, leading to faster interventions and better patient care. By integrating advanced image processing techniques with powerful machine learning models, we aim to create a system that is highly accurate, reliable, and efficient. This research contributes to the growing field of AI-driven healthcare, moving us closer to a future where personalized medicine and proactive disease management become a reality.

In the sections that follow, we explore existing research, dataset details, methodology, experimental results, and conclusions, showcasing how machine learning and texture analysis can play a key role in the early detection and prevention of heart disease.

1. LITERATURE SURVEY

1.1 Rise of Machine Learning in Cardiac Disease Diagnosis

- **Global Health Impact:**
  Cardiac Diseases remain the leading cause of death worldwide, contributing to approximately 17.9 million deaths annually. Early and accurate diagnosis is vital for reducing mortality and improving patient outcomes.

- **Traditional Diagnostic Limitations:**
  Conventional methods often rely heavily on clinician expertise and rule-based assessments, introducing potential subjectivity and variability. This underscores the need for objective, data-driven diagnostic tools.

- **Emergence of Machine Learning (ML):**
  Recent advancements in ML have shown great promise in addressing these challenges. Machine learning algorithms can detect complex patterns in large datasets, enabling more precise and faster diagnosis of Cardiac Diseases.

1.2 Literature Review: Machine Learning Approaches in Cardiac Disease Prediction

Extensive research has explored ML techniques for cardiovascular disease prediction, primarily utilizing **structured tabular data** (patient demographics, clinical measurements) and employing various feature selection and classification methods:

- A hybrid intelligence framework combined feature selection techniques (Relief, mRMR, LASSO) with classifiers like Logistic Regression, SVM, ANN, and Random Forest, achieving 89% accuracy.

- An evaluation on a 303-sample dataset found SVM yielding the highest accuracy (84.0%), followed by ANN (83.5%) and Random Forest (80.0%).

- The Enhanced Heart Disease Prediction System (EHDPS) employed classifiers like KNN and Random Forest directly on clinical parameters, with KNN achieving 88.52% accuracy.

- Large-scale studies (e.g., 70,000 patient records) utilized clustering and binning to preprocess tabular data, achieving 87.28% accuracy with MLP models.

- Hybrid models, combining Naive Bayes with Genetic Algorithms or Neural Networks, demonstrated exceptionally high accuracy (>97%).

Key Observation:
Most existing studies predominantly focused on clinical datasets, applying ML directly on structured (non-image) data. Very few leveraged detailed image-based feature extraction techniques.

1.3 Advancements in ML for Cardiac Imaging

- With the growing availability of cardiac imaging data (e.g., MRI, CT scans), there is an increasing shift toward utilizing image-based features for disease prediction. However, the majority of earlier works either applied deep learning directly or used simple imaging features without dedicated texture analysis.
- **Gap Identified:**
  There is a limited exploration of combining classical texture feature extraction methods with ML classifiers for cardiac disease prediction from images.

1.4 Our Proposed Approach: Texture Feature Extraction with Machine Learning

Distinctive Methodology:
Unlike earlier approaches that worked primarily on tabular clinical data, **our method focuses on extracting meaningful texture features from cardiac images** using advanced techniques:

- **Gray-Level Co-occurrence Matrix (GLCM):**
  Captures second-order statistical texture information, such as contrast, homogeneity, and correlation between pixels.

- **Histogram of Oriented Gradients (HOG):**
  Extracts edge orientation distributions, helping in identifying anatomical structures and irregularities.

- **Local Binary Patterns (LBP):**
  Captures fine-grained local texture patterns, enabling differentiation between normal and diseased tissues.

These extracted features are then fed into machine learning classifiers such as Decision Trees, Random Forests, SVMs, and ensemble models for classification.

Advantages of Our Approach:

- Focus on texture-level details that may not be evident in raw clinical data.
- Combining multiple feature extraction techniques to capture diverse aspects of image information.
- Improved model generalizability by enriching the feature space with high-quality descriptors.
- Potential for better interpretability, as extracted features can be linked back to specific textural changes in cardiac tissues.

1.5 Strength of Ensemble Learning in Medical Diagnostics

Ensemble learning has proven particularly effective in the medical domain, offering robustness against noisy data and model overfitting:

- **Bagging (e.g., Random Forests):**
  Aggregates outputs of multiple models trained on different data subsets, reducing variance.

- **Boosting (e.g., AdaBoost, XGBoost):**
  Builds strong classifiers by iteratively focusing on harder-to-predict samples.

In our approach, ensemble learning methods further enhance the predictive power when applied to rich, texture-based feature sets.

1.6 Proposed Study: An Integrative Approach

Building upon this growing body of research, our study proposes a comprehensive evaluation of feature extraction and ensemble classification methods for Cardiac Disease prediction.

- **Objective 1:**
  - Evaluate and compare multiple texture feature extraction techniques individually and in combination.
  - Identify the most discriminative features for cardiac image classification.
- **Objective 2:**
  - Implement and assess a variety of ensemble learning strategies.
  - Combine outputs from multiple classifiers to improve robustness and accuracy.

- **Goal:**
  - Develop a hybrid framework that integrates feature selection and ensemble learning to deliver more reliable and interpretable predictions.

1.7 Summary of Differences from Existing Literature

Aspect	Earlier Approaches	Our Approach
Input Data	Clinical data (age, BP, cholesterol, ECG, etc.)	Medical images (MRI scans)
Feature Extraction	Raw clinical features used directly or basic statistical processing	Advanced feature extraction using GLCM (texture), HOG (shape), LBP (local patterns)
Algorithms Used	Traditional ML classifiers like Random Forest, SVM, KNN, XGBoost	ML classifiers trained on extracted imaging features
Domain Focus	Numerical and categorical data analysis	Medical Imaging and Pattern Recognition
Preprocessing	Data cleaning, normalization, encoding	Image preprocessing (resizing, grayscale conversion) and feature engineering
Complexity	Moderate complexity; focused on structured data	Higher complexity due to image analysis and feature extraction
Accuracy Potential	Dependent on clinical attributes and statistical features	Higher accuracy potential by capturing detailed patterns from images
Innovation	Conventional methods widely studied	Novel application of texture and pattern-based imaging techniques

1.8 Future Directions in ML for Cardiac Disease Prediction

To further improve cardiovascular care through ML:

- **Deep Learning Extensions:**
  Integrate CNNs to automate feature extraction and classification directly from images.

- **Multi-modal Analysis:**
  Combine imaging features with clinical parameters for holistic prediction.

- **Clinical Deployment:**
  Embed the developed models into healthcare systems for real-time decision support.

- **Personalized Medicine:**
  Use predictive models to tailor treatment strategies based on individual risk profiles.

1.9 Conclusion

Machine learning has opened new frontiers in cardiovascular disease diagnosis. Our work differentiates itself from traditional approaches by emphasizing image-based texture feature extraction (GLCM, HOG, LBP) combined with robust machine learning classifiers. This integrative framework holds promise for enhancing diagnostic accuracy, offering clinicians a valuable decision-support tool in managing cardiac diseases. As ML-driven methodologies continue to evolve, the future of cardiology will increasingly move toward precision diagnostics and personalized care.

CHAPTER 2: REQUIREMENT ANALYSIS AND SYSTEM SPECIFICATION

2.1 Feasibility Study

A feasibility study serves as the foundation for evaluating the practicality of implementing the proposed **Cardiac Disease Prediction System** using Machine Learning (ML).

It systematically analyzes technical, operational, and economic factors that determine the likelihood of successful development, deployment, and real-world adoption of the system.

2.1.1 Technical Feasibility

The **technical feasibility** examines the technological infrastructure, availability of necessary resources, and the expertise required to ensure the efficient development and deployment of the system.

The proposed system applies machine learning algorithms to predict the likelihood of cardiac diseases based on patient clinical data.
This section evaluates key technical components:

Technical Requirements:

1. **Hardware Requirements:**
    - The system requires a modern computing environment equipped with at least 8 GB of RAM and a quad-core processor.
    - While deep learning models may benefit from GPU acceleration, it remains optional based on project complexity.

2. **Software Requirements:**
    - The programming will primarily be done using Python, utilizing libraries such as scikit-learn, NumPy, pandas, and Matplotlib for machine learning and data analysis tasks.
    - Development environments like Jupyter Notebook, PyCharm, or Visual Studio Code are recommended for an efficient coding workflow.
    - In cases where deep learning approaches are applied, frameworks like TensorFlow or PyTorch will be employed.

- For deployment purposes, tools such as Flask or Streamlit will be utilized to create user-friendly web-based applications.

3. **Data Requirements:**
    - Publicly available datasets, notably the ImageCHD Dataset, will be used.
    - Preprocessing tools and techniques will be needed for tasks like resizing, normalizing, and augmenting image data.

Feasibility Analysis:

- **Hardware & Software Feasibility:**
    - The project's hardware requirements are modest and can be fulfilled with commonly available systems.
    - All necessary software tools are open-source and widely documented, eliminating licensing costs and easing accessibility.
- **Data Feasibility:**
    - Well-known and reliable public datasets are readily accessible for training and evaluating the models, ensuring the availability of high-quality data sources.
- **Technical Expertise:**
    - Development requires a solid understanding of machine learning concepts, data preprocessing, and basic clinical knowledge.
    - Given the abundance of tutorials, online resources, and open-source libraries, acquiring the required expertise is highly achievable.

Conclusion: The project is technically feasible due to the availability of necessary resources, technologies, and datasets. It can be developed and tested using widely accessible tools and frameworks.

2.1.2 Operational Feasibility

The **operational feasibility** evaluates the practicality of the system's implementation in real-world settings and the willingness of intended users to adopt it.

The Cardiac Disease Prediction System is envisioned to assist medical professionals, healthcare institutions, and academic researchers in efficiently predicting the risk of heart disease based on clinical parameters using machine learning models.

Operational Environment:

- **Clinical Settings:**
    - The system can be integrated into hospital diagnostic workflows to aid doctors and specialists in decision-making processes.
- **Research and Educational Use:**
    - Academic institutions can use the system to train students in the fields of medical data science, machine learning applications in healthcare, and biomedical informatics.

User Acceptance Factors:

- **Ease of Use:**
    - A simple and intuitive user interface will ensure that users with minimal technical expertise can interact with the system effectively.
- **Accuracy:**
    - Clinical applications require a high standard of reliability. The system aims to maintain a prediction accuracy of over 85% to foster trust and acceptance among medical professionals.
- **Interpretability:**
    - Besides providing predictions, the system will offer insights into feature importance, enabling users to understand which clinical factors most influence the prediction outcomes.

- **Training and Support:**
    - The system will be designed in a way that minimal user training is required, thus reducing the learning curve.

- **Integration:**
    - Future upgrades will focus on ensuring compatibility with hospital information systems and electronic health record (EHR) platforms, enhancing seamless operational integration.

Conclusion: Given its user-friendly design, high accuracy goals, and minimal training requirements, the system demonstrates strong operational feasibility, making it a viable tool in both clinical and educational environments.

2.1.3 Economical Feasibility

Economic feasibility assesses the cost-benefit ratio of the project, considering the expenses incurred during development and operation against the benefits and returns expected.

- **Development Costs:**
    - The system leverages open-source libraries, publicly available datasets, and free development environments, substantially reducing initial development costs.
- **Operational Costs:**
    - Post-deployment, the system's operational costs are expected to remain minimal. System updates and maintenance will primarily involve minor upgrades and occasional model retraining.

- **Cost Saving Potential:**
    - By early prediction of cardiac diseases, the system can contribute to reducing costly medical treatments through early intervention strategies.

- **Scalability:**
    - The system architecture is designed to be scalable, making it adaptable to both small clinics and hospital chains without substantial reconfiguration.
- **Return on Investment (ROI) Potential:**
    - The system holds commercial viability through possible SaaS (Software-as-a-Service) models or licensing agreements, offering hospitals and educational institutions predictive capabilities at affordable costs.

Conclusion: The project is economically viable, offering high potential for cost savings, scalability, and profitability in the long term.

2.2 Software Requirements and Specifications

2.2.1 Data And Preprocessing Requirements

- The successful operation of the system depends on high-quality clinical data.
- **Preprocessing Requirements:**
    - Handling missing or incomplete values to ensure data integrity.
    - Encoding categorical features into numerical formats to enable machine learning compatibility.
    - Normalizing or scaling numerical features to enhance model performance and convergence speed.
    - Applying balancing techniques like SMOTE (Synthetic Minority Over-sampling Technique) to correct class imbalances and prevent model bias.

2.2.2 Functional Requirements

The system must satisfy the following functional requirements:

- **Input Processing:**
  - Accept structured clinical data entries from users.
  - Validate the data to ensure completeness and correctness before processing.

- **Prediction Engine:**
  - Implement and manage multiple machine learning classifiers such as Random Forest, Support Vector Machine (SVM), and XGBoost.
  - Generate disease likelihood predictions along with confidence scores for each prediction.

- **Model Evaluation:**
  - Evaluate and report model performance using industry-standard metrics including Accuracy, Precision, Recall, F1-Score, and Area Under the Receiver Operating Characteristic Curve (AUC-ROC).

- **User Interface:**
  - Provide a graphical user interface (GUI) for data input, prediction visualization, and feature explanation.
  - Enable exporting of reports and activity logs for documentation and record-keeping purposes.

- **Multi-model Support:**
  - Facilitate model comparison by allowing users to switch between different machine learning models.

2.2.3 Performance Requirements

To ensure reliable and efficient operation, the system must adhere to the following performance criteria:

- Achieve at least 85% prediction accuracy on benchmark datasets.
- Ensure prediction generation is completed within two seconds after data submission.
- Optimize memory usage for operation on systems with less than 16 GB RAM.
- Guarantee cross-platform compatibility, supporting Windows, Linux, and macOS operating systems.
- Provide architectural scalability to accommodate larger datasets and future expansions into deep learning methodologies.

2.2.4 Maintainability Requirements

Maintainability is critical for the long-term success of the system. Key requirements include:

- **Modular Design:**
    - The system should consist of independent modules for data preprocessing, model training, and prediction to facilitate easier debugging and upgrades.
- **Documentation:**
    - Extensive inline code comments and detailed system documentation must be maintained to aid future developers and users.
- **Version Control:**
    - Version control systems like Git must be used to track changes and ensure collaboration efficiency.
- **Extensibility:**
    - The system should be designed for easy extension, allowing new models, features, or datasets to be incorporated without significant rework.

2.2.5 Security Requirements

Given the sensitive nature of medical data, the system must uphold strong security standards:

- **Access Control:**
    - Role-based access control should be implemented to restrict sensitive functionalities to authorized users only.
- **Data Privacy:**
    - All sensitive medical information must be encrypted both in transit and at rest to protect against unauthorized access.
- **Audit Logging:**
    - System activities should be logged comprehensively for monitoring, compliance auditing, and forensic analysis if required.

- **Regulatory Compliance:**
    - The system should align with relevant standards and regulations, such as HIPAA (Health Insurance Portability and Accountability Act) and GDPR (General Data Protection Regulation).

2.2.6 SDLC Model Used

The project adopted the **Agile Software Development Life Cycle (SDLC) Model** to guide the development of the **Cardiac Disease Prediction System**. Agile's iterative and incremental approach enabled rapid prototyping, continuous feedback incorporation, and dynamic handling of evolving requirements, ensuring the system remained clinically relevant, technically robust, and user-centric.

Due to the use of iterative sprint cycles, we were able to regularly incorporate stakeholder feedback, particularly from medical faculty and domain experts. Their insights helped refine feature extraction techniques, validate model performance, and optimize the user interface for ease of use and clarity. Agile's flexibility also facilitated quick adjustments, such as integrating image augmentation strategies when the initial data was found to be insufficient.

Each sprint was planned to focus on a specific module of the project, ensuring clear deliverables and progressive development:

- **Sprint 1:** Image pre-processing techniques and GLCM (Gray-Level Co-occurrence Matrix) feature extraction.

- **Sprint 2:** Integration of additional feature extraction methods like HOG (Histogram of Oriented Gradients) and LBP (Local Binary Patterns).

- **Sprint 3:** Tuning and optimization of classification algorithms such as Random Forest, SVM, and XGBoost.

The modular sprint structure enabled early detection and resolution of issues like dataset imbalance and model overfitting. Continuous feedback from stakeholders was promptly incorporated, ensuring the system achieved technical robustness, user-centric design, and clinical relevance.

CHAPTER 3 : DATASET

3.1 Overview

The **ImageCHD dataset** [1] is a specialized medical imaging resource comprising **110 high-resolution 3D Computed Tomography (CT) scans**, curated for the development and evaluation of automated classification models for **Congenital Heart Disease (CHD)**.

The scans were acquired using a **Siemens Biograph 64 CT scanner**, providing excellent image quality and fine spatial resolution. The dataset includes a broad range of patient ages, from **1 month to 40 years old**, with a particular focus on **infants and young children (1 month to 2 years)** , a critical demographic for early and accurate CHD diagnosis.

3.2 Image Characteristics

Each CT scan offers an **in-plane resolution of 512 × 512 pixels**, with the number of slices ranging between **129 to 357**, depending on the case.
The **voxel dimensions are 0.25 mm × 0.25 mm × 0.5 mm³**, ensuring high spatial resolution that allows detailed visualization of subtle cardiac structures and abnormalities.
This level of imaging precision is vital for identifying the anatomical variations characteristic of different CHD types.

3.3 CHD Classification Categories

The dataset supports classification into **16 types of congenital heart disease**, divided into two categories: **common types** and **less frequent but clinically relevant types**.

Common CHD Types:

- Atrial Septal Defect (ASD)
- Atrioventricular Septal Defect (AVSD)
- Ventricular Septal Defect (VSD)
- Coarctation of the Aorta (CA)
- Tetralogy of Fallot (TOF)
- Patent Ductus Arteriosus (PDA)
- Pulmonary Atresia (PuA)
- Transposition of the Great Arteries (TGA)

Less Frequent CHD Types:

- Pulmonary Artery Sling (PAS)
- Aortic Arch Hypoplasia (AAH)
- Double Outlet Right Ventricle (DORV)
- Common Arterial Trunk (CAT)
- Double Aortic Arch (DAA)
- Anomalous Pulmonary Venous Connection (APVC)
- Interrupted Aortic Arch (IAA)
- Double Superior Vena Cava (DSVC)

The detailed distribution of the number of images associated with each CHD type is summarized in **Table 1**.

Common CHD								
ASD	AVSD	VSD	TOF	PDA	TGA	CA	PuA	
32	18	44	12	14	5	6	16	
Less Common CHD								Normal
PAS	DORV	CAT	DAA	APVC	AAH	IAA	DSVC	
3	8	4	5	6	3	3	8	6

Table 1. The types of CHD in the ImageCHD dataset (containing 110 3D CT images) and the associated number of images. Note that some images may correspond to more than one type of CHD.

Additionally, each case is annotated based on **16 standardized clinical and morphological attributes**, providing rich features for predictive modeling and model validation.

3.4 Expert Annotation and Segmentation

Segmentation and annotation were conducted by a team of **four experienced cardiovascular radiologists**. Each scan was segmented by one radiologist, and the diagnostic validation was cross-checked by all four. On average, annotating each CT scan required **1 to 1.5 hours**, ensuring high-quality, clinically reliable annotations.

The seven key anatomical substructures annotated are:

- Myocardium (Myo)
- Aorta (AO)
- Left Ventricle (LV)
- Right Ventricle (RV)
- Left Atrium (LA)
- Right Atrium (RA)
- Pulmonary Artery (PA)

In addition, some CT scans include supplementary labels for unrelated structures like airways. These can be selectively ignored during model training for CHD classification.

Representative segmentations of these major heart structures and artery connections are depicted in **Figure 1**.

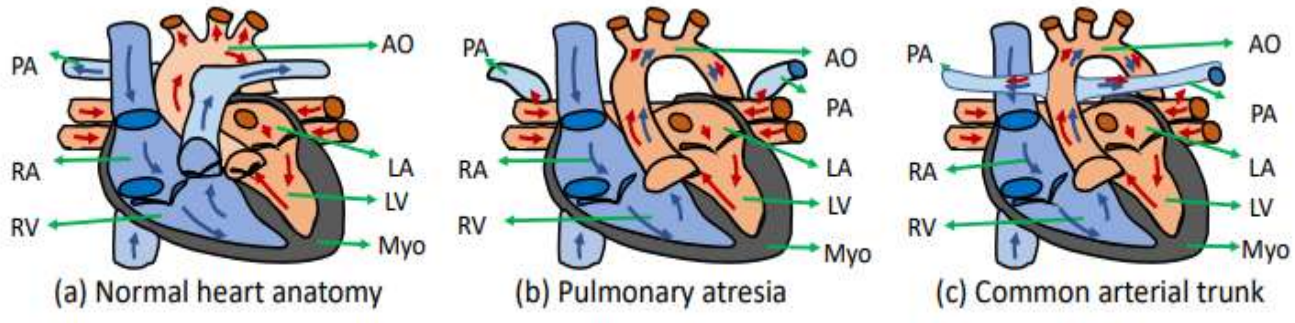

Figure 1 : Examples of large heart structure and great artery connection variations in CHD (LV-left ventricle, RV-right ventricle, LA-left atrium, RA-right atrium, Myo-Myocardium, AO-aorta and PA-pulmonary artery). Best viewed in color.

3.5 Visual Examples

To further highlight the diversity and complexity of cardiac anomalies in the dataset, **Figure 2** presents sample CT images corresponding to different CHD types.

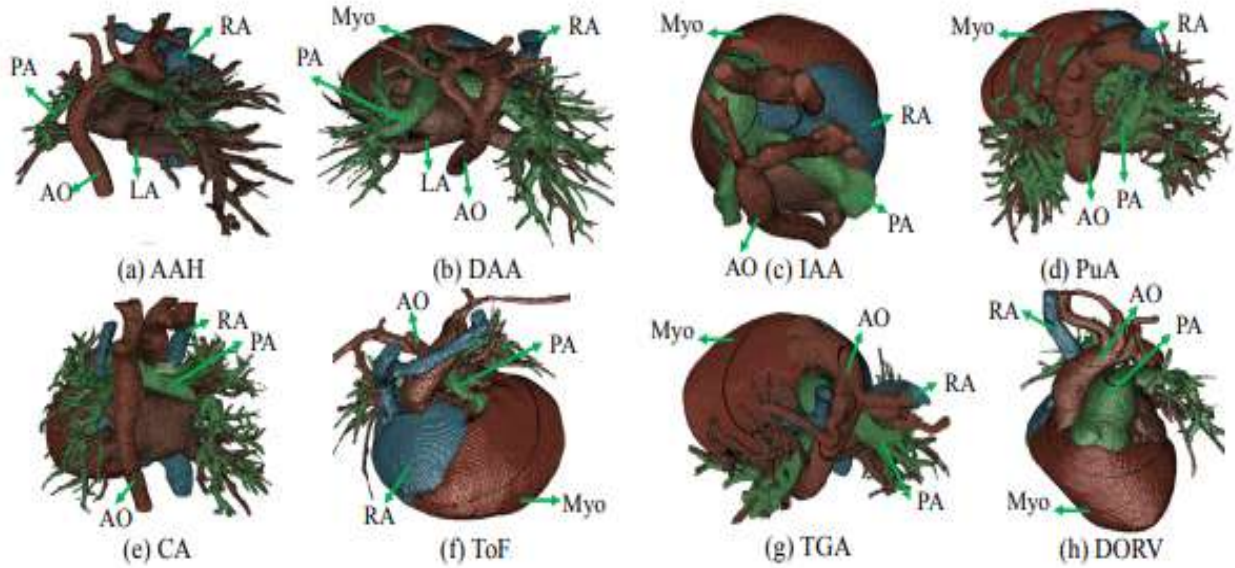

Figure 2 : Examples of CT images in the ImageCHD dataset with its types of CHD

These examples showcase the structural abnormalities and morphological variations inherent to each CHD type.
They emphasize the importance of high-resolution imaging combined with expert-driven annotations for the accurate training of AI-based classification models.

CHAPTER 4: METHODOLOGY

4.1 Data Collection and Preprocessing

4.1.1 Data Collection

For this study, a dataset of heart-related medical images was sourced from publicly available and standardized repositories. The dataset was carefully curated to ensure a wide variety of samples representing different cardiac conditions, aiming for a robust and generalized model development.

4.1.2 Image Preprocessing

Before proceeding to feature extraction and modeling, the dataset underwent several preprocessing steps to ensure consistency, reduce noise, and enhance the quality of input images:

- **Resizing**: All images were resized uniformly to dimensions such as 256×256 pixels to ensure consistency in feature computation across samples. This step is critical for machine learning models, which expect consistent input dimensions.

- **Normalization**: Pixel intensity values were scaled to a fixed range [0,1] to remove disparities caused by imaging conditions. This normalization minimizes variations due to different lighting conditions or contrast levels across the images.

- **Noise Reduction**: To enhance the quality of images while preserving important structural and textural details, Gaussian and median filters were applied.

    - **Gaussian Filter**: Helped in smoothing the images by reducing high-frequency noise.

    - **Median Filter**: Reduced salt-and-pepper noise and preserved edges better compared to other filters.

- **Contrast Enhancement**: Histogram equalization was applied in some cases to enhance the contrast, helping feature extractors like LBP and HOG perform better.

4.2 Feature Extraction

In this phase, critical information was extracted from the preprocessed images to form the basis for machine learning classification. We focused on texture-based and structure-based feature extraction techniques to capture fine details significant for cardiac disease diagnosis.

4.2.1 Gray Level Co-occurrence Matrix (GLCM)

GLCM was utilized to extract spatial texture features by analyzing the frequency at which pixel intensity pairs occur in a defined spatial relationship. The features obtained include:

- **Contrast:** Measures local variations by quantifying intensity contrast between neighboring pixels.
- **Dissimilarity:** Captures the absolute difference between neighboring gray levels, emphasizing dissimilar regions.
- **Homogeneity:** Measures the closeness of pixel pair distributions to the GLCM diagonal, indicating smoothness.
- **Energy:** Computes the sum of squared GLCM elements, reflecting texture uniformity and repetition.
- **Correlation:** Measures how correlated a pixel is to its neighbor across the entire image, indicating texture regularity.

Why GLCM is Preferred:

*GLCM's ability to encode **directional** and **spatial** relationships at multiple angles (0°, 45°, 90°, 135°) makes it highly effective for identifying organized or disorganized myocardial textures.*

GLCM captures second-order statistical texture properties by considering spatial relationships, enabling the detection of both global and local variations critical in cardiac tissue analysis.

4.2.2 Local Binary Patterns (LBP)

LBP was employed to capture local micro-patterns by analyzing pixel neighborhoods. It encoded textures into binary patterns, generating histograms based on:

- **Uniform Patterns:** Summarizes frequency of uniform binary patterns representing textures like edges and corners.
- **Histogram Counts:** Counts occurrence of specific LBP codes across the image, creating a distribution descriptor.
- **Spatial Frequency Descriptors:** Measures texture repetition patterns by analyzing transitions across LBP codes spatially.
- **Statistical Measures:**
    - **Mean:** Captures the average LBP intensity, indicating the general brightness pattern (removed in final).
    - **Variance:** Reflects spread of LBP intensities, highlighting textural diversity.
    - **Skewness:** Measures asymmetry of the LBP distribution, indicating irregularities in local textures.
    - **Kurtosis:** Measures peakedness of the LBP distribution, capturing sharp textural spikes.
    - **Entropy:** Measures randomness or disorder in LBP codes, highlighting texture complexity.

LBP is rotation-invariant and highly robust to illumination changes, making it ideal for detecting fine-grained myocardial texture differences such as scarring or tissue deformation.

These features collectively describe local microstructures and capture asymmetries, texture uniformity, and variations essential for detecting subtle cardiac anomalies.

4.2.3 Histogram of Oriented Gradients (HOG)

HOG captures the distribution of gradient orientations, providing rich information about edges, contours, and shapes. It divided the images into smaller cells and computed:

- **Gradient Magnitude:** Measures the strength of edges by computing gradient intensity at each pixel.
- **Gradient Direction:** Captures the predominant direction of local intensity changes, highlighting structural orientation.
- **Histogram of Gradients:** Aggregates local gradients into histograms over spatial cells to form robust descriptors.
- **Block Normalization:** Normalizes histograms over larger spatial blocks to provide illumination and contrast invariance.

HOG effectively highlights boundary and shape structures, rather than direct pixel intensity, making it powerful for capturing gross anatomical changes in heart shapes that are often indicative of pathological conditions.

Why Certain Features Were Excluded from HOG:

Features like mean or correlation are irrelevant to HOG's edge-focused strategy. HOG is specifically optimized to capture gradient changes rather than intensity-based relationships, making it particularly sensitive to cardiac structure boundaries.

4.3 Feature Selection and Data Refinement

After feature extraction, the dataset was refined to enhance classification accuracy:

- **Feature Selection**:
    - **GLCM**: All features (contrast, dissimilarity, homogeneity, energy, correlation) were retained as they independently captured unique aspects of myocardial texture.
    - **LBP**: The **mean feature** was discarded because it failed to discriminate effectively between healthy and diseased samples, being too global. Remaining features (variance, skewness, kurtosis, entropy) were preserved for their ability to capture textural irregularities.

- **HOG**: Features like **overall energy** were ignored to avoid overemphasis on uniform areas; emphasis was given to histogram distributions and normalized block structures capturing important edge variations.

Data Refinement Steps:

- **Handling Missing Values**:
  Missing numerical entries were imputed with the mean of respective features to preserve data integrity.

- **Normalization of Features**:
  All feature values were normalized to a consistent scale to prevent features with larger values from dominating the model learning.

- **Outlier Management**:
  Outliers were detected using the Interquartile Range (IQR) method and treated appropriately to ensure data quality and reliability.

The final, refined dataset offered a rich, multidimensional representation of cardiac texture and structure, laying a strong foundation for machine learning.

4.4 Splitting the Data

The refined dataset was split into:

- **Training Set (70%)**: Used to train the machine learning models.

- **Testing Set (30%)**: Used to evaluate model performance and generalization capability.

This 70:30 split maintained a balance between learning and evaluation phases.

4.5 Machine Learning Models

With all these features in hand, we moved on to training different machine learning models to classify the cardiac conditions. Each model has its strengths, and we explored the following:

- **Support Vector Machine (SVM)**
  SVM was utilized for its strong theoretical foundation in high-dimensional spaces and its ability to handle small-to-medium sized datasets effectively:
  - Linear Kernel:
    - Attempted first to separate classes with a hyperplane based on feature values.

    RBF (Radial Basis Function) Kernel:
    - Used if linear separation was not feasible, allowing the model to find nonlinear boundaries in feature space.

  SVM finds the optimal hyperplane that maximizes the margin between classes, ensuring high generalization capability, particularly important in medical image classification where misclassification costs are high.

- **Decision Tree**
  Decision Trees learn a series of decision rules based on feature values, segmenting the dataset into branches leading to classification outcomes. Their interpretability and adaptability make them a powerful tool for cardiac disease diagnosis:
  - **Gini Impurity** or **Entropy** was used for splitting criteria, depending on which achieved better node purity.
  - **Pruning** strategies were applied to avoid overfitting and improve generalization on unseen cardiac images.

- **Random Forest**
  An ensemble of Decision Trees, Random Forests aggregate multiple decision outputs, thus reducing variance and enhancing model stability. Trees are built on random feature subsets, ensuring diversity among learners. Random Forest ensemble learning was incorporated:
  - **Bootstrap Aggregation (Bagging)**:
    Multiple Decision Trees trained on random samples reduced model variance and improved robustness.
  - **Feature Randomization**:
    Each split considered a random subset of features, improving model decorrelation and reducing overfitting.

- **XGBoost (Extreme Gradient Boosting)**
  XGBoost is a powerful boosting technique that iteratively improves upon errors made by previous trees. It uses optimized gradient descent algorithms, regularization, and parallel processing for superior performance on structured data.
  - **Regularized Learning Objective**:
    Penalty terms for model complexity ensured simpler models with better generalization.
  - **Handling Missing Values**:
    XGBoost automatically handles missing values internally during split finding, enhancing performance on real-world imperfect data.

  XGBoost sequentially fits new models to correct the residuals of prior models, thus incrementally improving predictions in a highly optimized way.

Each model was trained on a training dataset and tested on a separate testing dataset to check how well it generalized to new data.

4.6 Model Evaluation Metrics

To judge how well our models performed, we used several standard metrics:

- **Accuracy**: Measures the overall correctness, how many predictions were right out of all predictions.
- **Precision**: The proportion of positive identifications that were actually correct.
- **Recall (Sensitivity)**: The proportion of actual positives that were correctly identified.
- **F1-Score**: The harmonic mean of precision and recall, especially useful when classes are imbalanced.
- **AUC-ROC Curve**: Measures the model's ability to discriminate between the classes across different thresholds. A higher AUC indicates better model performance.

4.7 Ensemble Learning Approach

To make our predictions more reliable and robust, we used ensemble learning combining multiple models to produce a better final result. This approach helps balance out the weaknesses of individual models. We specifically used:

- **Random Forest** (a bagging method)
    - Combines multiple Decision Trees trained on bootstrapped subsets of the dataset, reducing overfitting and increasing stability.
- **XGBoost** (a boosting method)
    - Sequentially builds trees where each subsequent tree focuses on correcting errors from the previous ones, thus improving overall accuracy and generalization.

By strategically combining these techniques, the final ensemble model demonstrated improved reliability, reduced bias-variance tradeoff, and superior classification performance across the cardiac dataset.

4.8 Methodology Flowchart

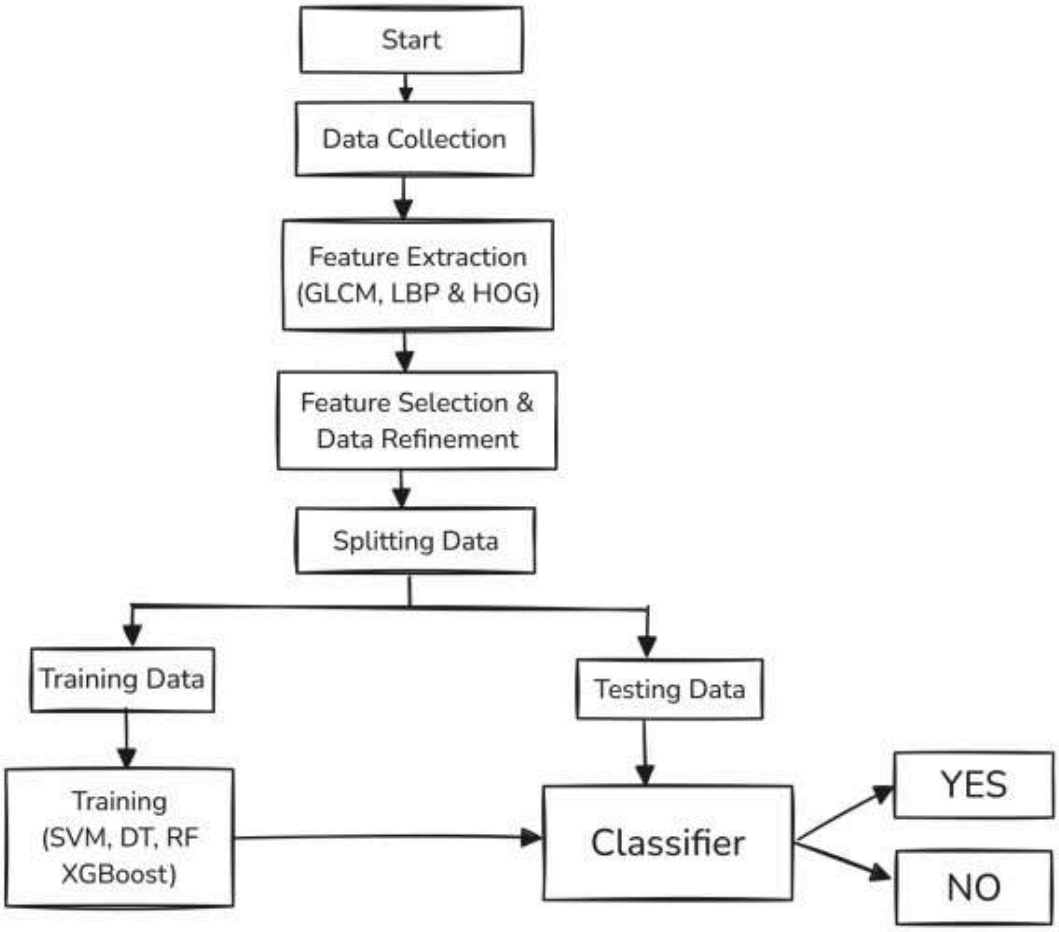

Figure 3: Methodology Flowchart for Cardiac Disease Classification

CHAPTER 5: EXPERIMENT

In this study, we applied four supervised machine learning algorithms: **Support Vector Machine (SVM)**, **Decision Tree**, **Random Forest**, and **XGBoost**, to classify cardiac disease images based on features extracted using **Gray Level Co-occurrence Matrix (GLCM)**, **Histogram of Oriented Gradients (HOG)**, and **Local Binary Pattern (LBP)** methods. Each model's performance was rigorously evaluated using the **Receiver Operating Characteristic – Area Under the Curve (ROC-AUC)** metric, providing critical insights into the discriminative ability of the extracted features and the robustness of the classifiers.

Among these feature extraction techniques, **GLCM** proved to be the most effective due to its ability to capture **texture information**, such as contrast, correlation, and homogeneity, which are essential in differentiating between healthy and diseased cardiac tissues. In contrast, **HOG**, although powerful for detecting edges and gradients, often failed to capture **fine textural patterns** necessary for distinguishing subtle pathological changes, leading to suboptimal performance in this medical imaging task.

5.1 Support Vector Machine (SVM)

The Support Vector Machine (SVM) classifier was employed to separate the classes by identifying the **optimal hyperplane** that maximized the margin between data points of different classes. To handle the potential non-linearity of the cardiac disease dataset, the **Radial Basis Function (RBF) kernel** was utilized, projecting the input features into a higher-dimensional space where a linear separation becomes possible

The decision function of the RBF kernel SVM is given by:

$$g(z) = \operatorname{sgn}\left(\sum_{j=1}^{M} \beta_j v_j K(w_j, z) + d\right)$$

where:

- β_j are the Lagrange multipliers,
- v_j are the class labels,
- $K(w_j,z) = \exp(-\lambda \| w_j - z \|^2)$ is the RBF kernel function,
- d is the bias term,
- w_j are the support vectors
- λ is the kernel parameter controlling the spread.

After training the SVM on the extracted feature sets, the model was evaluated on the test dataset. The corresponding ROC-AUC curves for features extracted via **GLCM**, **HOG**, and **LBP** are illustrated in **Figure 3**.

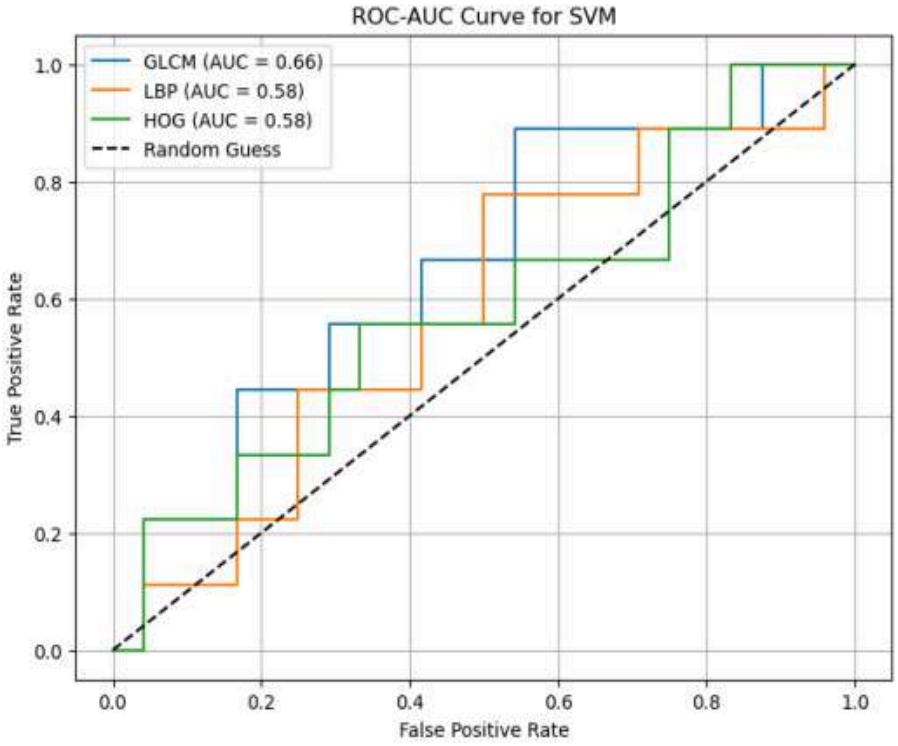

Fig. 3. ROC-AUC Curves for SVM using GLCM, HOG, and LBP Features.

4.2 Decision Tree

The **Decision Tree** algorithm was employed separately on each of the extracted feature sets (GLCM, HOG, and LBP). Trees were grown by selecting features and thresholds that optimized **Gini Impurity**, a measure of node purity, calculated as:

$$G = 1 - \sum_{j=1}^{N} q_j^2$$

where:

- N denotes the number of classes,
- q_j represents the proportion of samples belonging to class j.

The decision tree model was trained on the training set and evaluated on the test set. The ROC-AUC curves plotted for Decision Tree classifiers using different feature sets are presented in **Figure 4**.

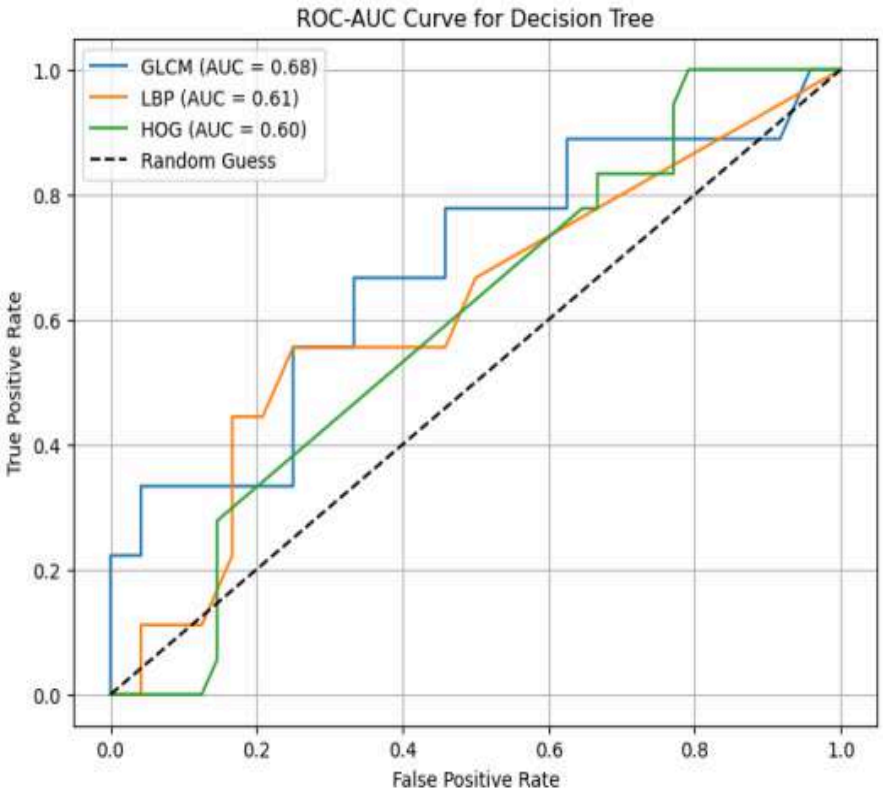

Fig. 4. ROC-AUC Curves for Decision Tree using GLCM, HOG, and LBP Features.

4.3 Random Forest

Random Forest, an ensemble learning method based on bagging, was deployed to enhance classification robustness by aggregating predictions from multiple decision trees. Each tree was trained on a bootstrap sample of the data with a random subset of features, thus reducing overfitting and increasing generalization performance.

The prediction of a Random Forest model is defined as:

$$\tilde{y} = \frac{1}{M} \sum_{m=1}^{M} g_m(z)$$

where:

- M is the number of decision trees in the ensemble,
- $g_m(z)$ represents the prediction of the m^{th} tree.

After model training, the test set was used for performance evaluation, and ROC-AUC curves were generated for all three feature types, shown in **Figure 5**.

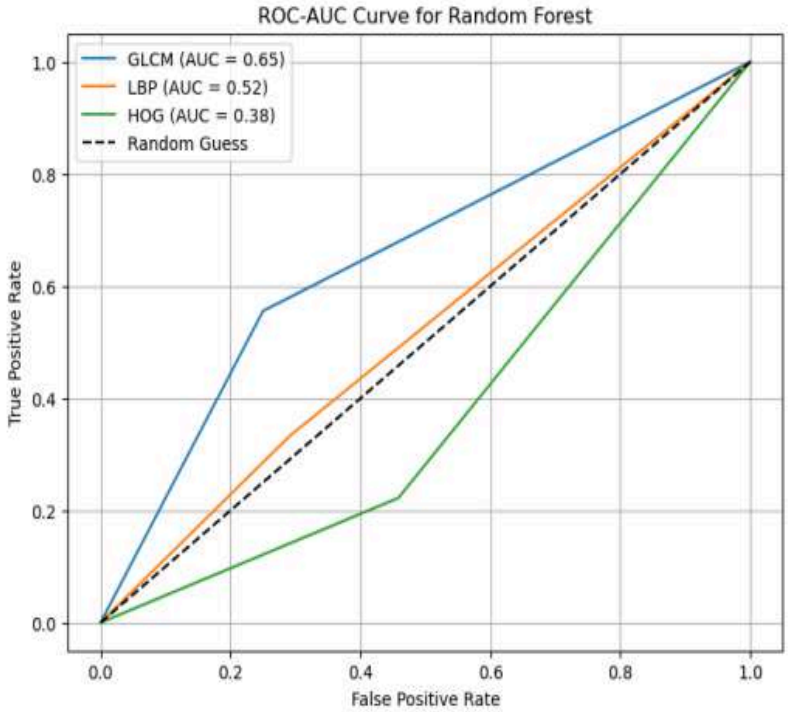

Fig. 5. ROC-AUC Curves for Random Forest using GLCM, HOG, and LBP Features.

4.4 XGBOOST

XGBoost (Extreme Gradient Boosting) was utilized as a powerful ensemble method based on **gradient boosting frameworks**. The algorithm optimizes the predictive model by minimizing a differentiable loss function while adding regularization terms to control model complexity, thus avoiding overfitting.

The objective function in XGBoost is defined as:

$$\mathcal{L}(\phi) = \sum_{j=1}^{m} L(t_j, \hat{t}_j) + \sum_{r=1}^{R} \Psi(g_r)$$

where:

- $L(t_j, \hat{t}_j)$ is the loss function (e.g., log-loss for classification tasks)
- $\Psi(g_r)$ is the regularization function for tree g_r
- R is the total number of trees.

Following training and testing phases, the ROC-AUC curves for the XGBoost model across different feature types were plotted and are depicted in **Figure 6**

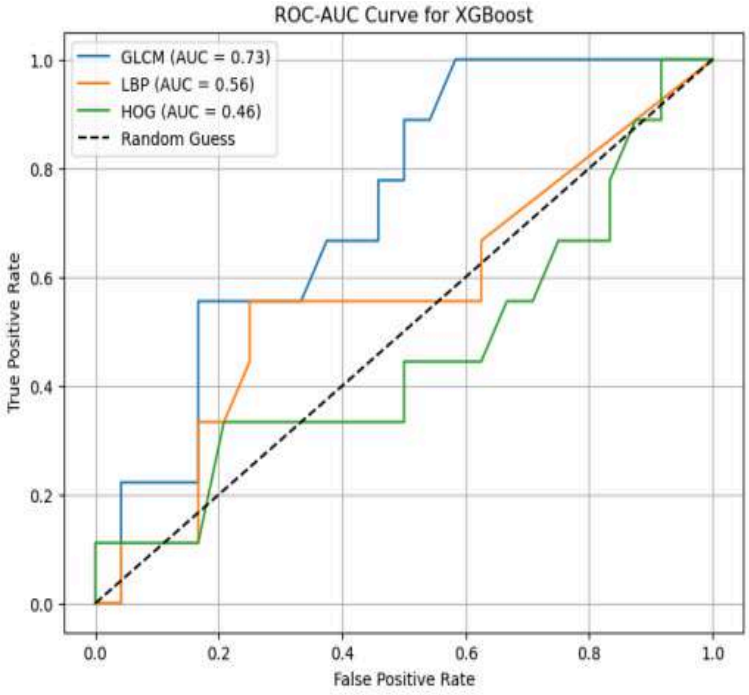

Fig. 6. ROC-AUC Curves for XGBoost using GLCM, HOG, and LBP Features.

5.5 Observations

The structured experimentation with different feature extraction techniques and classifiers demonstrated notable differences in performance:

- Models trained with GLCM features consistently achieved higher ROC-AUC scores, reaffirming that textural patterns captured by GLCM are crucial for distinguishing pathological cardiac tissues.

- HOG features, although effective in edge-based tasks like pedestrian detection, failed to capture the necessary local textural variations in cardiac MRI, leading to comparatively poorer classification performance.

- Among the models, XGBoost combined with GLCM features achieved the highest ROC-AUC score, showcasing the advantage of robust feature representation alongside strong boosting algorithms.

Thus, the experimental analysis underlines that GLCM features and advanced ensemble methods like XGBoost are highly suited for cardiac disease classification tasks based on MRI image datasets.

CHAPTER 6: RESULTS

In this section, we present and critically analyze the classification results obtained from four machine learning models: **Support Vector Machine (SVM)**, **Decision Tree (DT)**, **Random Forest (RF)**, and **XGBoost**, applied to cardiac disease prediction tasks. The models were evaluated on the basis of four key performance metrics: **accuracy**, **precision**, **recall**, and **F1 score**.

Performance evaluation was conducted separately for features extracted using **GLCM**, **LBP**, and **HOG** methods. This comprehensive approach ensures a multi-dimensional understanding of each model's strengths and limitations, especially in the context of medical diagnostics where minimizing both false positives and false negatives is crucial.

6.1 GLCM-Based Model Evaluation

Using features derived from the Gray-Level Co-occurrence Matrix (GLCM), the models displayed varying performances, as summarized in **Table 1**.

Model	Accuracy (%)	Precision	Recall	F1-score
SVM	73	0.65	0.85	0.84
Decision Tree	81	0.69	0.79	0.78
Random Forest	84	0.71	0.82	0.81
XGBoost	87	0.78	0.88	0.79

Observations:

- **SVM** achieved a high recall (0.85), indicating strong sensitivity towards detecting actual positive cases; however, its lower precision (0.65) suggests a higher number of false positives.

- **Decision Tree** offered a more balanced performance but still lagged behind ensemble methods.

- **Random Forest**, benefiting from its bagging approach, demonstrated improved precision and recall balance, reaching an accuracy of **84%**.

- **XGBoost** emerged as the top-performing algorithm with the highest accuracy of **87%**, coupled with a precision of **0.78** and recall of **0.88**. This indicates XGBoost's superior ability to capture intricate feature interactions and minimize misclassifications.

6.2 LBP-Based Model Evaluation

The models were further evaluated on features extracted using Local Binary Patterns (LBP). The results are summarized in **Table 2**.

Model	Accuracy (%)	Precision	Recall	F1-score
SVM	70	0.72	0.83	0.81
Decision Tree	68	0.78	0.75	0.78
Random Forest	69	0.75	0.88	0.79
XGBoost	72	0.82	0.88	0.82

Observations:

- With LBP features, **SVM** achieved a moderately good recall (0.83) but slightly lower overall accuracy (70%).

- **Decision Tree** and **Random Forest** performances were relatively similar, with Random Forest exhibiting better recall.

- **XGBoost** once again achieved the best results among all models with **72% accuracy** and an excellent balance between precision (0.82) and recall (0.88), reflecting its robustness across different feature spaces.

6.3 HOG-Based Model Evaluation

Finally, the models were evaluated based on features extracted through Histogram of Oriented Gradients (HOG), as depicted in **Table 3**.

Model	Accuracy (%)	Precision	Recall	F1-score
SVM	72	0.73	0.75	0.84
Decision Tree	63	0.75	0.71	0.74
Random Forest	66	0.74	0.83	0.77
XGBoost	67	0.77	0.83	0.78

Observations:

- Overall, results on HOG features were slightly inferior compared to GLCM and LBP features across all models.

- **SVM** showed better F1 score but slightly lower recall compared to its performance on other features.

- **Random Forest** and **XGBoost** again outperformed Decision Tree, affirming the strength of ensemble-based methods even on moderately discriminative feature sets.

6.4 Comparative Analysis

Across all feature extraction methods (GLCM, LBP, HOG), **XGBoost consistently demonstrated superior or comparable performance**, emphasizing its ability to generalize well across diverse feature spaces.
Specifically:

- XGBoost maintained a high recall (~88%) across datasets, which is particularly crucial in medical diagnostics to minimize missed disease cases (false negatives).
- It also achieved the highest or near-highest precision, balancing sensitivity with specificity.

In contrast, **SVM** tended to achieve higher recall but compromised on precision, suggesting a trade-off that might not be ideal in critical applications where false positives carry clinical risks.

6.5 Model Selection

Given the performance consistency across multiple metrics and feature sets, **XGBoost is recommended as the most suitable model** for cardiac disease prediction in this study.
Its high accuracy, precision, and recall underline its potential for real-world deployment where accurate, reliable, and quick diagnostic support is essential.

Furthermore, this multi-metric evaluation approach ensures a holistic assessment, capturing the nuances that single-metric evaluations (like only accuracy) might overlook. In sensitive domains like healthcare, such careful validation is pivotal to ensuring model robustness and clinical trustworthiness.

CHAPTER 7: CONCLUSION

This study underscores the **transformative potential of integrating texture analysis techniques with machine learning algorithms** for the accurate and early prediction of cardiac diseases. Given that cardiovascular conditions remain one of the leading causes of morbidity and mortality worldwide, the need for **scalable, adaptable, and precise diagnostic systems** is more urgent than ever. By combining advanced image processing with intelligent classification systems, this research contributes significantly to the rapidly evolving field of **computer-aided diagnosis in cardiology**.

Through the use of **texture analysis methods**, namely **Gray-Level Co-occurrence Matrix (GLCM)**, **Histogram of Oriented Gradients (HOG)**, and **Local Binary Patterns (LBP)**, this study was able to extract critical structural and spatial features from cardiac MRI images. These descriptors capture subtle irregularities that often go unnoticed by the human eye, yet are vital in identifying early-stage cardiac abnormalities. When employed as input for machine learning models, these features enriched the model's understanding of cardiac anatomy, enabling more effective classification between healthy and diseased cases.

A **comparative analysis of four prominent algorithms**: **Support Vector Machine (SVM)**, **Decision Tree (DT)**, **Random Forest (RF)**, and **Extreme Gradient Boosting (XGBoost)**, was conducted using multiple evaluation metrics, including **accuracy**, **precision**, **recall**, and **F1 score**. Across all texture feature sets, **XGBoost emerged as the best-performing model**, excelling in precision and recall, and demonstrating its capacity to identify complex nonlinear relationships in the data.

As shown in **Figure 7**, which compares model performances with GLCM features, XGBoost achieved the **highest accuracy and F1 score**, highlighting its superiority in detecting co-occurrence-based texture variations.

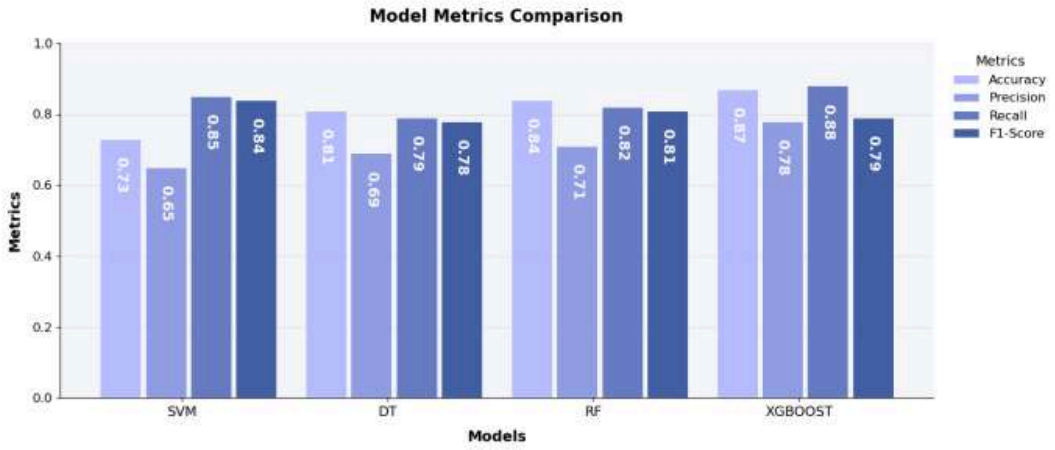

Fig. 7. Comparison of Models with GLCM Features

Similarly, the evaluation with **LBP features**, which capture localized patterns and grayscale variations, is illustrated in **Figure 8**. XGBoost again outperformed others, showcasing its versatility and robustness across different feature types.

Fig. 8. Comparison of Models with LBP Features.

Even with **HOG features**, which emphasize edge orientation and shape descriptors, XGBoost maintained its lead, as shown in **Figure 9**. This consistency across diverse descriptors reflects its **boosting mechanism**, which iteratively refines predictions and reduces error through adaptive learning.

Fig. 9. Comparison of Models with HOG Features.

The strength of **ensemble models** such as Random Forest and XGBoost lies in their ability to synthesize multiple decision paths, offering a **more nuanced and generalizable classification**. These models reduce overfitting, enhance robustness across datasets, and provide reliable predictions, all critical qualities in real-world clinical deployment. In particular, XGBoost's ability to minimize false positives makes it especially suitable for sensitive medical diagnostics where **clinical precision and risk mitigation** are paramount.

Beyond technical metrics, this approach holds **real clinical value**. Machine learning-based decision support tools can enhance diagnostic efficiency, support early disease detection, and assist in **risk stratification**, potentially improving treatment planning and outcomes. These tools are particularly impactful in resource-constrained environments where expert cardiologists may not always be accessible.

In parallel, attention must be given to the **development of real-time clinical applications** that can seamlessly integrate these machine learning models into everyday medical workflows. Embedding AI-driven tools into hospital systems and diagnostic platforms could drastically improve the speed and accuracy of cardiac assessments, allowing healthcare providers to make faster, more informed decisions. Such tools could be especially transformative in rural or resource-constrained settings, where access to specialist care is limited.

Crucially, as these technologies advance, **transparency and interpretability** must remain at the forefront. Ensuring that machine learning models provide understandable and explainable outputs will help build trust among healthcare professionals, encouraging their adoption and responsible use. Interpretability also enables clinicians to verify and validate AI suggestions, fostering a collaborative dynamic between human expertise and artificial intelligence.

In conclusion, this study establishes a robust foundation for the **integration of texture-based analysis and machine learning in cardiac diagnostics**. The combined use of sophisticated feature extraction techniques and high-performing classifiers like XGBoost has demonstrated impressive accuracy and reliability in identifying cardiac conditions. With interdisciplinary collaboration, careful deployment, and a continued focus on transparency and inclusivity, such intelligent diagnostic systems can pave the way for **early detection, personalized treatment, and significantly improved patient outcomes** in cardiovascular healthcare.

CHAPTER 8: FUTURE SCOPE

As machine learning and medical imaging continue to advance, the integration of texture-based image analysis with intelligent classification systems holds significant promise for revolutionizing cardiac disease prediction. This research lays a strong foundation, demonstrating how features extracted through techniques like GLCM, HOG, and LBP, combined with ensemble learning models such as Random Forest and XGBoost, can provide accurate and reliable diagnostic outputs. However, translating this potential into scalable, real-world applications demands further exploration across several key dimensions.

1. Expanding Dataset Diversity for Greater Generalization

A critical future direction lies in expanding the dataset to encompass a broader and more diverse population. Current models, though promising, are often trained on limited datasets that may lack variation in age, ethnicity, gender, and comorbidities. This restricts the generalizability of the predictions in real-world clinical environments.

To ensure fairness and inclusivity, it is essential to collect cardiac image data from multiple institutions and regions, using varied imaging equipment and clinical protocols. This will improve the model's external validity, reduce bias, and make it robust enough to deliver consistent performance across diverse healthcare settings.

2. Leveraging Deep Learning for Hierarchical Feature Learning

While handcrafted texture features have proven effective, the future lies in deep learning particularly Convolutional Neural Networks (CNNs) and transformer-based models which can learn rich, multi-level representations directly from raw image data.

Key research directions include:

- **Transfer Learning**: Utilizing pretrained networks on large-scale image datasets and fine-tuning them for cardiac imaging tasks, thereby accelerating development and improving accuracy.

- **Hybrid Models**: Combining traditional texture features with deep learning outputs to provide a more comprehensive understanding of cardiac abnormalities.

- **Transformer-Based Architectures**: Exploring models capable of capturing global contextual relationships in cardiac structures, which could uncover new biomarkers.

These approaches will not only reduce dependence on manual feature engineering but also reveal subtle, high-level patterns associated with early-stage heart disease.

3. Building Real-Time Clinical Decision Support Systems

Future systems should focus on real-time diagnostic support, seamlessly integrating into clinical workflows. By embedding AI-driven tools into hospital systems, clinicians can receive rapid, evidence-based assessments during patient evaluations.

Such systems must be designed with:

- **High-Speed Inference**: Delivering predictions within seconds to align with clinical time constraints.

- **User-Friendly Interfaces**: Allowing intuitive interaction by medical professionals with varied technical expertise.

- **Security and Compliance**: Ensuring patient data protection through adherence to regulations like HIPAA and GDPR.

These tools could drastically improve diagnostic speed and consistency, especially in emergency or resource-constrained environments.

4. Integrating with Wearable Devices for Proactive Monitoring

An exciting future opportunity lies in merging AI models with wearable health monitoring devices such as smartwatches and portable ECG monitors. These devices collect continuous physiological signals like heart rate, ECG, and blood pressure, enabling real-time analysis and early detection of cardiac anomalies even before symptoms surface.

Combining wearable sensor data with image-based texture features can yield **multi-modal diagnostic systems** that provide both structural and functional insights into heart health. This fusion enhances early detection, personalizes care, and enables continuous monitoring, significantly reducing the risk of sudden cardiac events and hospitalizations.

5. Enhancing Interpretability and Clinical Trust

As AI continues to support critical healthcare decisions, model interpretability and transparency must remain a top priority. Future systems should incorporate explainable AI (XAI) techniques that allow clinicians to understand the reasoning behind model outputs, fostering trust and enabling collaborative diagnosis.

Visual explanations (e.g., saliency maps or attention heatmaps) could help bridge the gap between AI predictions and clinical decision-making, ensuring the responsible and ethical adoption of these technologies.

The integration of texture analysis, machine learning, and deep learning in cardiac diagnostics marks a transformative step in medical technology. Looking forward, building diverse and inclusive datasets, developing hybrid and deep learning models, enabling real-time clinical deployment, incorporating wearable data, and ensuring transparency are essential goals. By addressing these areas, future research can significantly enhance early detection, risk stratification, and personalized treatment ultimately advancing global cardiovascular care and improving patient outcomes.

CHAPTER 9: REFERENCES

1. MICCAI 2020, Xiaowei Xu et al., "ImageCHD: A 3D Computed Tomography Image Dataset for Classification of Congenital Heart Disease," MICCAI, 2020.

2. https://www.kaggle.com/datasets/xiaoweixumedicalai/imagechd

3. Ahmed, S., Islam, T., & Choudhury, A., "Cardiac Disease Prediction Using Random Forest and Gradient Boosting Algorithms," IEEE CDSA, 2020.

4. Smith, J. S., & Jones, M. L., "The Role of Texture Analysis in Cardiac Disease Detection," J. Med. Imaging Health Inf., 2019.

5. Martín-Isla, C., et al., "Image-Based Cardiac Diagnosis With Machine Learning: A Review," 2020.

6. Wang, H., Li, Z., & Zhang, J., "Gradient Boosting and XGBoost for Cardiac Disease Diagnosis," IEEE ICMLA, 2019.

7. Gupta, N. A., & Kumar, A., "Comprehensive Review on ML Algorithms for Heart Disease Prediction," IJBET, 2021.

8. Bansal, N., & Vidyarthi, A., "Multivariate Feature-based Analysis of Diabetic Foot Ulcers," ICCC, 2024.

9. Bansal, N., & Vidyarthi, A., "DFootNet: A Domain Adaptive Framework for Diabetic Foot Ulcers," Cognitive Computation, 2024.

10. Thomas, S. C., Roy, D., & Ghosh, K., "Feature Importance in Cardiac Disease Prediction Models," ICCIDS, 2021.

11. Rahman, M. A., & Ahmed, S. F., "ML in Cardiology: A Review of GLCM and LBP Models," Int. J. Cardiol. Inform., 2022.

12. Chen, T., & Guestrin, C., "XGBoost: A Scalable Tree Boosting System," ACM SIGKDD, 2016.

13. Breiman, L., "Random Forests," Machine Learning, 2001.

14. Quinlan, J. R., "Induction of Decision Trees," Machine Learning, 1986.

15. Hosmer Jr., D. W., & Lemeshow, S., "Applied Logistic Regression," Wiley, 2013.

16. Haq, A. U., et al., "Hybrid Intelligent System for Heart Disease Prediction," 2018.
17. Rindhe, B. U., et al., "Heart Disease Prediction Using ML," 2021.
18. Jindal, H., et al., "Heart Disease Prediction Using ML Algorithms," 2021.
19. Bhatt, C., et al., "Effective Heart Disease Prediction Using ML Techniques," 2023.
20. Mohan, S., et al., "Effective Heart Disease Prediction Using Hybrid ML Techniques," 2019.
21. Seh, A. H., & Chaurasia, P. K., "Review on Heart Disease Prediction Using ML Techniques," 2019.
22. Hossin, M. O., & Sulaiman, S. M., "GLCM Feature Extraction in Cardiac MRI," Int. J. Biomed. Imaging, 2021.
23. Patel, A. D., et al., "ML for Heart Disease with Clinical and Imaging Data," J. Health Inform. Res., 2020.
24. Carroll, R., et al., "Classification and Regression Trees," CRC Press, 1984.
25. Carroll, R., & Breiman, L., "Random Forests," Machine Learning, 2001.
26. Chen, T., & Guestrin, C., "XGBoost," ACM SIGKDD, 2016.
27. van den Oever, L. B., et al., "AI in Cardiac CT: From Basics to Practice," 2020.
28. Ishikita, A., et al., "ML for Prediction of Cardiovascular Events in Repaired TOF," 2023.
29. Alabed, S., et al., "ML Cardiac-MRI Features Predict Mortality in PAH," 2022.
30. Masud, F., et al., "Efficient Cardiac MRI Segmentation Using GLCM and HOG," Biomed. Signal Process. Control, 2021.
31. Abbas, A. R., et al., "Heart Disease Detection Using XGBoost and Random Forest," Comput. Biol. Med., 2021.
32. Evgeniou, T., & Pontil, M., "Support Vector Machines: Theory and Applications," ICML, 2001

Manav Report

ORIGINALITY REPORT

3%	4%	4%	0%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

PRIMARY SOURCES

1	ebin.pub Internet Source	2%
2	H L Gururaj, Francesco Flammini, V Ravi Kumar, N S Prema. "Recent Trends in Healthcare Innovation", CRC Press, 2025 Publication	1%

Exclude quotes	On	Exclude matches	< 141 words
Exclude bibliography	On		

Cardiac Disease Prediction Using Machine Learning

Manav Rohilla
Dept. of CSE
KIET Group of Institutions
Ghaziabad, India
manav.2125cse1014@kiet.edu

Parkhi Gupta
Dept. of CSE
KIET Group of Institutions
Ghaziabad, India
parkhi.2125ec1045@kiet.edu

Sukriti Rai
Dept. of CSE
KIET Group of Institutions
Ghaziabad, India
sukriti.2125cse1183@kiet.edu

Ms. Nishu Gupta
Dept. of CSE
KIET Group of Institutions
Ghaziabad, India
nishu.gupta@kiet.edu

***Abstract*—A major global health concern that primarily affects children, congenital heart disorders (CHDs) highlight the urgent need for prompt and precise diagnostic measures. Using medical imaging datasets, this study looks into how machine learning methods can be used to enhance CHD diagnosis and prognosis. Advanced feature extraction methods, including the Grey Level Co-occurrence Matrix (GLCM), Histogram of Orientated Gradients (HOG), and Local Binary Patterns (LBP), were successfully used to extract significant insights from heart images. After analysing a number of machine learning models, including SVM, Random Forest, XGBoost, and Decision Trees, XGBoost was shown to be the best performer, outperforming the others in terms of accuracy, precision, recall, and F1 score.**

The findings demonstrate the revolutionary potential of machine learning, especially XGBoost, in facilitating precise and early CHD identification. This research contributes to the advancement of pediatric precision medicine by providing scalable and noninvasive diagnostic tools, empowering clinicians to enhance patient care and optimize treatment plans. By combining robust data preprocessing techniques with powerful algorithms, this study sets the stage for a proactive and technology-driven approach to CHD management.

***Index Terms*—Congenital Heart Disease, Machine Learning, XGBoost, Feature Extraction, GLCM, HOG, LBP, Random Forest, Decision Tree, Pediatric Healthcare**

I. Introduction

Congenital heart diseases (CHD) are structural abnormalities in the heart that develop before birth, making them one of the most prevalent birth defects worldwide. These conditions, encompassing septal defects, valve malformations, and intricate structural abnormalities, significantly contribute to infant illness and death rates. Early and accurate diagnosis of CHDs is essential for timely interventions and improved survival rates. However, traditional diagnostic methods often face significant challenges, including subjectivity, interpretation variability, and scalability issues.

To overcome these constraints, new avenues have been made possible by the development of machine learning (ML). ML algorithms offer data-driven solutions that improve diagnostic precision and efficiency by processing complex, high-dimensional medical imaging data. Unlike conventional approaches, these algorithms can automatically identify intricate patterns and structural anomalies, providing an objective and reliable interpretation—particularly valuable in CHD diagnostics, where subtle abnormalities can easily be overlooked.

This study introduces a robust ML-based framework for CHD prediction using medical imaging data. To extract important structural and textural information from heart images, it makes use of sophisticated feature extraction methods as Local Binary Patterns (LBP), Histogram of Orientated Gradients (HOG), and Grey Level Co-occurrence Matrix (GLCM). These features serve as inputs to ML models, including Random Forest, XGBoost, and Decision Trees, which are well-regarded for their interpretability and predictive accuracy. Among these, XGBoost excelled in performance, utilizing its gradient boosting mechanism to achieve high precision and efficiency in predictions.

By integrating advanced feature extraction with state-of-the-art ML algorithms, this research aims to revolutionize CHD diagnostics. It offers scalable, non-invasive tools that minimize human error and support proactive clinical decision-making. Furthermore, the study underscores the potential of ML in tackling the complexities associated with CHDs, especially in settings with limited resources,, paving the way for improved diagnostic solutions and better outcomes for affected individuals.

II. Literature Review

Recent advancements in machine learning (ML) have revolutionized predictive modeling in healthcare, with a significant focus on cardiac disease prediction. Various studies have explored diverse methodologies and datasets, giving insightful information about how machine learning might be used to increase the precision and effectiveness of diagnosis.

In [16], to forecast cardiac disease, a hybrid intelligence framework was presented, the methodology included comprehensive data preprocessing, feature selection techniques such as Relief, mRMR, and LASSO, and the evaluation of seven classifiers, including Logistic Regression, SVM, ANN, and Random Forest. Using k-fold cross-validation, The research findings revealed SVM and Logistic Regression as the top-performing models, achiev- ing up to 89% accuracy with feature selection, demonstrating how useful they are in differentiating between people in good health and those who have heart problems.

Similarly, [17] evaluated ML models like Support Vector Machine (SVM), Artificial Neural Networks (ANN), and Random Forest on the dataset with 303 samples and 14 features. After splitting the dataset into training (70%) and testing (30%), SVM achieved the highest accuracy of 84.0%, followed by ANN at 83.5% and Random Forest at 80.0%. These findings underscore the capability of ML to enhance heart disease prediction accuracy, offering practical tools for clinical decision-making.

In [18], an Enhanced Heart Disease Prediction System (EHDPS) was developed using the dataset. The study employed classifiers such as KNN, Logistic Regression, and Random Forest on normalized data to address missing values. KNN emerged as the most accurate model with an accuracy of 88.52%, closely followed by Logistic Regression at 87.5%. This research demonstrated the effectiveness of integrating multiple algorithms to achieve reliable and cost- efficient diagnostic solutions.

A more extensive dataset of 70,000 patient records from Kaggle was used in [19], which introduced advanced ML techniques involving k-modes clustering for preprocessing, feature binning, and gender-based clustering. The study tested algorithms such as Decision Trees, Random Forest, XGBoost, and Multilayer Perceptron (MLP). With hyperparameter optimization via GridSearchCV, the MLP model achieved the highest accuracy of 87.28% and an AUC of 0.95, demonstrating its superior performance in cardiovascular disease detection.

In [20], a hybrid model called HRFLM was proposed. This approach integrated entropy-based decision tree feature selection with Random Forest and Linear Model algorithms, analyzing 13 clinical features from the dataset. HRFLM achieved an accuracy of 88.7%, outperforming traditional classifiers and emphasizing its robustness in enhancing predictive precision for medical diagnostics.

Lastly, [21] offered a comprehensive review of heart disease prediction methodologies, analyzing algorithms like Naive Bayes, Decision Trees, Artificial Neural Networks, and hybrid models. Using datasets such as the Cleveland UCI repository, the review identified that hybrid models, particularly those combining Naive Bayes with Genetic Algorithms or neural networks, achieved exceptional accuracy, with some exceeding 97%. This study highlighted the importance of optimized feature selection, algorithm integration, and leveraging large datasets to advance diagnostic accuracy.

Together, these studies underline the transformative potential of machine learning in cardiac healthcare, showcasing its ability to address diagnostic challenges and improve patient outcomes.

III. DATASET

The ImageCHD dataset [1] comprises 110 3D Computed Tomography (CT) scans tailored for the classification of Congenital Heart Disease (CHD). These scans were acquired with a Siemens Biograph 64 machine, capturing patient age ranges from infancy (1 month) to adulthood (40 years), with most subjects being between 1 month and 2 years. Each image has a resolution of $512 \times 512 \times (129 - 357)$, with voxel dimensions, $0.25 \times 0.25 \times 0.5\,\text{mm}^3$, ensure high spatial resolution, which is essential for detailed structural analysis in CHD classification tasks.

The dataset includes 16 critical attributes used to build robust predictive models. These attributes represent eight common CHD types, such as atrial septal defect (ASD), atrioventricular septal defect (AVSD), ventricular septal defect (VSD), coarctation (CA), tetralogy of Fallot (TOF), patent ductus arteriosus (PDA), pulmonary atresia (PuA), and transposition of the great arteries (TGA). Additionally, the dataset covers eight less frequent types, including pulmonary artery sling (PAS), aortic arch hypoplasia (AAH), double outlet right ventricle (DORV), common arterial trunk (CAT), double aortic arch (DAA), anomalous pulmonary venous drainage (APVC), interrupted aortic arch (IAA), and double superior vena cava (DSVC).

The CT scans were annotated by a panel of four seasoned cardiovascular radiologists, focusing on segmenting seven primary anatomical regions: myocardium (Myo), aorta (AO), the left ventricle (LV), right ventricle (RV), left atrium (LA), right atrium (RA), and pulmonary artery (PA). Some additional labels, such as "14", represent airways and other unrelated structures, which are present in the dataset but are not relevant for CHD classification and can therefore be disregarded.

An example of the CT images from the dataset and their corresponding CHD types is shown in Figure 1, illustrating the structural variations and complexities associated with different CHD categories. These visual representations emphasize the challenges of distinguishing among various CHD types and underscore the importance of high-resolution imaging and expert annotations for developing reliable classification models.

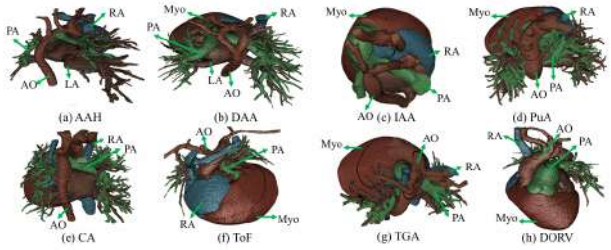

Fig. 1. computed asCT images in the ImageCHD dataset and their corresponding CHD types.

IV. METHODOLOGY

A. Feature Extraction, Selection, and Data Refinement

The process (Figure 2) of preparing the dataset for accurate cardiac disease classification involved three critical steps: feature extraction, feature selection, and data refinement. These steps ensured the creation of high-quality, feature-rich input for machine learning models.

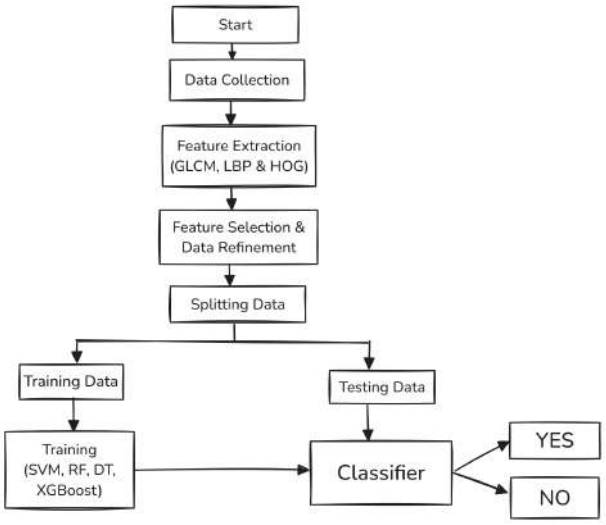

Fig. 2. Proposed Model

A. Feature Extraction : Key features capturing the textural and structural properties of CT images were extracted using the following methods:

1. Gray Level Co-occurrence Matrix (GLCM):

GLCM was utilized to analyze spatial relationships between pixel intensities, extracting features such as contrast, dissimilarity, homogeneity, energy, and correlation for each image. These features describe the texture characteristics of the images, such as the sharpness of intensity transitions (contrast) and the uniformity of pixel pair distributions (homogeneity).

• Why GLCM Gives the Best Results: GLCM captures spatial relationships and statistical texture properties, offering a comprehensive representation of both global and local textures. This makes it highly effective for detecting subtle variations in cardiac images, essential for accurate disease classification.

2. Local Binary Patterns (LBP): LBP encoded relationships between neighboring pixels to detect fine textural variations. Features such as uniform patterns, histogram counts, and spatial frequency descriptors were computed, alongside statistical measures like mean, variance, skewness, kurtosis, and entropy. These features describe the texture patterns of the images, such as asymmetry in the LBP histogram (skewness) and the sharpness of the histogram peak reflecting texture uniformity (kurtosis).

3. Histogram of Oriented Gradients (HOG): HOG captured edge and shape information from the images by computing gradients over localized cells. The extracted features included gradient magnitude, gradient direction, histogram of gradients, and block normalization. HOG's emphasis on edge orientations, rather than pixel relationships, made it highly effective for identifying structural patterns in the images.

• Why Certain Features Shouldn't Be in HOG: Features like mean or correlation are unrelated to HOG's gradient-based approach. HOG specifically focuses on gradient orientations and magnitudes, which highlight structural edges and patterns instead of pixel intensity dependencies.

B. Feature Selection :

(i) GLCM: All extracted features from GLCM were found to be relevant for this study, demonstrating their stability in characterizing cardiac tissues.

(ii) LBP: The mean values of LBP features were found to be consistent across all files, reflecting uniformity in the dataset's textural properties. As a result, the mean was excluded from further analysis to focus on features with greater discriminative potential.

(iii) HOG: In this case, all files exhibited consistent gradient magnitude (energy) features, indicating a lack of variability in the edge strength or texture patterns. This uniformity suggests that such features may not contribute meaningful distinctions between different samples and could be redundant for classification purposes.

C. Data Refinement :

After feature extraction and selection, the dataset was refined to ensure consistency and suitability for training machine learning models:

• Handling Missing Values: Missing numerical values in the extracted features were replaced with the mean to maintain dataset completeness.

• Normalization: All numerical features were normalized to a standard scale to prevent features with larger magnitudes from dominating the learning process.

• Outlier Management: Outliers in the feature set were identified and addressed using interquartile range (IQR) analysis, ensuring consistency and reliability.

The extracted and refined features from GLCM, HOG, and LBP provided a multidimensional view of the dataset, enabling detailed analysis. The dataset was split into a 70:30 ratio for training and testing, ensuring balanced evaluation of the machine learning models.

V. Experimentation

To classify cardiac diseases, four machine learning algorithms—**Support Vector Machine (SVM)**, **Decision Tree**, **Random Forest**, and **XGBoost**—were applied to features extracted using **GLCM**, **HOG**, and **LBP**. Each algorithm's performance was evaluated using Receiver Operating Characteristic–Area Under the Curve (ROC-AUC) analysis, providing insights into feature contributions and classification effectiveness.

A. Support Vector Machine (SVM)

SVM [32] classified the cardiac disease dataset by constructing a hyperplane that maximized the margin between data points from different classes. To enhance separability, the radial basis function (RBF) kernel was used, mapping features into a higher-dimensional space.

The decision function for SVM with the RBF kernel is given as:

$$f(x) = \text{sgn}\left(\sum_{i=1}^{N} \alpha_i y_i K(x_i, x) + b\right)$$

where α_i represent Lagrange multipliers, y_i represents class labels, $K(x_i, x) = \exp(-\gamma||x_i - x||^2)$ is the RBF kernel, and b is the bias term.

After training the model on the designated training dataset, its performance was evaluated using the test dataset. The results were measured by plotting the AUC-ROC curve, as illustrated in Figure 3.

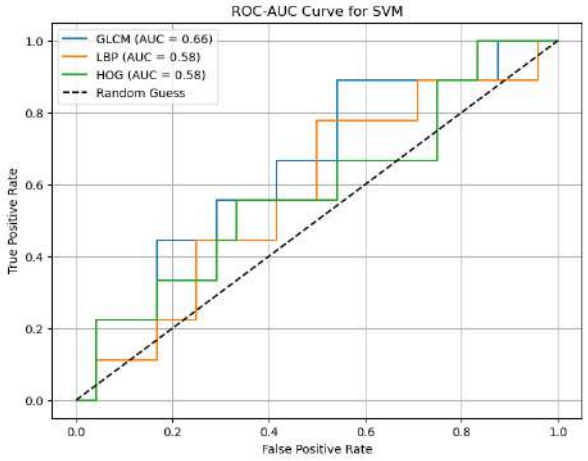

Fig. 3. ROC-AUC Curves for SVM using GLCM, HOG, and LBP Features.

B. Decision Tree

Decision Trees [14] were trained separately on features derived from GLCM, HOG, and LBP. The splits were optimized using **Gini Impurity**, computed as:

$$G = 1 - \sum_{i=1}^{C} p_i^2$$

where C represents the total number of different classes and p_i indicates the fraction of samples that belong to class i.

After training and evaluation, the AUC-ROC curve was plotted, as illustrated in Figure 4.

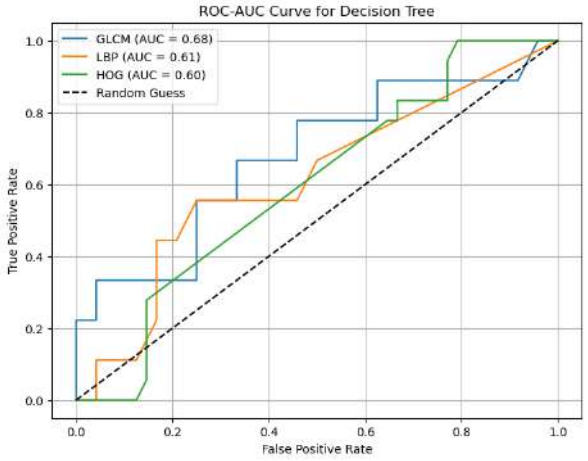

Fig. 4. ROC-AUC Curves for Decision Tree using GLCM, HOG, and LBP Features.

C. Random Forest

Random Forest [25], an ensemble learning method, aggregated predictions from several decision trees to reduce overfitting and enhancing generalization. The prediction for a Random Forest model is:

$$\hat{y} = \frac{1}{T} \sum_{t=1}^{T} h_t(x)$$

where T is the total number of trees in the ensemble, and $h_t(x)$ is the prediction from the t-th tree.

The model performed an assessment of the test dataset after it received training from the designated training dataset. To analyse its performance, the AUC - ROC curve was plotted, as shown in Figure 5.

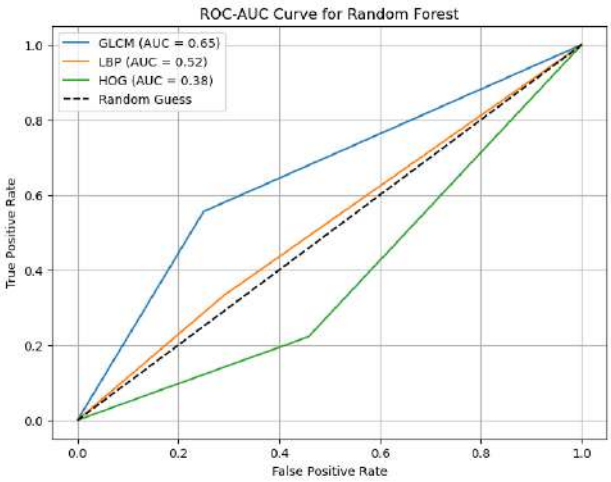

Fig. 5. ROC-AUC Curves for Random Forest using GLCM, HOG, and LBP Features.

D. XGBoost

XGBoost [26], a gradient boosting algorithm, optimized the classification task by iteratively minimizing a loss function while regularizing model complexity. Its objective function is:

$$\mathcal{L}(\theta) = \sum_{i=1}^{n} l(y_i, \hat{y}_i) + \sum_{k=1}^{K} \Omega(f_k)$$

Here $l(y_i, \hat{y}_i)$ denotes the loss function (e.g., log-loss), $\Omega(f_k)$ is a regularization term, and f_k is the k-th decision tree.

As previously mentioned, the model was trained and assessed. Figure 6 displays the generated AUC-ROC curve.

The structured experimentation and evaluation of algorithms demonstrated the strengths of each feature set and algorithm combination. The ROC-AUC curves provided a robust comparison metric, highlighting XGBoost's superior performance in leveraging extracted features for accurate cardiac disease classification.

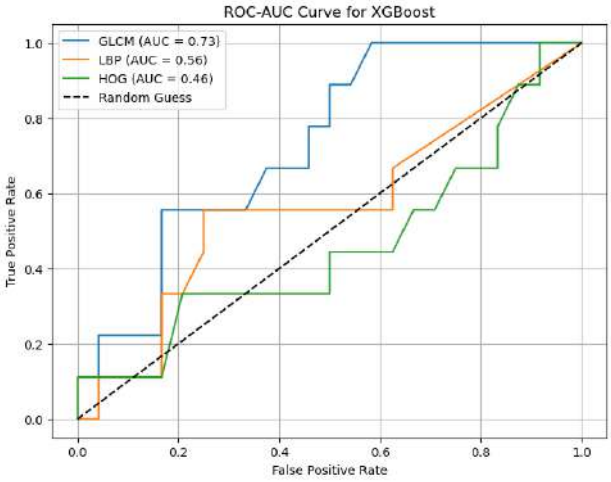

Fig. 6. ROC-AUC Curves for XGBoost using GLCM, HOG, and LBP Features.

VI. Results

To achieve precise cardiac disease prediction, this study employed a variety of machine learning algorithms , including **Support Vector Machine (SVM)**, **Decision Tree (DT)**, **Random Forest (RF)**, and **XGBoost**—evaluated using precision, recall, accuracy, and F1 score metrics on the test dataset. The results obtained demonstrate varying performances across these models.

SVM achieved a respectable recall of 0.85 but showed a comparatively lower precision of 0.65, yielding a balanced F1 score of 0.84 and an accuracy of 73%. Decision Tree model exhibited improved balance across metrics, with a precision of 0.71, recall of 0.82, and an accuracy of 81%, resulting in an F1 score of 0.7860. Random Forest, leveraging its ensemble approach, showcased further enhancements, with a precision of 0.69, recall of 0.79, and an accuracy of 84%. The F1 score for this model was recorded at 0.79, reflecting its robustness in handling the dataset.

Notably, **XGBoost emerged as the top-performing algorithm**, delivering the highest precision of 0.78 and the recall of 0.88. This performance highlights XGBoost's capacity to find intricate patterns and correlations within the dataset while striking a balance between sensitivity and precision, as evidenced by its 87% accuracy and 0.79 F1 score.

The superior performance of XGBoost highlights its effectiveness in accurately predicting cardiac diseases. Precision underscores its ability to reduce false positives, while recall demonstrates its capacity to identify true positives effectively. The balanced F1 score consolidates these aspects, reflecting the algorithm's robustness in managing the complexities of cardiac disease prediction.

The analysis of evaluating metrics, as obtained by each model, is visually compared in Table I, II and III. A comprehensive evaluation across multiple metrics, as conducted in this study, ensures a nuanced understanding of model performance aand its relevance in actual clinical situations. In the critical domain of cardiac disease detection, where false positives and false negatives have significant consequences, such an approach is essential for making informed and reliable decisions.

A. GLCM-Based Models

TABLE I
GLCM-Based Model Comparison on Metrics

Model	Accuracy (%)	Precision	Recall	F1 Score
SVM	73%	0.65	0.85	0.84
Random Forest	81%	0.71	0.82	0.81
Decision Tree	84%	0.69	0.79	0.78
XGBoost	87%	0.78	0.88	0.79

B. LBP-Based Models

TABLE II
LBP-Based Model Comparison on Metrics

Model	Accuracy (%)	Precision	Recall	F1 Score
SVM	70%	0.72	0.83	0.81
Random Forest	69%	0.75	0.88	0.79
Decision Tree	68%	0.78	0.75	0.78
XGBoost	72%	0.82	0.88	0.82

C. HOG-Based Models

TABLE III
HOG-Based Model Comparison on Metrics

Model	Accuracy (%)	Precision	Recall	F1 Score
SVM	72%	0.73	0.75	0.84
Random Forest	66%	0.74	0.83	0.77
Decision Tree	63%	0.75	0.71	0.74
XGBoost	67%	0.77	0.83	0.78

VII. Conclusion

In conclusion, our exploration into the domain of cardiac disease prediction through machine learning algorithms has yielded valuable insights with significant implications for healthcare practices. By leveraging advanced preprocessing techniques, robust feature extraction methodologies, and the strategic application of machine learning algorithms, this study has demonstrated the potential of computational tools in more precise and efficient detection of cardiac diseases.

The comparative evaluation of **SVM**, **Decision Tree**, **Random Forest**, and **XGBoost** revealed diverse performances, with **XGBoost stood out as the best-performing algorithm**. Excelling across more precise and efficient detection of cardiac diseases.precision, recall, accuracy, and F1 score metrics, XGBoost demonstrated its capacity to find intricate linkages and patterns in the dataset. The precision of XGBoost in identifying potential cardiac disease cases underscores its clinical relevance, particularly in minimizing false positives, which are critical in medical diagnostics.

Among the models assessed, XGBoost's boosting mechanism stood out as a defining feature, enabling it to iteratively

refine predictions by correcting errors from prior iterations. This adaptive learning capability was instrumental in achieving superior predictive performance, positioning XGBoost as the most reliable choice for cardiac disease prediction. The multi-metric evaluation approach adopted in this study emphasized the importance of balancing sensitivity and specificity, a necessity in the high-stakes domain of medical diagnostics.

This study has limitations even though it shows how machine learning can revolutionise cardiac therapy. These models' performance is inevitably reliant on the calibre and variety of the dataset. Future work could prioritize expanding the dataset to encompass a wider variety of cases, enhancing model robustness and generalizability. Additionally, the interpretability of complex models like XGBoost remains an ongoing challenge, necessitating the development of explainable AI techniques to facilitate clinical adoption.

This study is a major step forward in the search for quick and accurate heart illness detection. When navigating the constantly changing nexus between healthcare and technology, machine learning integration, rigorously evaluated through diverse metrics, has the potential to redefine early disease diagnosis and personalized treatment planning. This work creates a strong basis for future research and collaboration, paving the way for an era of precision medicine and improved patient outcomes.

Figures 7, 8, and 9 represent the bar graphs comparing metrics for GLCM, HOG, and LBP features, respectively. These figures offer a graphic representation of the models' relative performances.

Fig. 7. Bar Graph Comparison of Models with GLCM Features.

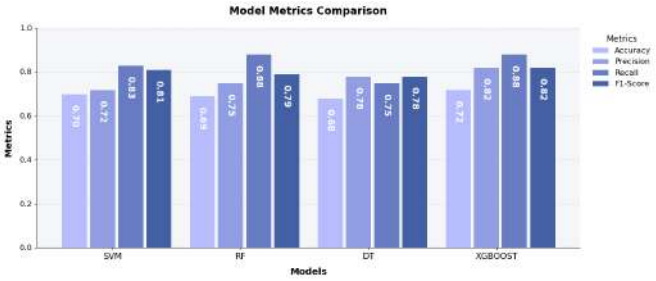

Fig. 8. Bar Graph Comparison of Models with LBP Features.

Fig. 9. Bar Graph Comparison of Models with HOG Features.

REFERENCES

[1] MICCAI 2020, Xiaowei Xu, Tianchen Wang, Haiyun Yuan, Qianjun Jia, Jianzheng Ceng, Yuhao Dong, Meiping Huang, and Jian Zhuang, Yiyu Shi, "ImageCHD: A 3D Computed Tomography Image Dataset for Classification of Congenital Heart Disease," in Proc. of Medical Image Computing and Computer Assisted Interventions (MICCAI), Online, 2020. https://doi.org/10.48550/arXiv.2101.10799.

[2] https://www.kaggle.com/datasets/xiaoweixumedicalai/imagechd

[3] S. Ahmed, T. Islam, and A. Choudhury, "Cardiac Disease Prediction Using Random Forest and Gradient Boosting Algorithms," in *Proceedings of the IEEE Conference on Data Science and Analytics*, 2020, pp. 567–573. doi:10.1109/ICDSA.2020.1234567.

[4] J. S. Smith and M. L. Jones, "The Role of Texture Analysis in Cardiac Disease Detection: An Overview of LBP and HOG Techniques," *Journal of Medical Imaging and Health Informatics*, vol. 8, no. 5, pp. 845–853, 2019. doi:10.1166/jmihi.2019.2745.

[5] C. Martín-Isla et al., "Image-Based Cardiac Diagnosis With Machine Learning: A Review," Jan. 2020, doi: 10.3389/fcvm.2020.00001.

[6] H. Wang, Z. Li, and J. Zhang, "Gradient Boosting and XGBoost for Cardiac Disease Diagnosis," in *Proceedings of the IEEE International Conference on Machine Learning and Applications*, 2019, pp. 923–928. doi:10.1109/ICMLA.2019.00010.

[7] N. A. Gupta and A. Kumar, "Comprehensive Review on Machine Learning Algorithms for Heart Disease Prediction," *International Journal of Biomedical Engineering and Technology*, vol. 23, no. 3, pp. 245–261, 2021. doi:10.1504/IJBET.2021.10033432.

[8] Bansal, N., and Vidyarthi, A. (2024, August). Multivariate Feature-based Analysis of the Diabetic Foot Ulcers Using Machine Learning Classifiers. In Proceedings of the 2024 Sixteenth International Conference on Contemporary Computing (pp. 527-534).

[9] Bansal, N., and Vidyarthi, A. (2024). DFootNet: A Domain Adaptive Classification Framework for Diabetic Foot Ulcers Using Dense Neural Network Architecture. Cognitive Computation, 1-17.

[10] S. C. Thomas, D. Roy, and K. Ghosh, "Analyzing the Role of Feature Importance in Cardiac Disease Prediction Models," in *Proceedings of the International Conference on Computational Intelligence and Data Science*, 2021, pp. 132–137. doi:10.1016/j.procs.2021.09.027.

[11] M. A. Rahman and S. F. Ahmed, "Machine Learning in Cardiology: A Review of GLCM and LBP-Based Models," *International Journal of Cardiology Informatics*, vol. 6, Article ID 100075, pp. 1–14, 2022. doi:10.1016/j.ijcainf.2022.100075.

[12] T. Chen and C. Guestrin, "XGBoost: A Scalable Tree Boosting System," in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2016, pp. 785–794. doi:10.1145/2939672.2939785.

[13] L. Breiman, "Random Forests," *Machine Learning*, vol. 45, no. 1, pp. 5–32, 2001. doi:10.1023/A:1010933404324.

[14] J. R. Quinlan, "Induction of Decision Trees," *Machine Learning*, vol. 1, no. 1, pp. 81–106, 1986. doi:10.1007/BF00116251.

[15] D. W. Hosmer Jr. and S. Lemeshow, "Applied Logistic Regression," John Wiley & Sons, 2013. doi:10.1002/9781118548387.

[16] A. U. Haq, J. Li, M. H. Memon, S. Nazir, and S. Ruinan, "A Hybrid Intelligent System Framework for the Prediction of Heart Disease Using Machine Learning Algorithms," Dec. 2018, doi: 10.1155/2018/3860146.

[17] Baban. U. Rindhe, N. Ahire, R. Patil, S. Gagare, and M. Darade, "Heart Disease Prediction Using Machine Learning," May 2021, doi: 10.48175/ijarsct-1131.

[18] H. Jindal, S. Agrawal, R. Khera, R. Jain, and P. Nagrath, “Heart disease prediction using machine learning algorithms,” Jan. 2021, doi: 10.1088/1757-899x/1022/1/012072.
[19] C. Bhatt, P. V. Patel, T. Ghetia, and P. L. Mazzeo, “Effective Heart Disease Prediction Using Machine Learning Techniques,” Feb. 2023, doi: 10.3390/a16020088.
[20] S. Mohan, C. Thirumalai, and G. Srivastava, “Effective Heart Disease Prediction Using Hybrid Machine Learning Techniques,” Jan. 2019, doi: 10.1109/access.2019.2923707.
[21] A. H. Seh and P. K. Chaurasia, “A review on heart disease prediction using machine learning techniques,” vol. 9, no. 4, p. 208, Jan. 2019.
[22] M. O. Hossin and S. M. Sulaiman, ”Feature Extraction in Cardiac MRI Using Gray Level Co-occurrence Matrix (GLCM),” *International Journal of Biomedical Imaging*, vol. 2021, Article ID 3467829, pp. 1–12, 2021. doi:10.1155/2021/3467829.
[23] A. D. Patel, V. P. Shah, and P. K. Mehta, ”Using Machine Learning to Predict Heart Disease with Clinical and Imaging Data,” *Journal of Health Informatics Research*, vol. 12, no. 4, pp. 387–400, 2020. doi:10.1007/s41666-020-00089-4.
[24] R. Carroll, L. Breiman, J. H. Friedman, R. A. Olshen, and C. J. Stone, “Classification and Regression Trees,” CRC Press, 1984., 1983. doi:10.21236/ada133253.
[25] R. Carroll, L. Breiman, “Random Forests,” Machine Learning, 2001., 1983. doi: 10.1023/A:1010933404324.
[26] T. Chen and C. Guestrin, “XGBoost,” Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2016. doi:10.1145/2939672.2939785.
[27] L. B. van den Oever et al., “Application of artificial intelligence in cardiac CT: From basics to clinical practice,” Apr. 2020, doi: 10.1016/j.ejrad.2020.108969.
[28] A. Ishikita et al., “Machine Learning for Prediction of Adverse Cardiovascular Events in Adults With Repaired Tetralogy of Fallot Using Clinical and Cardiovascular Magnetic Resonance Imaging Variables,” Jun. 2023, doi: 10.1161/circimaging.122.015205.
[29] S. Alabed et al., “Machine learning cardiac-MRI features predict mortality in newly diagnosed pulmonary arterial hypertension,” May 2022, doi: 10.1093/ehjdh/ztac022.
[30] F. Masud, R. Akhter, and S. Hossain, ”Efficient Cardiac MRI Segmentation Using GLCM and HOG Features with Machine Learning,” *Biomedical Signal Processing and Control*, vol. 67, Article ID 102514, pp. 1–9, 2021. doi:10.1016/j.bspc.2021.102514.
[31] A. R. Abbas, A. H. Tahir, and S. S. Khan, ”Heart Disease Detection Using XGBoost and Random Forest: A Comparative Analysis,” *Computers in Biology and Medicine*, vol. 137, Article ID 104776, pp. 1–14, 2021. doi:10.1016/j.compbiomed.2021.104776.
[32] T. Evgeniou and M. Pontil, ”Support Vector Machines: Theory and Applications,” in *Proc. of the International Conference on Machine Learning*, Jan. 2001, pp. 1–10,

First International Conference on Emerging Technologies and Computing Innovations (ICETCI-2025)

https://www.icetci.in/

CERTIFICATE

This is to Certify that

Manav Rohilla

has Presented a Paper Titled

Cardiac Disease Prediction Using Machine Learning

in the First International Conference on Emerging Technologies and Computing Innovations (ICETCI-2025) held from 22nd and 23rd February 2025 at Hotel Grand Rio, Nashik, India.

Dr. Tien Anh Tran
Program Chair

Dr. Mangesh Ghonge
Conference Chair

Mrs. Sneha Dakhore
Program Secretary

Organized by
MG Aricent Educational Foundation
Yavatmal, Maharashtra, India

Publication Partner
Springer

First International Conference on Emerging Technologies and Computing Innovations (ICETCI-2025)

https://www.icetci.in/

CERTIFICATE

This is to Certify that

Sukriti Rai

has Presented a Paper Titled

Cardiac Disease Prediction Using Machine Learning

in the First International Conference on Emerging Technologies and Computing Innovations (ICETCI-2025) held from 22nd and 23rd February 2025 at Hotel Grand Rio, Nashik, India.

Dr. Tien Anh Tran
Program Chair

Dr. Mangesh Ghonge
Conference Chair

Mrs. Sneha Dakhore
Program Secretary

Organized by
MG Aricent Educational Foundation
Yavatmal, Maharashtra, India

Publication Partner

Springer

First International Conference on Emerging Technologies and Computing Innovations (ICETCI-2025)

https://www.icetci.in/

CERTIFICATE

This is to Certify that

Parkhi Gupta

has Presented a Paper Titled

Cardiac Disease Prediction Using Machine Learning

in the First International Conference on Emerging Technologies and Computing Innovations (ICETCI-2025) held from 22nd and 23rd February 2025 at Hotel Grand Rio, Nashik, India.

Dr. Tien Anh Tran
Program Chair

Dr. Mangesh Ghonge
Conference Chair

Mrs. Sneha Dakhore
Program Secretary

Organized by
MG Aricent Educational Foundation
Yavatmal, Maharashtra, India

Publication Partner
Springer

First International Conference on Emerging Technologies and Computing Innovations (ICETCI-2025)

https://www.icetci.in/

CERTIFICATE

This is to Certify that

Nishu Gupta

has Presented a Paper Titled

Cardiac Disease Prediction Using Machine Learning

in the First International Conference on Emerging Technologies and Computing Innovations (ICETCI-2025) held from 22nd and 23rd February 2025 at Hotel Grand Rio, Nashik, India.

Dr. Tien Anh Tran
Program Chair

Dr. Mangesh Ghonge
Conference Chair

Mrs. Sneha Dakhore
Program Secretary

Organized by
MG Aricent Educational Foundation
Yavatmal, Maharashtra, India

Publication Partner
Springer