Heart Disease Prediction Using ECG Signal Analysis and Machine Learning

Mrs. M. Mounika[1], Grandhi Ranga Madhupriya[2], A. Vivekanand[3], Y. Swarageeth[4], G. Shashank[5]
[1]*Assistant Professor, Dept. of Computer Science & Engineering Malla Reddy University, Hyderabad*
[2,3,4,5]*Department of Computer Science & Engineering Malla Reddy University, Hyderabad*
doi.org/10.64643/IJIRTV12I12-206826-459

***Abstract*—The heart disease detection system is a web-based application that seeks to analyze the images of an electrocardiogram (ECG) and detect potential heart conditions with the help of machine learning algorithms. It is a hybrid model that combines both image processing and supervised learning in order to convert ECG images into numerical signals to be analyzed. Initial steps that are followed by the system after the uploaded ECG image is processing involve: transformation to grayscale, Gaussian filtering followed by thresholding having used Otsu thresholding that makes the images clear. The ECG has been preprocessed then divided into 12 standard leads as well as a longer lead of one. Waveform patterns are then determined by using contour detection strategies and translated into one dimensional numerical signal which represent heart activity. Min-Max scaling is then used to normalize these signals and Principal Component Analysis (PCA) is used to trim off the features and this helps in enhancing efficiency and model performance. The refined information is inputted in a trained machine learning model to classify. The algorithms that were tested during development are logistic regression, random forest, Naive Bayes, and K-nearest neighbors. The last model estimates four conditions, which include normal functioning of the heart, arrhythmia, infarction of the myocardium, and the past history concerning attacks. This system was written in Python by using NumPy, Pandas, Scikit-learn and Scikit-image libraries. The interface was created on the basis of the Streamlit framework, thus allowing users to input ECG pictures and get immediate estimates. This project proves to be a good method of early heart disease detection through image processing methods with machine learning models.**

***Index Terms*—ECG Classification, Machine Learning, Predictive Modeling, Grayscale Conversion, K-Nearest Neighbors (KNN), Image Processing**.

I. INTRODUCTION

Cardiovascular diseases remain among the major causes of mortality globally, and the promptness of the disease diagnosis and timely, precise diagnosis is of paramount importance towards the successful treatment and prevention of the disease. Electrocardiogram (ECG) signals are most important as they give crucial information about the electrical activity of the heart and are important in diagnosis of any cardiac abnormalities. Nevertheless, the process of medical professionals manually interpreting ECGs reports can be extensive, subjective and prone to error especially when working with a large number of patients at the same time.

Over the last few years, machine learning and image processing have developed tremendously, which has provided the prospects of change in the healthcare sector especially in automated medical diagnosis. With the fusion of these two disciplines, intelligent systems to study ECG data efficiently and identify patterns that are related to various heart conditions are now possible. By means of machine learning, useful features of ECG signals can be obtained and applied to classify and predict the presence of various cardiovascular disorders with near accuracy.

In this project, a method of detecting heart diseases is developed, which uses conventional ECG images as an input. The method uses a sequence of picture processing methods to draw valuable numerical data out of the visual waveforms. It begins with very fundamental preprocessing stages like noise elimination and image refinement. It is followed by the division of ECG image into separate leads whence the individual signals are derived. Figure 1: The extracted data is further elevated using feature processing and

Principal Finally, the processed features will be inputted into a supervised machine learning model that has been trained to label against various heart diseases such as normal ECG, arrhythmia, and myocardial infarction. The whole system is executed as a web-based application based on the Streamlit framework, which provides the end-user with the simplicity of the interface and the possibility to get predictions instant. This publication shows how imaging processing and machine learning approaches can be combined to provide an opportunity to detect heart diseases in time and with high accuracy, which would lead to better clinical outcomes eventually.

II. LITERATURE REVIEW

In the recent years, the detection of heart-related illnesses using machine learning algorithms based on ECG signals has been a topic of significant concern due to the necessity to support this process by the clear need to gain faster and more efficient diagnostics. Initial studies indicated that multi-model combining or ensemble methods could produce a significant improvement on prediction accuracy of medical data. The subsequent research found that Random Forest performs well, particularly in the situations of complex and high-dimensional data, such as ECG data, which has more insidious patterns and can be detected using alternative statistical methods.

Advancements in the field were made through deep learning architectures, especially neural networks of convolutional type (CNNs), used directly on raw ECG signals. A lot of literature demonstrated the accuracy of arrhythmia classification rivaling the average cardiologist, and the development of artificial intelligence as an automated diagnostics tool is in place. A few studies have also utilized large open ECG databanks to train deep networks, and reported reliable results on a variety of cardiac anomalies.

Simultaneously, there is an increasing number of research works done on image-based methods of ECG analysis. Under such approaches, ECG signals are considered as images undergoing classical steps in computer vision along with classifiers like Support Vector Machines (SVM) and K-Nearest Neighbors (KNN). Generally, these works emphasis on several occasions that good classification results can be achieved through effective noise elimination and proper segmentation and careful extraction of features. More studies have been conducted on adaptive neural network models to provide real-time or patient-specific ECG monitoring to address the need to deliver more individualized care. Approaches to reducing the dimensionality of data and enhancing the quality of features have been valuable in the creation of efficient systems. One of the most commonly applied methods that have found application in this respect is Principal Component Analysis (PCA) that minimizes on the features enjoyed, but retains the most informative details on the data.

It has been experimentally demonstrated that PCA on the ECG characteristics is applicable in reducing the redundant information in the characteristics and minimizing training and inferences. It has been proved that integration of PCA and standard classifiers is more accurate at less cost to the calculation thus making it profitable to the real-world heart disease detecting systems.

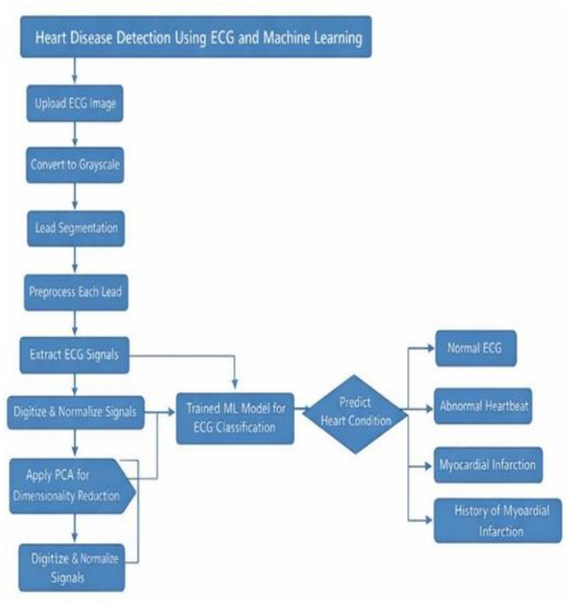
Nonetheless, it is still greatly disadvantaged. Much of published algorithms are only tested when using clean and laboratory-type data, and so become less robust on the actual data when trying with real-world ECG images. Problems of imbalance in classes, dissimilarity in image quality and lack of common norms still corrupt credibility. Deep-learning-based systems do not always need large datasets and a significant amount of powerful hardware, which, to some extent, can limit their use in smaller clinics or other environments that can be resource-intensive. In the future, their trends will be towards more generalizable models whose models can be run in real-time, and the model could still be applicable in the everyday clinical practice.

III. METHODOLOGY

Overview:

The suggested solution is focused on the recognition of the heart disease using ECG pictures with the help of image processing solution and machine learning technologies. Some steps that the general process entails have been identified, and they include: data collection, data preprocessing, signal extraction, relevant features generation, reduction of data dimensions, model training and system performance evaluation and eventually end the process as implementation of the system. In this strategy, ECG images are coded to numerical data, which are sampled by a trained learning model in order to label all the different types of heart conditions.

3.1 Data Collection



The system primarily takes ECG images as its input. The dataset used in this approach includes several components. It contains ECG images collected from medical records or diagnostic centers, including standard 12-lead ECGs and, in some cases, additional extended leads. Each image is associated with labels that indicate specific heart conditions, such as normal activity, irregular heartbeat patterns, or myocardial infarction. During preprocessing, the ECG waveforms are also converted into numerical signal data, which is later used for analysis and model training.

3.2 Data Preprocessing

There is the need to carefully process raw ECG images prior to analysis. The stages would start by synthesizing color images into grayscale, which makes it less complex. The Gaussian filtering is then done to reduce noise and Otsu thresholding is followed to reduce the image to a binary one to enhance separation of waveforms and background.

Once this has been done, the ECG is separated into its intended elements, typically incorporating the 12 standard leads along with a long lead. Lastly, there is the application of normalization so that all pictures are on the same scale which makes it easier to be more similar when further analysis is required.

3.3 Signal Extraction and Feature selection.

The procedure starts by drawing outlines in the ECG images after using the processed ECG images in order to determine the outlines of the waveform patterns. These derived waveforms are again transformed into one dimensional numeric array, and can therefore be analyzed with computation. The numerical data is then structured into feature sets of significance of the signals. Lastly, a normalization process is implemented in which normalization is performed by methods like MinMax scaling which balances the values of the features in a normal range of 0 to 1 which not only ensures consistency but also enhances the model performance.

3.4 Dimensionality Reduction

The system uses the dimensionality reduction method, Principal Component Analysis (PCA). Its primary aim is to ease the feature space by ensuring that it lessens the variables but maintains the most significant data. This method allows to remove redundant or less important features and increases the general effectiveness and performance of the model.

3.5 Model Selection

The system uses supervised machine learning capabilities to do classification. Several algorithms are employed in this, one of them being Logistic Regression being a core model, another being random forest because it is able to deal with nonlinear relationships, and all the other algorithms are Nave bayes, and K-Nearest Neighbors (KNN) where the data is classified with the help of similarity conscience. The best model is selected through the performance comparison and selection of the best model that would have the best accuracy and overall performance.

3.6 Model Training

The data is first divided into two groups one of them is used to train the model and the other is used to test the performance of the model. Significant aspects are then marked off and chosen to increase the degree of prediction. Once this has been done the selected algorithms are then trained through the processed ECG data allowing them to learn patterns and thus make sound classifications.

3.7 Model Evaluation

Within the model, the evaluation is measured with a number of measures. The extent of correct prediction made by the model on average is determined by accuracy. Precision and recall can also be used to assess its performance on individual classes. F1 score is the most balanced score which is a combination of recall and precision. Moreover, a confusion types table is employed to visually illustrate the comparison between the prediction done by the model and the true results.

3.8 Deployment and Prediction

The implementation of the system is done through the Streamlit framework to offer an interactive interface. It is a system in which the user inserts an image of the ECG and the system works on extracting the signal information. The PCA is used to reduce the extracted features and forwarded to the trained model to make predictions.

According to the analysis, the ECG is categorized in the model into various types including normal condition, abnormal heartbeat, myocardial infarction or a history of myocardial infarction history.

3.9 Tools and Technologies

The development is made in Python as the main programming language. It utilizes various libraries, such as NumPy and Pandas to process data, Scikit-learn to perform

machine learning, Scikit-image to process images, and Matplotlib to visualize and depict data. TheStreamlit framework is used to build its user interface. The project uses different file formats like Python code, saved models, data, and picture inputs in .py, .pkl, and.csv, respectively.

3.10 System Architecture

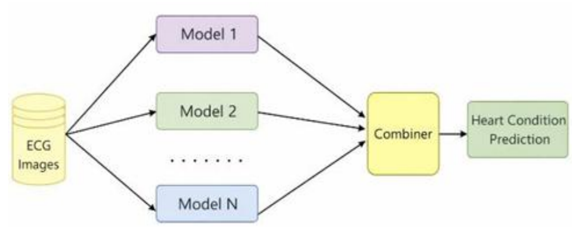
The general system design is that which is designed in such a way that it deals effectively with the ECG images and makes the appropriate predictions. It starts with the input layer which takes ECG images by the user. These images are then transferred to the processing layer where they undergo activities like preprocessing, segmentation and signal extraction.

The extracted data is then normalized and decreased in the feature layer by employing methods such as PCA in order to simplify the features. The extracted data is refined and then sent through to the model layer which is trained using a machine learning algorithm to perform the classification. Lastly, the output layer displays the forecasted state of the cardiac findings to the user.

The data flow between the input and the final output is facilitated in this structured architecture in a smooth manner whilst the processing time and efficiency are high.

3.11 Limitations and Future Enhancement.

Although the system is good, it still possesses certain limitations. Its results highly rely on the quality of images of the input ECG because low or noisy images can influence the results. The size and the diversity of the dataset, with which the model is trained, also have an impact on the effectiveness of the model. The image processing and machine learning methods also consume some amount of computing resources. The future has a number of possibilities of improving. The use of the methods of deep learning, including convolutional neural networks (CNNs), would improve accuracy further. The system would be made more practical by moving towards the use of real-time analysis of ECG signals as opposed to the use of only static images. The creation of the mobile-based application might enhance the level of access and including the system into the hospital settings would facilitate its implementation into clinical practice.



IV. DISSCUSSION

Impact of Data Preprocessing

The correct preprocess of ECG images and signals is crucial to enhancement of the overall system performance. Grayscale conversion, Gaussian filter to remove noise, and Otsu thresholding also improve the image depiction and visibility of the waveforms to a great extent, which are essential in the extraction of reliable signals. Min-Max normalization is a form of feature scaling, which normalizes all the features to a proportional range, and in the process, no feature is allowed to occupy the focus of the model demonstrating the final decision. Also, Dimensionality Reduction is undertaken using Principal Component Analysis (PCA), which is effective in ensuring that redundant data is done away with and computational complexity is minimized. Such steps do not only help in enhancing the processing time but also help in attaining a better accuracy of the prediction.

A. Metrics of Evaluation: Explanations.

The standard measures that were used to determine model performance consisted of a number of measures. Accuracy is a measure of the total percentage of accurate predictions of all test cases and will give a rough idea of the reliability of the model. The problem with accuracy is however, deceptive with imbalanced datasets.

Precision shows what percentage of positive projections made by the model are correct. It is especially crucial in medical applications where false positives are very important like missing out giving a heart disease diagnosis to a healthy patient. The ability of the model to recognize all the real positive cases is known as sensitivity which is also termed as recall. The measure is particularly critical in medicine, since the failure to identify a true case of a disease (false negative) may be clinically severe.

Aliso, to visualize how the model worked out, a confusion matrix was employed which presented the number of correct and incorrect outpourings of each class. ROC-AUC metric also tests the ability of the model to discriminate among various classes with higher values indicating greater general discriminative power.

Table: Performance Metrics of Classifiers

Classifier	Accuracy (%)	Precision	Recall	F1-Score
Logistic Regression	85	0.84	0.83	0.83
Random Forest	92	0.91	0.92	0.91
Naïve Bayes	82	0.80	0.81	0.80
KNN	88	0.87	0.88	0.87

B. Detailed Performance Analysis

The machine learning models assessed in this paper exhibited varying degrees of their performance in the heart disease classification task involving ECG. The overall results on the tested algorithms saw the highest results on random Forest. It can be considered effective to capture complex and non-line relationships inherent in ECG data with its ensemble structure that is constructed using several decision trees to minimize the likelihood of overfitting.

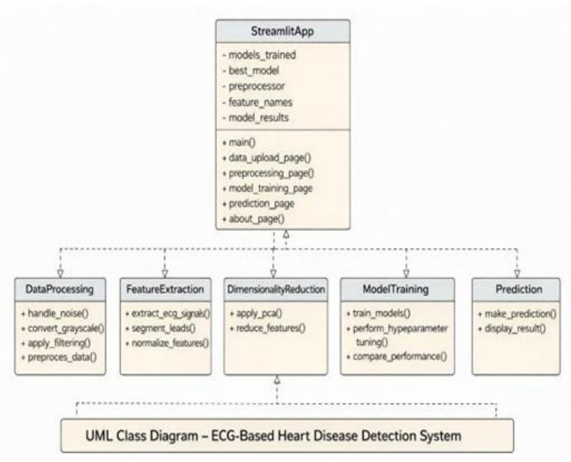
This makes it especially appropriate when dealing with high-dimensional medical data. K-Nearest Neighbors (KNN) was also competitive, particularly when it was followed by Min-Max normalization. Since KNN is a distance-based method, appropriate feature scaling contributed to its accuracy to a considerable extent. The performance of the model is however sensitive to the parameter k used as well as the distance measure. Naive Jones provided the ability of computation and quick prediction. However, its high performance on the independence of features diminished its performance in its ability to deal with correlated features, such as those in multi-lead ECG signals, making its accuracy relatively lower. The basic model was that of Logistic Regression which gave consistent and interpretable results. Although it is efficient with regard to linearly separable data, it is not as good at modeling the complicated nonlinear trends that ECG waveforms manifest. Generally, the best choice was found to be the most effective in terms of accuracy, resistance to overfitting, and ability to generalize effectively, making the Random Forest a significant choice. The comparative analysis of all the four models offered useful information recourse to the final choice of the most appropriate method of this heart disease prediction task.

V. UML CLASS DIAGRAMS

A. Class Diagram

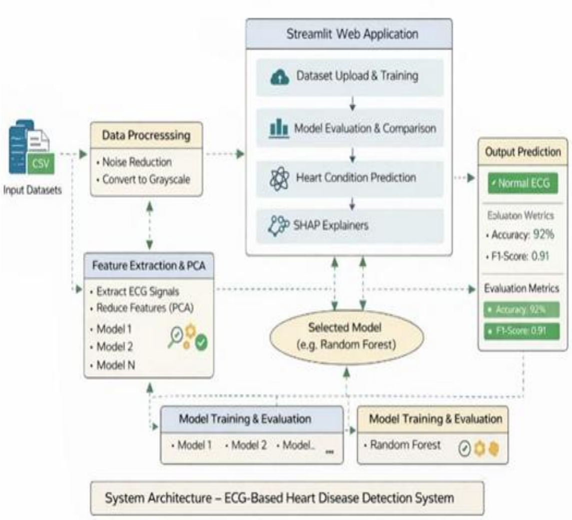
UML class diagram is the general architecture of the ECG-based system of detection of heart diseases. The main part is the Streamlit App object, which manages the user interface, along with the application flow.

It stores such crucial information as trained models, preprocessing information and prediction outcomes and provides such functions as main (), data_upload_page, preprocessing page as well as prediction page to control the various stages of operation within the system.



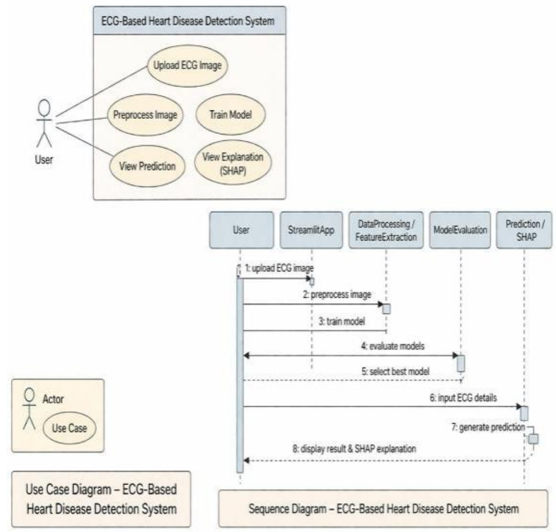
The task of constructing and training the machine learning models lies in the Model Training class, whereas the performance of the more specific models can be evaluated by the Model Evaluation class based on such metrics as accuracy and F1 score. At the last step, the Prediction class is applied to create the results of the classification of heart diseases and to present them.

B. System Architecture



The system architecture diagram shows the entire workflow of the ECG-based system of detecting heart diseases, beginning with the input and concluding with the ultimate forecasting. The procedure starts when the user feeds a picture of the ECG, is an contains valuable data about the actions of the heart, through the preprocessing stage, the image is fed through filters like noise removal, grayscale conversion, and filtering are processes done to improve the picture. Then feature extraction techniques are used to extract some meaningful information in the ECG data. It is also worth noting that the dimensionality reduction techniques are employed to structure the information into a simplified and structured format which can be utilized in machine learning model.

C. Use Case Diagram & Sequence Diagram



The use case diagram is a reflection of the interaction between the user and the system of heart disease detection. The web application is primarily utilized by the patient who carries out various operations through the application. The process starts with the uploading of an ECG image which is subsequently uploaded to undergo analysis which includes preprocessing processes to minimise noise and improve the quality of the image. The user can then go on to execute the model where machine learning techniques analyse the ECG data and identify significant features. After the analysis is done the results are presented in the system as either the ECG is a normal result or there are indications of heart-related abnormality.

The sequence diagram details the interactions between various components of the system in appropriate succession of the prediction process. It begins when the user sends an ECG image by using Streamlit. Then this image is fed into the preprocessing and feature extraction module whereby the data is cleaned and then the ECG signals are extracted.

The refined data is then sent to the module that performs the task of training and evaluating the model after preprocessing whereby various models are compared and the best model selected. The chosen model then uses the ECG data as inputs to come up with a prediction. Lastly, the system shows the user the identified cardiac condition.

This flow chart shows a clear way how the data flows through the system and outlines the sequence in which each operation was used in the system.

VI. RESULTS

A. Uploaded ECG Image

The process begins when the system accepts an ECG report

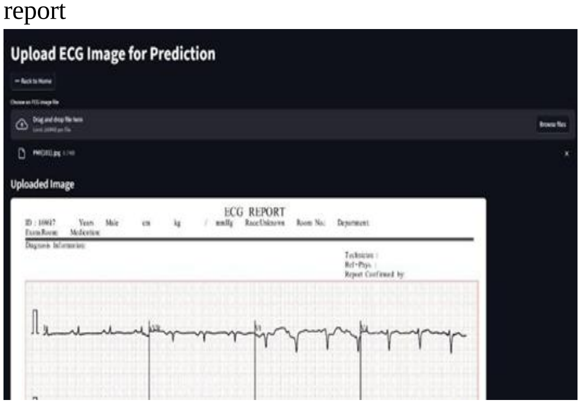


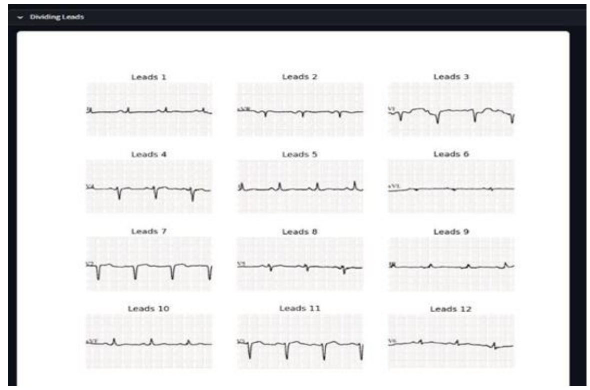
image through the web interface. After uploading, the image is shown on the screen to confirm that it has been received correctly. This step helps verify that a proper multi-lead ECG report is available for further analysis. The clarity and resolution of the image play a key role in determining the accuracy of the following stages.

B. Gray Scale Image

Once the ECG image is uploaded, it is transformed into a

grayscale format. This step simplifies the image by removing color details while retaining the essential waveform patterns. Converting to grayscale makes further processing more efficient and prepares the image for steps like noise removal and segmentation. It plays a key role in ensuring accurate extraction of ECG signals.

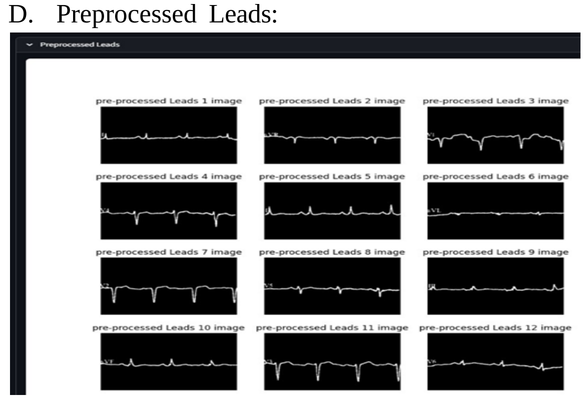
C. Divide Leads (Lead Segmentation)



The ECG image is separated into 12 distinct leads, with each lead capturing the heart's electrical activity from a different viewpoint. By isolating these leads, the system is able to examine each signal individually, which helps improve the accuracy of feature extraction.

This step is important because certain cardiac conditions may appear only in specific leads, making detailed analysis essential.

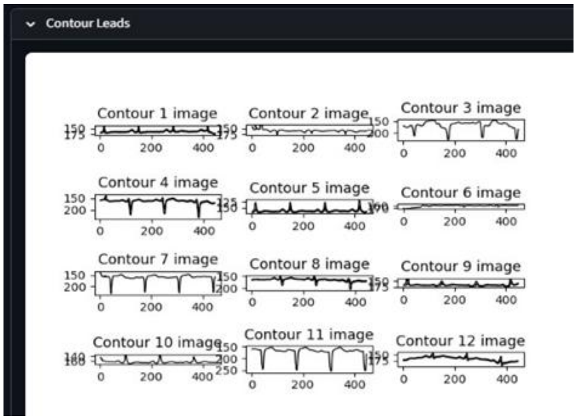
D. Preprocessed Leads:



Each extracted lead undergoes further processing steps, including noise removal, thresholding, and edge detection. These operations help produce cleaner signals with well-defined waveforms against a dark background.

By reducing unwanted disturbances and enhancing key patterns such as P waves, QRS complexes, and T waves, this stage improves the clarity of the signals and supports more accurate classification.

E. Signal Extraction (Contour Detection)



The system retrieves the ECG waveform from the processed images using contour detection techniques. The extracted waveform is then transformed into numerical data that represents changes in amplitude over time. This conversion enables the machine learning model to work with the ECG information in a numerical format rather than as an image.

a. 1D Signal Conversion

1D Signals

	0	1	2	3	4	5	6	7	8	9	10	11	12
0	0.9993	1	0.9434	0.9069	0.8297	0.7932	0.809	0.7626	0.6496	0.5292	0.5487	0.6134	0.6359

The signals obtained from each lead are transformed into one-dimensional arrays and saved in a CSV file. This format organizes the ECG waveform data in a clear and structured numerical form. The data from all 12 leads are then merged into a single dataset, which is used as the input for the prediction model.

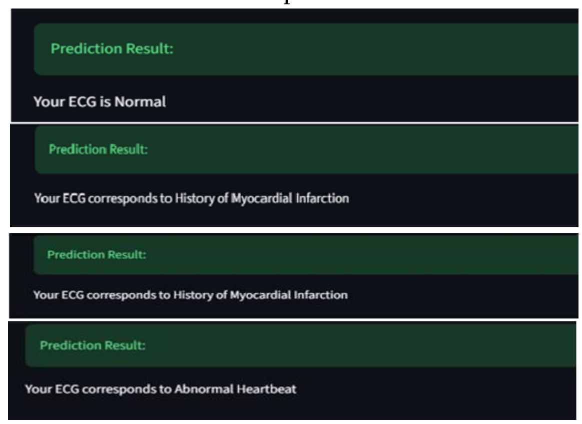
F. Feature Scaling and Dimensionality Reduction

Dimensional Reduction

	0	1	2	3	4	5	6	7	8	9	10
0	3.1476	7.0697	-1.4893	-2.8059	-1.5944	0.2949	-0.4786	0.7191	-0.7898	2.3014	-2.8122

The combined ECG signal data is first scaled using a min–max normalization method so that all values lie within a consistent range. After this, Principal Component Analysis (PCA) is applied to reduce the number of features while preserving the most important information. This process enhances the efficiency of the model, speeds up computation, and helps minimize the risk of overfitting.

G. Final Prediction Output



The prepared data is then provided to the trained machine learning model for analysis, allowing it to determine the condition of the heart. Based on the evaluation, the system categorizes the ECG into one of several classes, including normal condition, irregular heartbeat, myocardial infarction (MI), or a previous history of myocardial infarction.

VII. CONCLUSION

The article has succeeded to show that the hybridization of conventional processing on an image of a patient and the machine learning with supervision is an effective tool of identifying heart disease with the pictures of an ECG report only. The achieved scheme processes input images in a predetermined sequence of levels: picture upgrading, lead directed separation, strain amplification recognizancy, signal remodeling and feature elimination by PCA to generate a reliable identification between normal cardiac rhythm, arrhythmia, myocardial infarction, and instances of pre-infarction.

When the tracks of visually recorded ECG are digitized to digital time-series data, the model provides more predictable and accurate results compared to the manual process. The entire automated pipeline may be run within a short time and on the one hand; it is useful especially when emergency health care evaluations are to be conducted. Furthermore, the application is the convenient auxiliary solution, which could help the doctors and clinicians to discover the potential problematic issues faster and, more to the point, more confidently.

It is easy to program this solution since it is coded as a simple web app with Streamlit and then start airing an image of an ECG and instantly receive imminent feedback. In conclusion, the given solution is a powerful, fast, and convenient substitute to contribute to the screening of heart disease in its early stages. The question of how deep learning architecture can be integrated and how bigger and more varied collections of actual clinical ECG images can be employed to increase generalization and real-world performance will be researched in the future.

REFERENCES

[1] Islam et al. (2015) presented a detailed survey on the role of the Internet of Things (IoT) in healthcare, discussing its applications, benefits, and challenges in medical systems.

[2] Piepoli and colleagues (2016) provided European guidelines focused on the prevention of cardiovascular diseases, offering recommendations for clinical practice.

[3] Nahum-Shani et al. (2016) introduced the concept of Just-in-Time Adaptive Interventions (JITAIs), highlighting their design principles for supporting health behavior through mobile technologies.

[4] Mohan, Thirumalai, and Srivastava (2019) proposed a hybrid machine learning approach aimed at improving heart disease prediction accuracy.

[5] Kelly et al. (2019) discussed the major challenges involved in applying artificial intelligence effectively in clinical environments.

[6] Yu and co-authors (2020) developed an AI-based system for stroke prediction using real-time electromyography (EMG) signals.

[7] Wagner et al. (2020) introduced PTB-XL, a large-scale and publicly accessible ECG dataset widely used for research purposes.

[8] Abdou, Ngom, and Niang (2020) explored arrhythmia prediction using a hybrid model that combines convolutional neural networks with nonlinear regression techniques.

[9] Saini and Gupta (2021) reviewed various artificial intelligence methods for analyzing ECG signals, along with current challenges and future directions.

[10] Khan and Kim (2021) applied a Long Short-Term Memory (LSTM) deep learning model for classifying cardiac arrhythmias.

[11] Alzubaidi et al. (2021) provided a comprehensive review of deep learning concepts, architectures such as CNNs, and their applications and challenges.

[12] Mastoi and collaborators (2021) proposed an integrated machine learning and data mining approach for predicting premature ventricular contractions.

[13] KS (2021) presented a machine learning-based system for heart disease prediction.

[14] Patil et al. (2022) developed a heart disease prediction model using machine learning techniques to improve diagnostic accuracy.

[15] Kumar and Ramachandran (2022) reviewed compressive sensing methods for ECG signal processing within IoT-based deep learning frameworks.

[16] Ardeti et al. (2023) provided an overview of modern ECG signal processing methods, ranging from traditional approaches to AI-based techniques.

[17] Ashtaiwi and colleagues (2024) focused on improving heart disease diagnosis through classification methods based on ECG image vectorization.

[18] Anand et al. (2025) proposed a fast detection approach for heart disease using ECG signals and chest X-ray images, incorporating advanced transformation techniques.