IJCRT.ORG ISSN : 2320-2882

INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS (IJCRT)

An International Open Access, Peer-reviewed, Refereed Journal

Cardiovascular Disease Prediction System Using Ensemble Algorithm

Manas Chagi, Abhishek G, Theju T S, Bhargava K R

B.E Students, Dept. of ISE, BIT, Bengaluru, Karnataka, India

Dr. Hema Jagadish

Associate Professor, Dept. of ISE, BIT, Bengaluru Karnataka, India

Abstract: —Cardiovascular disease (CVD) remains one of the leading causes of mortality globally, highlighting the need for early and accurate detection systems. This project presents a comprehensive approach for predicting cardiovascular disease using ensemble machine learning techniques. The system comprises an image processing module that pre processes ECG images and a machine learning module to analyze the extracted features. The image processing phase involves converting ECG images to grayscale, applying Gaussian smoothing to reduce noise, and contour detection for waveform extraction. The next phase integrates Principal Component Analysis (PCA) for dimensionality reduction, aimed at enhancing the model's efficiency by minimizing redundant data while retaining crucial features. A Voting Classifier ensemble method is then employed, combining multiple machine learning classifiers including Support Vector Machine (SVM), k Nearest Neighbors (kNN), Random Forest, Gaussian Naive Bayes, and Logistic Regression. The use of a soft voting strategy aggregates predictions based on probability scores, improving robustness and prediction accuracy. This modular architecture leverages the power of ensemble learning, utilizing ECG data to make reliable predictions. Initial results demonstrate the effectiveness of the proposed model in identifying patterns associated with cardiovascular disease. The final system is implemented using the scikit-learn and Streamlit libraries, providing an intuitive front-end for medical professionals to upload ECG images and view predictive results. This work underscores the potential of combining image processing and ensemble learning for non-invasive diagnosis of heart conditions**.**

Keywords—Cardiovascular disease prediction, ECG signal processing, Principal Component Analysis, Ensemble Learning, Voting Classifier, Machine Learning, Image Processing.

I. Introduction

Cardiovascular disease (CVD) is a broad term for conditions affecting the heart and blood vessels, and it remains one of the most common causes of death globally. Early diagnosis of CVD is critical for reducing the risks of severe complications, such as heart attacks and strokes. Traditional diagnostic methods rely heavily on clinical expertise and manual analysis of ECG signals, which can be time-consuming and prone to subjective errors. As a result, there is a growing need for automated solutions that can analyze ECG data with high accuracy and efficiency.

The Cardiovascular Disease Prediction Using Ensemble project aims to address these challenges by developing a robust system for the prediction of cardiovascular diseases based on ECG images. The system combines image processing techniques and machine learning algorithms to preprocess ECG data and predict the likelihood of CVD. By utilizing an ensemble technique, which combines multiple models to make predictions, the approach aims to improve prediction accuracy and minimize errors that may arise from using a single model. The process begins by converting ECG images into a format suitable for analysis, followed by preprocessing steps that enhance the quality of the data. The system then extracts relevant features from the ECG signals using techniques such as contouring and signal scaling. These features are reduced to lower dimensions using Principal Component Analysis (PCA), and finally, a trained ensemble model is used to predict the presence of cardiovascular diseases. The ensemble approach ensures that the system can leverage the strengths of multiple models, leading to more reliable and accurate predictions. This approach not only enhances diagnostic precision but also provides a fast, scalable, and automated method for cardiovascular disease prediction, which can be invaluable in clinical settings.

II. LITERATURE REVIEW

M.J. McAuliffe et al proposed the MIPAV (Medical Image Processing, Analysis, and Visualization) program, a platform-independent, general-purpose tool for clinical and quantitative analysis of medical images. The system facilitates 3D visualization and quantitative analysis of diverse image types such as confocal microscopy, MRI, CT, and PET scans on standard desktop computers, eliminating the need for expensive UNIX workstations. Strengths include enabling remote collaboration, enhancing data sharing and analysis, and providing an affordable solution for studying, diagnosing, monitoring, and treating medical disorders [1].

Ingrid Scholl et al highlighted the challenges of medical image processing due to the increasing resolution and volume of digital imaging data. The paper addresses Kilo- to Terabyte challenges in medical image management, bioimaging, virtual reality in medical visualization, and neuroimaging. Scalable algorithms and advanced parallelization techniques using graphical processing units (GPUs) have been developed to handle large data volumes. Strengths include the ability to adapt algorithms to manage big data efficiently and prepare for emerging Petabyte-level challenges, ensuring medical image processing remains a crucial field of research [2].

Felix Ritter et al emphasized the critical role of medical image processing in leveraging data from imaging modalities like CT, MRT, PET, and ultrasound. The field has evolved significantly since the discovery of X-ray radiation, offering unprecedented opportunities for diagnosis, therapy planning, and therapy assessment. Strengths include enhancing the extraction and presentation of relevant information from medical images, facilitating more effective and taskspecific applications in clinical practice [3].

Sigurd Angenent et al explored the central mathematical problems in medical imaging, focusing on advancements driven by geometric partial differential equations, signal/image processing, and computer graphics. These methods provide a rigorous mathematical foundation for developing software integrated into therapy delivery systems. Strengths include enhanced image processing techniques for tasks like image enhancement, registration, and segmentation, enabling more effective image-guided procedures such as radiation therapy, biopsy, and minimally invasive surgery [4].

Dibyadeep Nandi et al presented a study on the applications of Principal Component Analysis (PCA) in medical image processing. PCA transforms correlated variables into fewer principal components, enabling dimensionality reduction while retaining significant data characteristics with minimal information loss. The paper highlights PCA's efficiency in various medical image applications, including feature extraction, image fusion, compression, segmentation, registration, and denoising. Strengths include its ability to reduce data dimensions effectively, enhance computational efficiency, and prove its utility across diverse medical imaging tasks [5].

Tahmida Tabassum and Mohiuddin Ahmad proposed a method for converting printed ECG data into digital signals using image processing techniques. The system digitizes scanned ECG strips from the MIT- BIH database and real patient records, enabling effective reproduction of ECG images with high accuracy and reduced data size. Strengths include high accuracy in digitization, applicability in diagnosing cardiac conditions, and enabling further processing for automatic analysis and early detection of heart abnormalities. This research highlights the importance of digitizing ECG records for enhanced storage, analysis, and clinical application [6].

Ansumana F. Jadama and Modou K. Toray provided an in-depth overview of ensemble learning techniques, including boosting, bagging, and stacking. The study highlights the superiority of ensemble methods, particularly XGBoost, in applications such as intrusion detection, traffic accident severity prediction, and kidney disease diagnosis. Strengths include reducing over-fitting (bagging), addressing under-fitting (boosting), and achieving optimal predictions by combining diverse models (stacking). The research emphasizes ensemble approaches for mitigating bias, reducing variance, and enhancing model stability and performance, with future directions focusing on improving prediction accuracy, expanding applications, and minimizing computational costs [7].

Palak Mahajan et al reviewed ensemble learning techniques for disease prediction, focusing on four methods: bagging, boosting, stacking, and voting. The study analyzed 45 articles (2016–2023) that applied these techniques to five diseases: diabetes, skin disease, kidney disease, liver disease, and heart conditions. Stacking demonstrated the highest accuracy, particularly for skin disease and diabetes, followed by voting. Bagging performed best for kidney disease, while boosting excelled in liver and diabetes predictions. Strengths include improved prediction accuracy, variability in performance across datasets, and insights into selecting suitable ensemble models for disease analytics [8].

Behzad Naderalvojoud and Tina Hernandez-Boussard explored the effectiveness of ensemble learning in observational healthcare data, focusing on scenarios where individual models may perform inconsistently. The study proposed an ensemble model for predicting patients at risk of prolonged postoperative opioid use, integrating two machine learning models trained with different covariates. Strengths include significant improvement in AUROC and AUPRC metrics, enhanced precision, and providing insights into conditions under which ensemble approaches outperform individual models in healthcare prediction tasks [9].

Bauer and Kohavi (1999) conducted an empirical study comparing voting classification algorithms, including bagging, boosting, and their variants. The research evaluated the performance of these ensemble techniques across various datasets, analyzing their strengths and weaknesses. Strengths include bagging's robustness in reducing variance, boosting's ability to reduce bias while enhancing accuracy, and insights into the trade-offs of different variants. This foundational work highlighted the effectiveness of ensemble methods in improving classification performance and provided guidance on selecting suitable algorithms for specific tasks [10].

III. PROBLEM STATEMENT

Cardiovascular diseases (CVDs) are a leading cause of mortality worldwide, affecting millions annually and requiring early detection to prevent severe outcomes like heart attacks and strokes. Current diagnostic methods, reliant on manual ECG interpretation, are time-intensive, expertise-dependent, and prone to human error. The growing complexity and volume of ECG data further challenge accurate diagnosis, especially in resource-limited settings. The objective is to develop an automated system that can quickly and accurately analyze ECG data to detect subtle changes indicative of early CVD onset, improving diagnostic efficiency, reducing human effort, and enabling timely intervention to prevent severe health outcomes.

IV. OBJECTIVES

The primary objective of the project is to develop an efficient and automated system for predicting cardiovascular diseases from ECG images. To achieve this, several specific objectives need to be met:

o Develop an automated system to predict cardiovascular diseases from ECG images using ensemble techniques.

o Preprocess ECG images with grayscale conversion, Gaussian smoothing, and thresholding to remove noise.

o Extract 12 standard ECG leads and 1 long lead for focused signal analysis.

o Perform feature extraction using contour extraction, PCA, and normalization.

o Build an ensemble model combining Random Forest, Gradient Boosting, and SVM for accurate prediction.

o Enable early diagnosis to support timely clinical decisions.

.

V. MOTIVATION

Cardiovascular diseases (CVDs) are a leading cause of mortality, requiring early and accurate diagnosis. Traditional manual ECG interpretation can be time-consuming and error-prone. This project uses machine learning ensemble techniques to analyze ECG images, enhancing prediction accuracy and reliability. The goal is to assist healthcare professionals in diagnosing CVDs efficiently, reduce workload, and improve heart health management, especially in resource-limited settings.

VI. SYSTEM DESIGN

The system design for Cardiovascular Disease (CVD) detection represents an innovative integration of multiple components, working in harmony to achieve accurate and reliable outcomes. At the heart of this system lies the use of Electrocardiogram (ECG) images, which are collected as input data from trusted datasets and medical sources. These images form a well-organized datastore, serving as the foundation for all subsequent processes. To ensure that the data is primed for analysis, the system implements a series of meticulous preprocessing steps aimed at enhancing the quality of the ECG images. This stage is crucial, as it helps eliminate noise, improve clarity, and extract essential features from the images.

The preprocessing pipeline includes tasks such as dataset cleaning, gray scaling to standardize image formats, Gaussian smoothing to minimize noise, and dividing ECG leads for better focus on individual segments. Advanced techniques like finding threshold intensity using the Otsu method, binarizing images, and contour detection are applied to highlight critical aspects of the ECG signals. These steps collectively prepare the images, transforming raw input into a form that is ideal for detailed analysis. Once the preprocessing phase is complete, the images are divided into two distinct subsets: 80% for training and 20% for testing. This ensures that the system has ample data to learn from while reserving a portion for unbiased evaluation. During the training phase, the system employs ensemble classifiers, which are sophisticated Machine Learning (ML) models designed to capture patterns and relationships within the preprocessed images. Multiple classifiers are evaluated, with the goal of selecting the one that delivers the best performance based on accuracy metrics. This rigorous selection process ensures that the final predictive model is both robust and reliable. The chosen model then becomes the centerpiece of the system, capable of analyzing new ECG images and classifying them into specific categories. These categories include "Normal Person," "Abnormal Heartbeat," "Myocardial Infarction," and "History of Myocardial Infarction."

The classification results are not only precise but also carry significant clinical value, offering healthcare professionals valuable insights into a patient's cardiac health. This information can play a pivotal role in the early diagnosis of conditions, as well as in creating tailored treatment plans that address individual needs. The modular design of the system is another standout feature, offering flexibility for future enhancements. Whether it involves incorporating additional preprocessing techniques, integrating more advanced classifiers, or expanding the range of detectable conditions, the system is built to evolve alongside advancements in technology and medical science. Ultimately, this innovative approach underscores the potential of technology-driven solutions in transforming healthcare. By enabling early detection and providing actionable insights, the system has the power to contribute significantly to preventive healthcare and personalized medicine. It not only highlights the promise of integrating technology into medical practices but also brings hope for a future where Cardiovascular Diseases can be detected and managed with unparalleled precision.

The proposed framework combines image preprocessing techniques, machine learning algorithms, and classification processes to deliver accurate predictions. Model development is supported by essential libraries such as scikit-learn, scikitimage, joblib, pandas, matplotlib, natsort, and streamlit. These tools facilitate efficient model training, testing, and performance evaluation by streamlining data analysis and visualization. The predictive model is fine-tuned to ensure robust performance and accuracy, enabling it to effectively handle complex and subtle variations in ECG data. Once the classification results are generated, they can be used to provide actionable insights for early diagnosis and intervention.

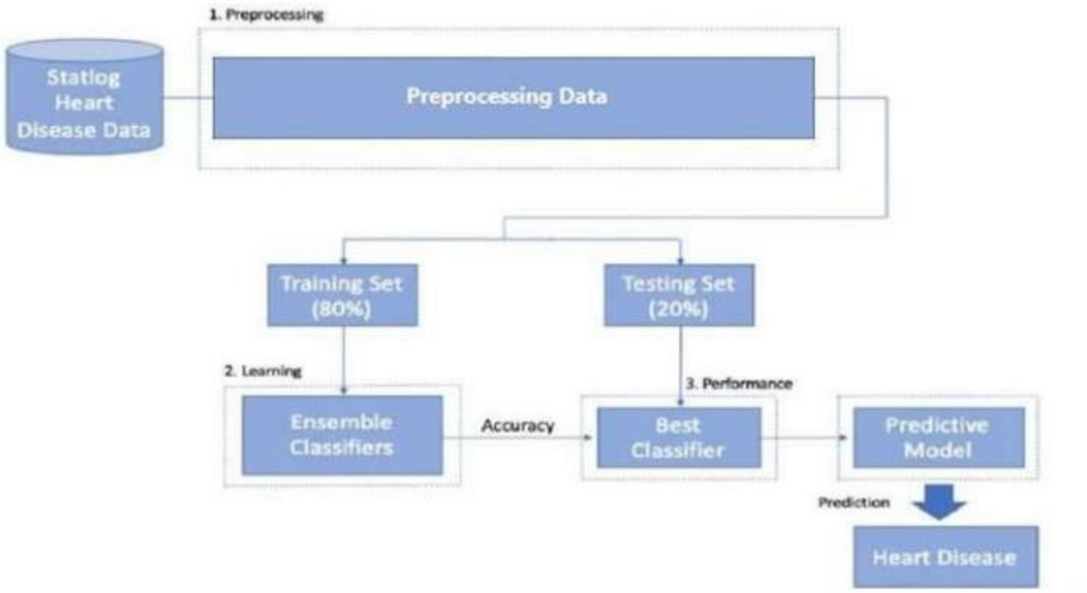


Fig 4.1 Architectural Design

Furthermore, the system reduces the reliance on manual ECG interpretation, which is often time-consuming and expertisedependent, by automating the detection of abnormalities. This not only enhances diagnostic efficiency but also significantly reduces human error, particularly in resource-limited healthcare settings. By integrating preprocessing, model training, and predictive analysis into a cohesive workflow, the system offers a scalable and reliable solution for CVD detection. Ultimately, it supports healthcare professionals in delivering timely and accurate diagnoses, contributing to improved patient outcomes and overall healthcare efficiency.

VII. IMPLEMENTATION

A. Ensemble Modeling for Cardiovascular Disease Prediction

Ensemble modeling is a powerful technique that combines predictions from multiple base learners to form a robust and accurate prediction model. It leverages the collective wisdom of individual models, reducing bias and variance while enhancing overall prediction accuracy. For this project, we employed soft voting within the ensemble framework, a method suitable for cardiovascular disease classification tasks by aggregating probabilistic outputs of base classifiers.

1) Voting Mechanisms:
Hard vs. Soft Voting

- Hard Voting: Aggregates predictions using majority voting. The final output is determined by the mode (majority class) of predictions from all base learners.
- Soft Voting: Aggregates predicted probabilities from all base classifiers and selects the class with the highest summed probability. In this study, soft voting was preferred as it provides probabilistic estimates, leading to improved accuracy and confidence in predictions.

Input Images: The ensemble model for cardiovascular disease prediction takes patient data as input for classification tasks. This data includes a variety of features such as age, gender, blood pressure, cholesterol levels, and ECG readings. The input data can vary in format, but it is typically structured as numerical or categorical values, ready for preprocessing before feeding into the model.

B. Base Learners

Base learners are individual models that contribute to the ensemble prediction. Their combination captures various aspects of the
data, enhancing robustness and reducing errors. The base learners employed in this study are as follows:

1. Support Vector Machine (SVM)
   SVM constructs a hyperplane (or multiple hyperplanes) in a high-dimensional space to classify data points by maximizing the margin between classes.
2. Logistic Regression (LR)
   Logistic Regression models the probability of a data point belonging to class y=1
3. K-Nearest Neighbors (KNN)
   The KNN algorithm classifies data based on the similarity of nearby data points. For a given query point xxx, distances to all training points xi are computed using a distance metric such as Euclidean distance
4. Random Forest
   Random Forest is an ensemble learning method that combines multiple decision trees to improve predictive accuracy and robustness. For a given input x, the prediction is made by aggregating the outputs of N trees.
5. Gaussian Naive Bayes (GNB)
   Gaussian Naive Bayes is a probabilistic classification algorithm based on Bayes' theorem, assuming that the features are conditionally independent given the class label.

C. System Workflow

The workflow of the ensemble-based cardiovascular disease prediction system involves the following steps

1. Data Acquisition: Collecting a dataset containing patient records, including relevant features for cardiovascular disease prediction.
2. Data Preprocessing: Cleaning and transforming the raw data, handling missing values, and scaling numerical features to ensure model compatibility.
3. Model Training: Training individual base learners, namely SVM, Logistic Regression, and KNN, on the prepared dataset.
4. Soft Voting Implementation: Combining the probabilistic outputs of all base learners to predict the final class label using the soft voting mechanism.
5. Model Evaluation: Assessing the performance of the ensemble model using evaluation metrics such as accuracy, precision,
   recall, and F1-score to ensure reliability.
6. Fault Detection: Validating the model's ability to detect anomalies and identify patients with cardiovascular disease
   effectively.

Output: The model classifies the input data into one of four categories: Normal, Abnormal Heartbeat, Myocardial Infarction, or History of Myocardial Infarction. Each prediction is accompanied by a confidence score indicating the model’s certainty.The final classification is based on the aggregated probabilities from multiple base learners, ensuring robust and accurate predictions for cardiovascular disease risk.

VII.RESULT AND PERFORMANCE ANALYSIS

The performance of various models was evaluated using the experimental dataset, and their accuracy, precision, recall, and F1-scores were analyzed. Individual models, including Logistic Regression, Support Vector Machines (SVM), K-Nearest Neighbors (KNN), and XGBoost, were tested and compared against the ensemble approach.

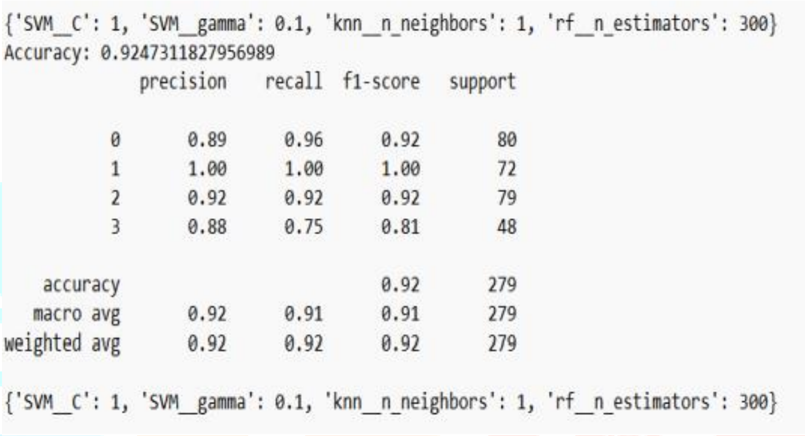
Individual Model Performance

1. Logistic Regression: The model achieved an accuracy of 79.3%, with a weighted F1-score of 0.79. While Logistic Regression showed strong performance on Class 1 with a precision of 0.95, its results for Class 0 and Class 3 lagged behind with F1-scores of 0.76 and 0.69, respectively.
2. Support Vector Machine (SVM): Tuned with parameters {'SVM__C': 10, 'SVM__gamma': 0.01}, the SVM model attained an accuracy of 77.7% and a weighted F1-score of 0.77. It demonstrated better recall for Class 1 (0.91) but struggled with Class 3, yielding an F1-score of 0.67.

3. K-Nearest Neighbors (KNN): With the optimal parameter {'knn__n_neighbors': 1}, the KNN model delivered an accuracy of 85.3%, with balanced performance across all classes. However, Class 3 had a slightly lower F1-score of 0.81, indicating room for improvement in handling this class.
4. XGBoost: Among the individual models, XGBoost performed the best with an accuracy of 90.5% and a weighted F1- score of 0.91. Its performance was particularly notable for Class 1, achieving perfect scores in precision, recall, and F1. However, Class 3 presented challenges, with a reduced recall of 0.78.

Ensemble Model Performance

The ensemble model, which integrates predictions from multiple classifiers, demonstrated superior performance compared to individual models. With an accuracy of 92.47%, the ensemble approach outperformed even the best individual model (XGBoost). The weighted F1-score of 0.92 highlights the robustness of the ensemble strategy in capturing complex patterns and reducing errors across classes. Class-wise analysis showed that the ensemble model achieved near-perfect scores for Class 1, with an F1-score of 1.00, and strong performance across all other classes, with F1-scores exceeding 0.81.



```
{'SVM__C': 1, 'SVM__gamma': 0.1, 'knn__n_neighbors': 1, 'rf__n_estimators': 300}
Accuracy: 0.9247311827956989
              precision    recall  f1-score   support

           0       0.89      0.96      0.92        80
           1       1.00      1.00      1.00        72
           2       0.92      0.92      0.92        79
           3       0.88      0.75      0.81        48

    accuracy                           0.92       279
   macro avg       0.92      0.91      0.91       279
weighted avg       0.92      0.92      0.92       279

{'SVM__C': 1, 'SVM__gamma': 0.1, 'knn__n_neighbors': 1, 'rf__n_estimators': 300}
```

Fig 6.1 Performance Metrics of Ensemble Model

By combining the strengths of multiple models, the ensemble approach mitigated the weaknesses observed in individual classifiers, particularly for challenging classes such as Class 3. The ensemble model's ability to balance precision and recall across all classes underscores its adaptability to diverse and imbalanced datasets. Furthermore, its scalability and integration potential make it a promising choice for real-world applications requiring high accuracy and reliability. This approach also demonstrated resilience to overfitting, benefiting from the diverse perspectives of its constituent classifiers. The results suggest that the ensemble model can serve as a robust framework for tasks demanding consistent and generalizable performance. Its success highlights the importance of leveraging complementary model strengths, paving the way for more innovative ensemble techniques in future research.

VIII. CONCLUSION

The development and implementation of the cardiovascular disease prediction system have demonstrated the effectiveness of ensemble learning methods in achieving high predictive accuracy. The ensemble model, utilizing a combination of SVM, KNN, and Random Forest classifiers, achieved a remarkable accuracy of 92.47%, surpassing the performance of individual models such as SVM, KNN, XGBoost, and logistic regression. The evaluation metrics, including precision, recall, and F1-score, further emphasized the robustness and reliability of the ensemble approach. These results validate the hypothesis that integrating multiple classifiers can leverage their unique strengths, resulting in a more accurate and generalized predictive model. The system features a well-structured modular design, with each module playing a critical role in the pipeline. The preprocessing and image processing modules ensure that the ECG data is cleaned, normalized, and transformed into a suitable format for feature extraction. The Streamlit-based frontend provides an intuitive interface for users, enabling seamless interaction with the system for uploading ECG images and receiving predictions. The integration of these components makes the system both accessible and scalable for real-world applications. This work highlights the potential of combining machine learning techniques and medical imaging for early detection and diagnosis of cardiovascular diseases. By providing a non-invasive, efficient, and accurate diagnostic tool, the project can significantly contribute to reducing the burden of cardiovascular diseases globally. Future enhancements, such as incorporating larger and more diverse datasets or exploring advanced deep learning architectures, can further improve the system's predictive capabilities and applicability in broader healthcare settings.

ACKNOWLEDGEMENT

We convey our sincere thanks to Rajya Vokkaligara Sangha, Bangalore and our guide Dr. Hema Jagadish, Associate Professor, Department of Information Science and Engineering, Bangalore Institute of Technology, without whose direction, this would not have been possible. We also express our gratitude to our team members whose team participation resulted in successful completion of the paper. We are also grateful to our institution and its faculty for providing the necessary resources and a conducive environment for research. Lastly, we extend our heartfelt thanks to our families and friends for their unwavering support and encouragement throughout this endeavor.

REFERENCES

[1] McAuliffe, Matthew J., et al. "Medical image processing, analysis and visualization in clinical research." Proceedings 14th IEEE symposium on computer-based medical systems. CBMS 2001. IEEE, 2001.

[2] Scholl, Ingrid, et al. "Challenges of medical image processing." Computer science-Research and development 26 (2011): 5-13.

[3] Ritter, F., et al. "Medical Image Analysis." IEEE Pulse, vol. 2, no. 6, Nov.-Dec. 2011, pp. 60–70, https://doi.org/10.1109/MPUL.2011.942929

[4] Angenent, Sigurd, Eric Pichon, and Allen Tannenbaum. "Mathematical methods in medical image processing." Bulletin of the American mathematical society 43.3 (2006): 365-396.

[5] Nandi, Dibyadeep, et al. "Principal component analysis in medical image processing: a study." International Journal of Image Mining 1.1 (2015): 65-86.

[6] Tabassum, Tahmida, and Mohiuddin Ahmad. "Numerical data extraction from ECG paper recording using image processing technique." 2020 11th international conference on electrical and computer engineering (ICECE). IEEE, 2020.

[7] Jadama, Ansumana F., and Modou K. Toray. "Ensemble Learning Techniques: Boosting, Bagging, and Stacking." Journal of Machine Learning Research and Applications, vol. 15, no. 3, 2023,

[8] Mahajan, Palak, et al. "Ensemble learning for disease prediction: A review." Healthcare. Vol. 11. No. 12. MDPI, 2023.

[9] Naderalvojoud, Behzad, and Tina Hernandez-Boussard. "Improving machine learning with ensemble learning on observational healthcare data." AMIA Annual Symposium Proceedings. Vol. 2023. 2024.

[10] Bauer, Eric, and Ron Kohavi. "An Empirical Comparison of Voting Classification Algorithms: Bagging, Boosting, and Their Variants." Machine Learning, vol. 36, no. 1-2, 1999, pp. 105–139. Springer, doi:10.1023/A:1007515423169.