AI-Powered Early Diagnosis and Personalized Health Recommendations for Coronary Artery Disease (CAD) using Predictive Analytics

Siddharth Vats
Department of Computing Technologies, School of Computing
SRM Institute of Science and Technology
Tamil Nadu, India
siddharthvats44@gmail.com

Khayati Sharma
Department of Computing Technologies, School of Computing
SRM Institute of Science and Technology
Tamil Nadu, India
sharmakhayati0123@gmail.com

M. Revathi
Department of Computing Technologies, School of Computing
SRM Institute of Science and Technology
Tamil Nadu, India
revathim@srmist.edu.in

***Abstract*—Cardiovascular diseases, especially coronary artery disease, have been the leading cause of deaths worldwide. Hence, the early detection and personalized treatment strategies for coronary artery disease (CAD) are now needed more than ever. This paper suggests an AI-driven framework based on machine learning algorithms for the timely risk detection of CAD based on 13 clinical parameters. The framework utilizes an ensemble learning method that includes logistic regression, random forest and support vector machine classifiers which improve the prediction accuracy. The proposed system also comes equipped with an AI chemist assistant which helps users specifically with medication-related queries, such as drug interactions, administration instructions and prescription explanations. This feature managed textual and visual inputs using Google's generative AI (Gemini), thus allowing users to upload photos or ask questions and receive instant assistance. Adding to the suite of features, the system also comes with a health recommendation engine, which provides users with certain lifestyle changes to be adopted based on the risk assessment, and a chatbot which answers more general questions that the user might have related to his diet, exercise, sleep or any symptoms being experienced. This approach provides an integrated preventative care solution by combining CAD risk prediction, AI-based medication assistance, a coronary health chatbot and lifestyle recommendations. Apart from improving accessibility, the suggested solution also improves patient education and enables more informed decision-making in cardiovascular disease care. The experimental results show the effectiveness of the ensemble model in risk prediction of CAD and AI chemist's capacity to help users with medication-related concerns. This integrated system aims to combine AI-based diagnostic techniques with medicative and general heart-health assistance, thus improving the efficiency and accessibility of healthcare services.**

Keywords—coronary artery disease (CAD), ensemble learning, generative AI, coronary health chatbot, AI medication assistant

I. Introduction

Cardiovascular diseases (CVDs), and coronary artery disease (CAD) in particular, continue to be a major cause of morbidity and mortality worldwide [1]. Early diagnosis and risk stratification are essential to avoid serious complications. However, standard diagnostic procedures often involve long clinical assessment, advanced equipment, and specialist interpretation. Advances in artificial intelligence (AI) and machine learning (ML) have revolutionized predictive analytics into a valuable tool in medical diagnosis, with enhanced accuracy, efficiency, and accessibility [2].

The system in this case is built based on research with artificial intelligence to facilitate predictive analytics for coronary artery disease risk assessment and individualized health recommendations. To improve the prediction accuracy, the system integrates a hybrid ensemble approach that combines three distinct machine learning models: logistic regression, random forest, and a support vector machine (SVM). This combination leverages the strengths of each algorithm to produce more robust and reliable results. The models leverage clinical result data and integrate basic cardiovascular risk predictors to yield patient-specific risk measurements. This research goes beyond typical predictive modeling by incorporating an AI-enhanced chemist instrument that employs LLMs to facilitate users' understanding of medicines, drug interactions, and life changes. Integrating a multimodal AI system that can analyze text-based input and visual cues enhances user engagement and usability.

The primary objective of this research is to develop an AI-powered system capable of early diagnosis and risk assessment for Coronary Artery Disease (CAD) using predictive analytics. Additionally, the system integrates an AI Chemist Assistant to provide medication-related guidance and answer users' queries about drugs, side effects, and interactions. The proposed framework leverages machine learning algorithms for CAD prediction and natural language processing (NLP) for AI-driven patient interaction, enhancing accessibility and engagement in healthcare.

II. Related Work

Cardiovascular disease prediction has been a major focus of research with several studies using machine learning to improve diagnostic performance and detect diseases early. Many have attempted various models, datasets, and methods to enhance predictive systems. Seckeler and Hoke [3] presented a detailed overview of the epidemiology of rheumatic heart disease and its burden and long-term public health impact. Early detection remains a recurring theme in heart disease research, particularly within vulnerable and underserved populations. Several studies underscore the urgency of diagnosing cardiovascular conditions before they progress to severe stages. For instance, Gaziano et al. [4] raised concern over the growing burden of coronary artery disease (CAD) in low- and middle-income countries, drawing attention to how limited access to timely care often exacerbates patient outcomes. Their findings emphasize the need for scalable and proactive diagnostic solutions in such regions.

Building on this Boukhatem et al. [5] investigated machine learning-based frameworks for predicting cardiac disease, illustrating how heterogeneous feature extraction techniques may stabilize models. Their findings suggested that advanced machine learning techniques, when used appropriately, may advance early diagnosis and treatment. In a related effort, Jindal et al. [6] examined a diverse set of machine learning algorithms for predicting heart disease, identifying critical variables with a strong impact on prediction accuracy. The study conducted by Ramalingam et al. [7] focused on balancing performance with interpretability, a crucial consideration in medical decision-making environments where transparency is key. Their findings emphasized the trade-off between model complexity and usability. Weng et al. [8] described the integration of machine learning with traditional clinical information to estimate cardiovascular risk, underlining how artificial intelligence-based methods can recognize patients at risk prior to clinical symptoms appearing.

Fatima and Pasha [9] conducted a comparative analysis evaluating a wide range of machine learning methods, offering insight into accuracy trends and practical limitations across different use cases. Their paper emphasized the pros and cons of various algorithms in terms of accuracy, computational complexity, and interpretability. Vembandasamy et al. [10] explained the applicability of the Naïve Bayes algorithm to detect cardiac disease. From their findings, Naïve Bayes offers an efficient and interpretable strategy but its accuracy is frequently surpassed by more sophisticated models. Chaurasia and Pal [11] applied data mining to identify cardiac conditions and established the capability of ensemble learning methods to enhance predictive accuracy. Bhatt et al. [12] subsequently employed a 70,000 sample dataset and compared different models to determine the most appropriate method to predict heart disease. A number of works, including those by Patel et al. [13], explored the impact of robust pre-processing and feature selection techniques on predictive accuracy. One of their key contributions was a novel approach to filtering out redundant features, ultimately leading to more refined and computationally efficient models. Rindhe et al. [14] investigated different machine learning architectures for the prediction of the cardiac disease. Their study relied on the Cleveland Heart Disease dataset [15], a benchmark in cardiovascular research. The study proved that decision tree and support vector machine models were of high accuracy when they were trained with hyperparameters that were optimized. Tithi et al. [16] concentrated their research on the assessment of ECG data and the prediction of cardiovascular disease based on the implementation of six supervised learning models. Their research highlighted the implementation of feature selection and pre-processing methods in enhancing the performance of the model. Garg et al. [17] focused their attention on supervised learning techniques, evaluating how different model architectures and pre-processing pipelines influenced diagnostic precision. Their research compared many classification models and concluded that ensemble-based models have an incredible impact on enhancing predictive accuracy compared to conventional machine learning models.

These papers together show the shifting paradigm of cardiac disease prediction using machine learning approaches. The diversity of the methodologies and data sets utilized in these papers improve the insight into the problems and advancements in this field.

III. Methodology

A. Dataset and Preprocessing

This study utilizes the Cleveland Heart Disease dataset [15] obtained from the UCI Machine Learning Repository (DOI: 10.24432/C52P4X). The dataset contains 303 patient records, with 13 clinical parameters and one target variable indicating the presence of coronary artery disease (CAD). The major features used for prediction include:

- Demographic factors: age, gender
- Cardiovascular health indicators: chest pain type, resting blood pressure, cholesterol levels, fasting blood sugar
- ECG and stress test results: resting ECG, maximum heart rate achieved, exercise-induced angina
- Imaging and genetic factors: number of major blood vessels colored by fluoroscopy, thalassemia

The target variable is coded from 0 to 4, where 0 represents no CAD and values 1 to 4 indicate varying severity levels of CAD. This study reformulated the classification task as a binary task, so values greater than zero are labeled as "presence of CAD." Prior to model training, an extensive workflow of preprocessing steps was applied to improve the quality of the data and model performance:

- Handling of missing values: the dataset was examined for any possible missing values, which were appropriately addressed with regards to data quality.
- Categorical encoding: features like chest pain type and thalassemia were converted into numerical form through appropriate encoding techniques to enable the ML models to process the categorical data effectively. This ensured that the model could interpret and utilize the these variables during training without introducing bias or misinformation.
- Scaling features: continuous parameters, such as cholesterol and resting blood pressure, were normalized to create comparable feature distributions and model performance.

These pre-processing steps were significant for the dataset to be properly used for training and evaluating of the predictive models.

B. Model Selection and Training

In order to enhance accuracy and robustness of prediction, a voting classifier ensemble was developed via three models of machine learning, with each model having unique advantages in the classification process.

- Logistic regression: as a linear model, logistic regression is an effective binary classifier, which makes it a simple baseline model that is interpretable and computationally efficient.
- Random forest classifier: it is built on an ensemble of decision trees that work collectively to improve

prediction performance. Unlike a single decision tree, which may be prone to variance or overfitting, a random forest introduces randomness in both data and feature selection (bagging), making the model more robust. It aggregates the outputs of multiple trees through a voting mechanism, allowing the final decision to reflect a broader consensus across different tree perspectives.

- Support vector machine: it is a strong classifier that provides optimal decision boundaries, is focused on high dimension space, and non-linear detection relationships.

The ensemble employed a technique of soft voting, where each model's prediction represented a probability of membership of classes. The output was made via a weighted average of the models' probability membership. This meant that a stronger level of confidence from a model would have more influence on the final prediction from the ensemble, and improved level of classification overall. Once the data was trained, it was split 80% training to 20% test, which allowed for balanced evaluation of the ensemble. The training data was used to fit each model to; followed by an ensemble aggregation where each model's predictions were jointly utilized to make the classification. The method has assumed that complementing advantages are relied on.

C. System Architecture

The system is implemented as a streamlit-based web application, providing users with an interactive interface for comprehensive heart disease risk assessment. Users can input their clinical parameters, and the system processes this data to generate coronary artery disease (CAD) risk predictions using a pre-trained ensemble learning model. Additionally, the application offers personalized lifestyle recommendations based on predefined medical guidelines, assisting users in managing their cardiovascular health.

A unique feature of this system is the integration of an AI-powered chemist, designed to provide medical and pharmaceutical insights. This functionality is powered by Google Gemini, enabling the chatbot to answer user queries regarding medications, side effects, and potential drug interactions. Furthermore, an AI-driven heart health assistant leverages OpenAI's GPT model to provide additional guidance, ensuring users receive real-time and context-aware medical insights.

The backend architecture is designed for efficiency and scalability. The trained machine learning model is stored as a joblib object, allowing for fast and consistent predictions without requiring retraining during runtime. The system is structured to enable seamless integration with APIs, ensuring smooth AI-driven interactions and chatbot functionalities.

D. AI Chemist Assistant for Medication Guidance

To enhance the system's usability, an AI Chemist Assistant is integrated, leveraging large language models (LLMs) to provide medication-related insights. This assistant enables users to:

- Understand prescribed medications, including dosage, side effects, and interactions.
- Upload images of medication labels for AI-driven recognition and detailed explanations.
- Receive AI-generated responses for general health inquiries, improving accessibility to healthcare knowledge.

The AI Chemist module leverages Google's Gemini 1.5 Pro, a versatile generative AI model that can handle both textual and visual information. This capability allows users to interact naturally by either typing their queries or uploading images, making the assistant particularly helpful for tasks like identifying medicines and explaining their usage.

E. Model Evaluation

To evaluate the effectiveness of the trained ensemble model, a set of widely accepted classification metrics were employed, each offering a unique perspective on performance.

- Performance Indicators: Different principal metrics used for evaluating the model include:
  - Accuracy: it provides an overall indication of how many predictions the model got right across all classes. However, in healthcare scenarios—especially involving heart disease—accuracy alone can be misleading.
  - Precision and Recall: precision answers the question: of all patients predicted to have CAD, how many actually do? On the other hand, recall focuses on the model's ability to detect all actual cases of CAD, which is particularly important to minimize the risk of overlooking high-risk individuals.
  - F1-Score: it combines both precision and recall into a single number, making it especially valuable in situations where the dataset may have slightly imbalanced classes.
- ROC-AUC (Receiver Operating Characteristic - Area Under Curve) Score: this analysis was conducted to assess the model's ability to distinguish between the presence and absence of disease across different classification thresholds. A higher AUC indicates that the model consistently performs well at this separation task.
- Confusion matrix analysis: the confusion matrix (see Fig.1) was used to visualize how many predictions fell into each category—true positives, false positives, true negatives, and false negatives—offering granular insight into the model's strengths and shortcomings.
- Precision-Recall curve: the precision-recall (PR) curve (see Fig. 2) holds more significance during an imbalanced classification case as it focuses on model performance with the positive class. The PR curve visualization shows how precision and recall trade off at distinct classification thresholds, allowing the threshold to be tuned according to the precise clinical needs.

- Receiver operating characteristic curve: the ROC curve (see Fig. 3) offers a visual interpretation of how the model balances sensitivity and specificity across varying thresholds. This helps in selecting the most appropriate cutoff value when deploying the model in clinical decision-making, ensuring that the trade-off between false alarms and missed diagnoses is well managed.

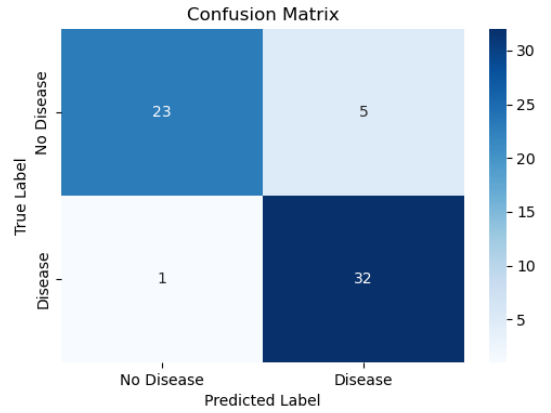

Fig. 1. Confusion matrix of the ensemble model

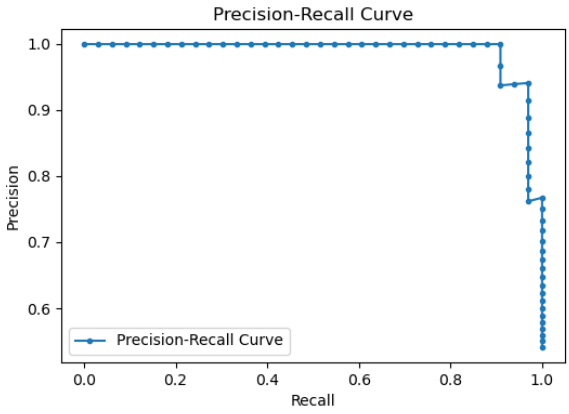

Fig. 2. Precision Recall curve of the ensemble model

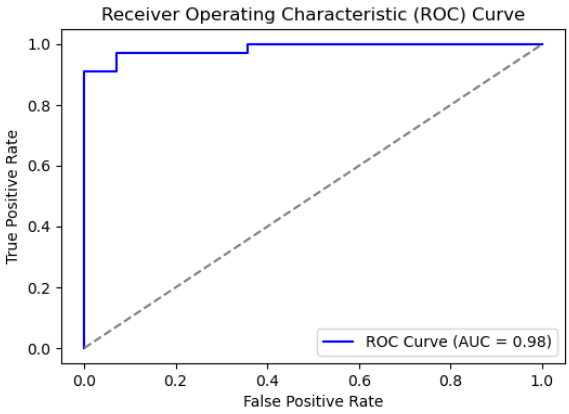

Fig. 3. ROC curve of the ensemble model

TABLE I. EVALUATION RESULTS FOR THE MODEL

Accuracy	0.9016
Presicion	0.8649
Recall	0.9697
F1-Score	0.9143
ROC-AUC Score	0.9848

F. Model Deployment

The deployment process follows a structured and modular approach to ensure replicability and ease of use. The system is implemented using python with scikit-learn for machine learning functionalities and streamlit for an intuitive, web-based user interface. The pre-trained model is stored using joblib, facilitating quick loading and inference.

To enhance usability, the system supports API-based interactions for integrating AI-powered assistants. The modular design allows for future enhancements and scalability, enabling the inclusion of additional functionalities such as expanded health analytics or integration with wearable health-monitoring devices. This approach ensures that the system remains efficient, explainable, and adaptable to evolving medical and technological advancements.

IV. RESULTS

The developed system successfully integrates machine learning-based CAD risk prediction with an interactive web application powered by streamlit. The performance evaluation of the trained ensemble model demonstrated high accuracy (90.16%), indicating reliable classification of patients at risk of coronary artery disease (CAD). The model achieved a precision of 86.49%, ensuring that most patients predicted to have CAD truly have the condition. Additionally, the recall of 96.97% highlights the model's effectiveness in identifying at-risk patients, minimizing the chances of missing a potential CAD case. The F1-score (91.43%) and ROC-AUC (98.93%) further confirm the model's robustness in distinguishing between healthy and at-risk individuals.

Beyond model performance, the streamlit-based web application provides an accessible and user-friendly interface for CAD risk assessment. Users can input their clinical parameters, and based on the model's prediction, they receive personalized health recommendations (see Fig. 4). These recommendations are generated to promote a heart-healthy lifestyle, offering tailored advice on diet, exercise, and other preventive measures. The web application is further enhanced by AI-driven conversational agents. Users can consult the AI-powered health chatbot (see Fig. 5), which assists them in understanding their symptoms, potential risk factors, and general heart health concerns. Additionally, the AI Chemist (see Fig. 6) provides detailed pharmaceutical insights, answering queries related to medications, including their composition, uses, and possible side effects. This integrated approach ensures that users not only receive risk assessments but also gain

valuable medical knowledge to make informed health decisions.

The results confirm that this system is not just a predictive model but a comprehensive health advisory tool, bridging the gap between AI-driven diagnosis and patient education. The combination of high model performance, real-time health recommendations, and AI-powered assistance makes this system a promising solution for CAD risk assessment and patient engagement.

By leveraging a collective model strategy, the system was able to achieve strong performance metrics in predicting CAD risk, highlighting the strength of ensemble-based methods in clinical data classification. It demonstrates strong potential for real-world deployment in both clinical and telemedicine environments, offering early risk assessment, improved patient engagement, and AI-assisted medication guidance. Further refinements and validation efforts could help transition this model into a fully integrated clinical decision-support system in the future.

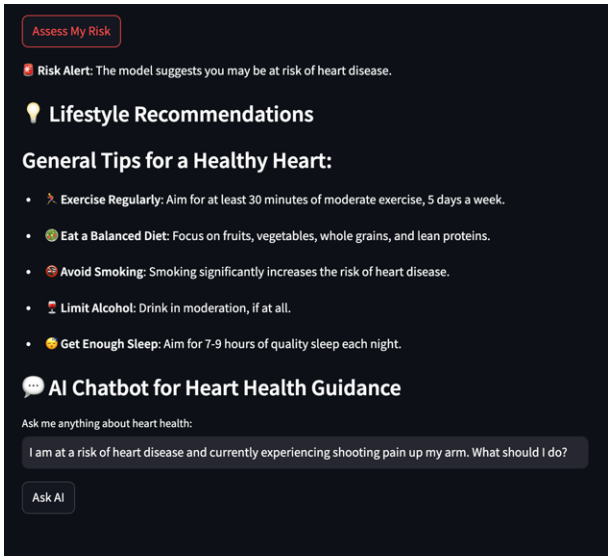

Fig. 4. Personalized health recommendations based on the level of risk

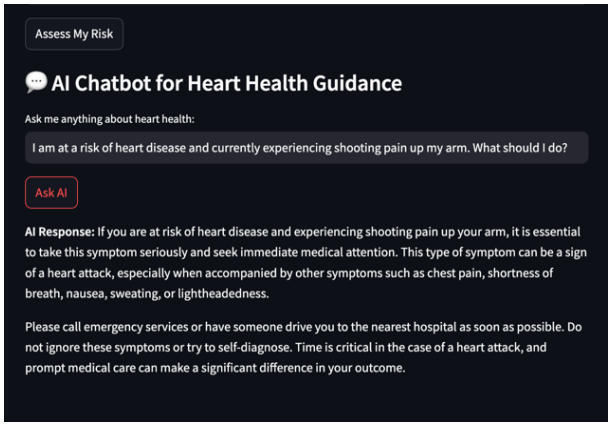

Fig. 5. AI chatbot for patient's query resolution

Fig. 6. AI chemist module for resolving drug related queries

V. Conlusion

This research presents an integrated AI-driven framework for coronary artery disease (CAD) risk prediction and patient support, combining machine learning-based diagnostics with an AI-powered chemist assistant. The model effectively combines logistic regression, random forest, and support vector machines into an ensemble structure, which enhances its ability to correctly identify individuals at potential risk of coronary artery disease with strong predictive precision. The system's integration into a user-friendly web application enhances accessibility, allowing users to receive real-time risk assessments and personalized health recommendations.

A key innovation of this research is the incorporation of an AI Chemist Assistant powered by Google Gemini, which provides users with pharmaceutical guidance, including medication explanations, drug interactions, and dosage instructions. This feature enhances patient education and supports informed decision-making, making the system a comprehensive healthcare tool. Evaluation of the model yielded an accuracy of 90.16%, alongside high precision, recall, and F1-score values. These consistent outcomes

highlight the system's robustness and its potential for practical deployment in real-world healthcare settings. Beyond its predictive capabilities, the system contributes to improving healthcare accessibility by offering AI-driven assistance in disease management. This framework brings together coronary risk prediction and AI-guided medication support, creating a more holistic tool that not only assists in clinical decision-making but also supports patients in understanding and managing their cardiovascular health.

Future work will focus on expanding the dataset, incorporating additional biomarkers, and refining the AI assistant's capabilities to further enhance diagnostic accuracy and usability. With continuous improvements, this AI-powered system has the potential to evolve into a valuable clinical decision-support tool, benefiting both healthcare professionals and patients in the early detection and management of CAD.

REFERENCES

[1] World Health Organization, *"Cardiovascular diseases (CVDs)",* WHO, Jun. 11, 2021. [Online]. Available: https://www.who.int/news-room/fact-sheets/detail/cardiovascular-diseases-(cvds) [Accessed: Apr. 10, 2025].

[2] ForeSee Medical, *"Artificial Intelligence in Healthcare: Transforming the Future of Medicine",* ForeSee Medical, 2023. [Online]. Available: https://www.foreseemed.com/artificial-intelligence-in-healthcare [Accessed: Apr. 10, 2025].

[3] Seckeler, M. D., & Hoke, T. R. (2011). The worldwide epidemiology of acute rheumatic fever and rheumatic heart disease. Clinical epidemiology, 67-84.

[4] Gaziano, T. A., Bitton, A., Anand, S., Abrahams-Gessel, S., & Murphy, A. (2010). Growing epidemic of coronary heart disease in low-and middle-income countries. *Current problems in cardiology*, *35*(2), 72-115.

[5] Boukhatem, C., Youssef, H. Y., & Nassif, A. B. (2022, February). Heart disease prediction using machine learning. In *2022 Advances in Science and Engineering Technology International Conferences (ASET)* (pp. 1-6). IEEE.

[6] Jindal, H., Agrawal, S., Khera, R., Jain, R., & Nagrath, P. (2021). Heart disease prediction using machine learning algorithms. In *IOP conference series: materials science and engineering* (Vol. 1022, No. 1, p. 012072). IOP Publishing.

[7] Ramalingam, V. V., Dandapath, A., & Raja, M. K. (2018). Heart disease prediction using machine learning techniques: a survey. *International Journal of Engineering & Technology*, *7*(2.8), 684-687.

[8] Weng, S. F., Reps, J., Kai, J., Garibaldi, J. M., & Qureshi, N. (2017). Can machine-learning improve cardiovascular risk prediction using routine clinical data?. *PloS one*, *12*(4), e0174944.

[9] Fatima, M., & Pasha, M. (2017). Survey of machine learning algorithms for disease diagnostic. *Journal of Intelligent Learning Systems and Applications*, *9*(01), 1-16.

[10] Vembandasamy, K., Sasipriya, R., & Deepa, E. (2015). Heart diseases detection using Naive Bayes algorithm. *International Journal of Innovative Science, Engineering & Technology*, *2*(9), 441-444.

[11] Chaurasia, D. V., & Pal, S. (2014). Data mining approach to detect heart diseases. International Journal of Advanced Computer Science and Information Technology (IJACSIT) Vol, 2, 56-66.

[12] Bhatt, C. M., Patel, P., Ghetia, T., & Mazzeo, P. L. (2023). Effective heart disease prediction using machine learning techniques. *Algorithms*, *16*(2), 88.

[13] Patel, J., TejalUpadhyay, D., & Patel, S. (2015). Heart disease prediction using machine learning and data mining technique. *Heart Disease*, *7*(1), 129-137.

[14] Rindhe, B. U., Ahire, N., Patil, R., Gagare, S., & Darade, M. (2021). Heart disease prediction using machine learning. *Heart Disease*, *5*(1).

[15] A. Janosi, W. Steinbrunn, M. Pfisterer, and R. Detrano. "Heart Disease," UCI Machine Learning Repository, 1989. [Online]. Available: https://doi.org/10.24432/C52P4X.

[16] Tithi, S. R., Aktar, A., Aleem, F., & Chakrabarty, A. (2019, June). ECG data analysis and heart disease prediction using machine learning algorithms. In *2019 IEEE Region 10 Symposium (TENSYMP)* (pp. 819-824). IEEE.

[17] Garg, A., Sharma, B., & Khan, R. (2021). Heart disease prediction using machine learning techniques. In *IOP Conference series: materials science and engineering* (Vol. 1022, No. 1, p. 012046). IOP Publishing.