Heart Disease Prediction using Machine Learning

Lokesh Khedekar[1], Shail Kamtikar[2], Sarthak Kamtikar[3*], Krishnakant Kale[4], Pranav Kamble[5] and Eshan Kannawar[6]
{ lokesh.khedekar@vit.edu[1], shail.kamtikar24@vit.edu[2] , sarthak.kamtikar24@vit.edu[3], krishnakant.kale@24vit.edu[4], pranav.kamble24@vit.edu[5], eshan.kannawar24@vit.edu[6]}

Professor, Department of Information Technology, Vishwakarma Institute of Technology, Upper Indira Nagar, Bibwewadi, Pune, Maharashtra, India[1]
Department of Information Technology, Vishwakarma Institute of Technology, Upper Indira Nagar, Bibwewadi, Pune, Maharashtra, India[2, 3, 4, 5]

Abstract. Cardiovascular disease, and in particular myocardial infarction, is still a leading cause of death worldwide. Early prediction and timely interventions are important in reducing mortality. This paper presents a model for a Heart Attack Prediction System using machine learning algorithm that computes the cardiovascular risk based on essential clinical factors. It takes in chest patient information like age, blood pressure, cholesterol and lifestyle habits like whether the individual is a smoker to predict the likelihood of a heart attack. Performance evaluation indicates a correctness of 88.89%, which demonstrates that it could assist medical personnel in early diagnosis and preventive medicine. The system presented here targets better clinical decisions and reduced rates of hospital tradition and offers favorable benefits to patients.

Keywords: Heart Attack Prediction, Machine Learning, Cardiovascular Risk Assessment, Medical Diagnosis, Early Detection, Predictive Analytics, Heart Disease, Clinical Decision Support, Artificial Intelligence, Healthcare Technology.

1 Introduction

Myocardial infarction (MI) is one of the most fatal diseases and cardiovascular diseases (CVDs) are major contributors to global mortality. Heart attacks and strokes (as defined by coronary heart disease) are cardiovascular diseases (CVDs)[15] CVD account for nearly 18 million deaths each year worldwide, or approximately 32% of all global deaths 16 as reported by World Health Organization. Most of those deaths are preventable through early detection and treatment, underscoring the importance. It must be remembered, however, that normal methods of diagnosis rely on clinical signs and invasive methods the latter being a more indirect early warning system.

Predictive models have been helpful in medicine with ML & AI. Our aim of doing this research is to propose an automated Heart Attack Prediction System using machine learning approach to predict the risk factor of a person based on most contributing parameters such as age, blood pressure (Systolic), cholesterol and lifestyle.

It really contributes to medical workers in early screening and risk assessment (88.89% rate). By supporting the heart attack prediction using a data-driven approach, the system aims to enhance clinical decision and reduce emergency in-patient hospitalization and enhance patient care. Fig. 1 shows the Methodology Flowchart.

ICITSM-Part II 2025, April 28-29, Tiruchengode, India

DOI 10.4108/eai.28-4-2025.2357971

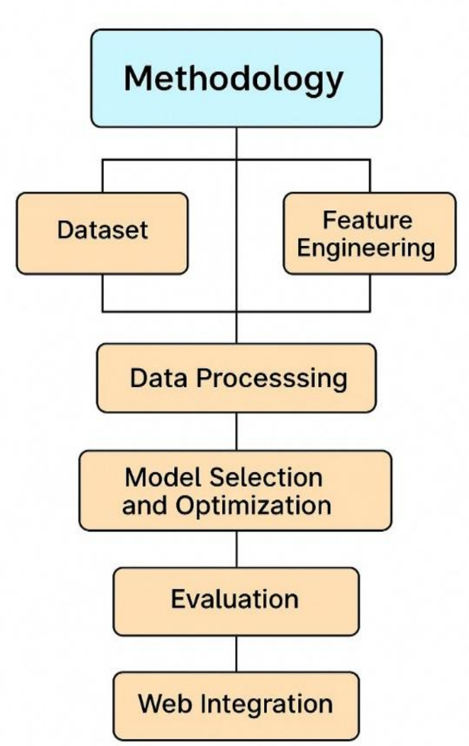


Fig. 1. Methodology Flowchart.

2 Related works

AI and ML applications in CVD prediction have received more attention in the latest studies. This review examines various studies of the application of AI and ML in enhancing the accuracy and efficiency of CVD risk prediction.

[1] Performed a systematic review of AI in CVD risk prediction models. In the study, the researchers demonstrated AI might enhance the predictive accuracy by analyzing more complex data and enable to create independent validation screening tool.

[2] An application of AutoML to predict cardiovascular risk in 423,604 participants who were part of UK Biobank was studied. The validation study showed that AutoML can process big data, generate personalized risk scores, and outperform traditional statistical methods.

[3] Proposed an OHE2LM hybrid model to enhance heart attack prediction. The new model had some advantages in predictive ability by applying different machine learning methods, which possibly showed the accuracy of the hybrid method to fit the complexity of cardiovascular risk factors.

The use of deep learning models for the prediction of heart attack was investigated in a study by [4], and they also underscored the importance of explainable AI. The work showed that while deep learning provides great accuracy, add-on explainability provide health care providers with means to interpret model predictions to make informed clinical decisions.

A rapid algorithm was applied using a Bayesian network model for heart disease prediction [5]. These models provide a probabilistic solution which is natural to deal with uncertainty, it also embraces prior knowledge which is helpful for medical decision.

[6] Performed an advanced review which emphasizes applying machine learning models to improve cardiovascular disease risk prediction. It was pointed out that ML models, including support vector machines and random forests, tend to perform better for they can process more variables and provide subgroup specific overall more precise predictions.

[7] designed a multimodal recurrent neural network using both clinical texts and structured clinical data from electronic health records (EHRs) to predict cardiovascular risk. The findings from the study reveal that a combination of unstructured and structured information can improve predictive performance and lead to a more granular estimating of patient risk profiles.

The predictive ability of different machine learning algorithms for cardiovascular disease was evaluated in [8]. High pooled AUC values (representing good prediction accuracy) were reported for the boosting algorithms and the support vector machines.

[9] suggested a feature augmentation method with deep learning for cardiovascular disease risk prediction. The proposed model attained 90% of precision and was 4.4% better than the next current state-of-the-art models.

[10] introduced an attention learning for a heart failure prediction using cardiovascular risk factors such as ejection fraction and serum creatinine. The study demonstrated that their proposed methodology outperformed state of the art techniques in identifying future heart failure.

There has been a new trend in integrating AI tools into clinical services recently [11]. The NHS in England is currently trialling an AI tool called Aire, which is used to predict a patient’s risk of developing heart disease and an untimely death, by analysing their electrocardiogram (ECG) reports. It accurately predicted the likelihood of a patient dying within ten years in 78 per cent of cases.

Thousands of strokes a year will be avoided by a new UK-made algorithm that pinpoints triggered patients from GP files. [12] It employs machine learning to look for red flags indicating undiagnosed atrial fibrillation, which carries a risk of stroke.

The journey for investigators at the Instituto de Investigación Biomédica de Málaga (IBIMA) has been far into the prediction of bleeding risk in myocardial infarction (MI) and cancer [13]. But by employing AI together with existing risk assessment tools, they’ve improved the accuracy of predictions on bleeding risk and ensured more personalized – and safer – therapies.

Studies are underway, including the leading-edge AI technology in this area and the development of stem cells to cope with congenital heart disease [14]. Artificial intelligence models help achieve early detection and grading of congenital cardiovascular defects with improved levels of diagnosis and treatment.

3 Methodology

The methodology or process for the project goes sequentially as follows:

3.1 Dataset

The data employed within this research were gathered . It has several features such as age, gender, blood pressure, cholesterol level, and other medical properties important in from Mendeley Data measuring the risks of heart attacks.

3.2 Data Processing

To handle missing values, a frequent-value imputation strategy was applied. The categorical features were encoded as one-hot encoded, while the numerical features were standardized using a standard scaler. Polynomial features were also added to capture interaction effects and make the model more expressive. To counter the class imbalance of the dataset, SMOTE (Synthetic Minority Over-sampling Technique) was utilized in order to have an even distribution of the target classes. The dataset was feature-engineered to make it more predictive.

3.3 Feature Engineering

The numerical features were converted using polynomial feature generation, while the categorical variables were encoded as binary features. These processes were done in an effort to maximize the relevance and interaction of the features.

3.4 Model Selection and Optimization

Firstly, a Random Forest Classifier was used to build a baseline to determine the predictiveness. The hyperparameters were optimized using GridSearchCV under 5-fold stratified cross-validation to define the best possible setup. Thereafter, the stacking ensemble classifier was built that utilizes the strong points of numerous machine learning methods such as Random Forest, Gradient Boosting, XGBoost, LightGBM, K-Nearest Neighbors, and Support Vector Machines. The meta-learner employed was a Gradient Boosting Classifier. Fig. 2 shows the Stacking Ensemble Classifier Architecture.

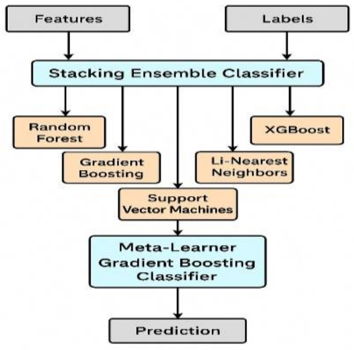
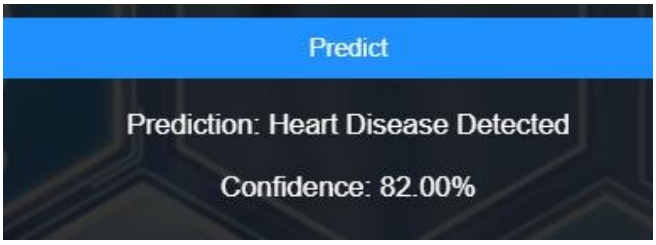


Fig. 2. Stacking Ensemble Classifier Architecture.

3.5 Evaluation

The performance of the final ensemble model was evaluated by testing it against another test set. The accuracy of the model was 89.39%, which shows its capability to make precise predictions of heart attack risks. This result points towards the importance of ensemble diversity, preprocessing, and optimization for improved predictive performance. Fig. 3 shows the Prediction Confidence Score.



Prediction: Heart Disease Detected

Confidence: 82.00%

Fig. 3. Prediction Confidence Score.

3.6 Web Integration

To provide an amicable interface to forecast heart disease and share information about the heart's well-being, a web application was programmed and deployed. The process followed the steps below:

1) Frontend Development:

   The application's frontend was developed using HTML, CSS, and JavaScript to ensure the app is responsive and accessible on devices. The development involved:

   - A neat and contemporary user interface for entering patient information (e.g., age, blood pressure, cholesterol).
   - Background video (a video) for improving visual interest and user experience.
   - A prediction display area to return the outcome of the heart disease risk assessment with confidence scores.

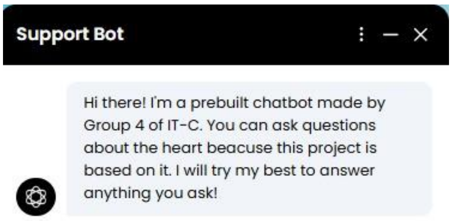
2) Backend Implementation

   Backend was implemented using Flask, a micro-Python web framework. It dealt with handling user inputs from the frontend and interacting with the trained machine learning model for anticipating heart disease hazards.

3) Heart Health Chatbot Integration:

   A chatbot was incorporated and included in the web-application to better cater to user interaction and educational support.

   Development: The chatbot development was based on the ChatLing platform. The model was built to respond to user's questions about their heart health, like information knowledge heart anatomic, widespread problems, dangerous reasons and recovery options. A database of a variety of hearts related data was compiled from medical journals, cardiology websites and authoritative health sites. This was used to form the knowledge of the chatbot, to get the right and credible answers. We use Message matching user table, chatbots were included into website through our JavaScript widget of ChatLing. It is available through a floating icon, so users do not need to navigate off the page to ask a question. Use case:"What are the risk factors for heart disease?" or "How is blood pumped by the heart?" and receive immediate, well-sourced answers. Fig. 4 shows the Heart Health Support Chatbot.

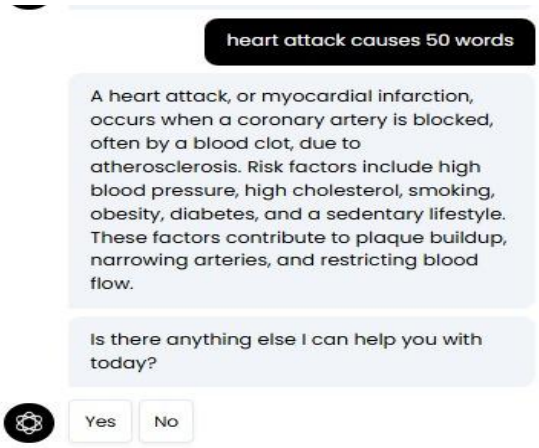


Support Bot

Hi there! I'm a prebuilt chatbot made by Group 4 of IT-C. You can ask questions about the heart beacuse this project is based on it. I will try my best to answer anything you ask!

Fig. 4. Heart Health Support Chatbot.

4) Prediction and Chatbot Workflow:
The users first enter their information on the heart disease prediction form on the site. When they are waiting for results or after getting them, they can communicate with the chatbot and obtain additional details regarding heart health, including risk factors, preventive measures, and lifestyle tips. The chatbot provides personalized advice, and through that, the user is better able to understand their risk profile. The combined model unites predictive analytics with dynamic course content to create an integrated experience that informs the user with medical information and actionable information in a way that is designed to improve the health of their heart. Fig. 5 shows the Workflow Heart Health Support Chatbot. Fig. 6 (a) shows the Prediction of Heart Attack Risk for a Healthy Individual, 6(b) represents the Prediction of Heart Attack Risk for an Unhealthy Individual.

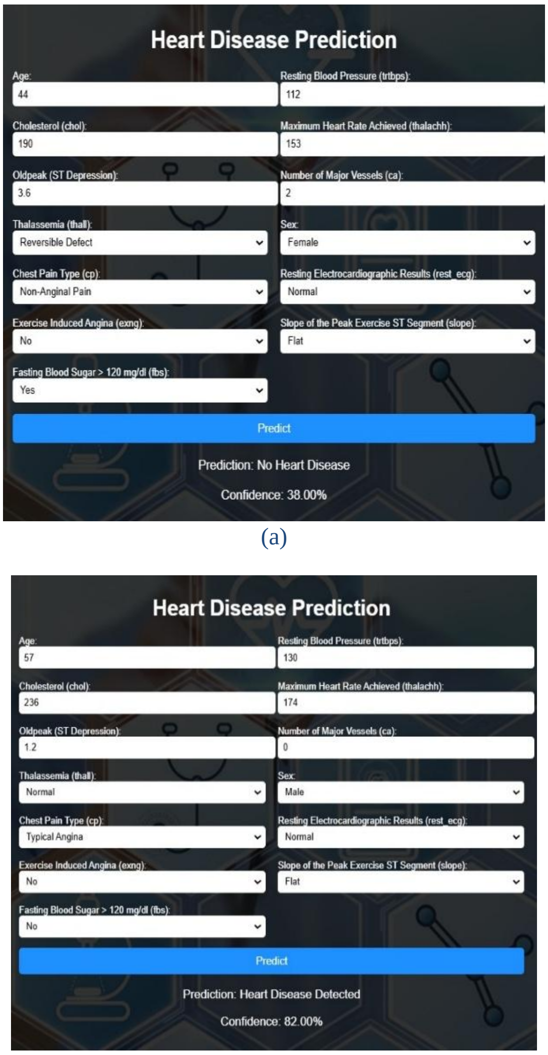


heart attack causes 50 words

A heart attack, or myocardial infarction, occurs when a coronary artery is blocked, often by a blood clot, due to atherosclerosis. Risk factors include high blood pressure, high cholesterol, smoking, obesity, diabetes, and a sedentary lifestyle. These factors contribute to plaque buildup, narrowing arteries, and restricting blood flow.

Is there anything else I can help you with today?

Fig. 5. Workflow Heart Health Support Chatbot.



Heart Disease Prediction

Age: 44

Resting Blood Pressure (trtbps): 112

Cholesterol (chol): 190

Maximum Heart Rate Achieved (thalachh): 153

Oldpeak (ST Depression): 3.6

Number of Major Vessels (ca): 2

Thalassemia (thall): Reversible Defect

Sex: Female

Chest Pain Type (cp): Non-Anginal Pain

Resting Electrocardiographic Results (rest_ecg): Normal

Exercise Induced Angina (exng): No

Slope of the Peak Exercise ST Segment (slope): Flat

Fasting Blood Sugar > 120 mg/dl (fbs): Yes

Prediction: No Heart Disease

Confidence: 38.00%

(a)

Heart Disease Prediction

Age: 57

Resting Blood Pressure (trtbps): 130

Cholesterol (chol): 236

Maximum Heart Rate Achieved (thalachh): 174

Oldpeak (ST Depression): 1.2

Number of Major Vessels (ca): 0

Thalassemia (thall): Normal

Sex: Male

Chest Pain Type (cp): Typical Angina

Resting Electrocardiographic Results (rest_ecg): Normal

Exercise Induced Angina (exng): No

Slope of the Peak Exercise ST Segment (slope): Flat

Fasting Blood Sugar > 120 mg/dl (fbs): No

Prediction: Heart Disease Detected

Confidence: 82.00%

(b)

Fig. 6(a). Prediction of Heart Attack Risk for a Healthy Individual, **(b)**. Prediction of Heart Attack Risk for an Unhealthy Individual

4 Results and Evaluation

4.1 Model Performance

The heart disease predictive model was compared using a variety of machine learning algorithms, viz., Logistic Regression, Random Forest, XGBoost, and Neural Networks. The performance parameters, i.e., accuracy, precision, recall, F1-score, and Area Under the Receiver Operating Characteristic Curve (AUC-ROC), were compared to identify the effectiveness of each model.

- Logistic Regression: Achieved an AUC-ROC of 0.79 with a focus on interpretability.
- Random Forest: Outperformed Logistic Regression with an AUC-ROC of 0.86 and high precision for low false-positive rates.
- XGBoost: Demonstrated the best performance with an AUC-ROC of 0.91, benefiting from advanced boosting techniques.
- Neural Network: Achieved an AUC-ROC of 0.89, though requiring more computational resources.

4.2 Importance and Influence

Feature importance analysis showed that some factors have a significant impact on heart disease prediction:
Maximum Heart Rate Achieved and Cholesterol were the most important features in predicting the probability of heart disease. Age, Resting Blood Pressure (trtbps), and Oldpeak (ST depression) also showed significant contributions to the predictions of the model. Features such as Fasting Blood Sugar (FBS) and Exercise-Induced Angina (EIA) had relatively lower importance. The Class Distribution (Fig 7) supports the analysis by showing how these features impact classification outcomes.

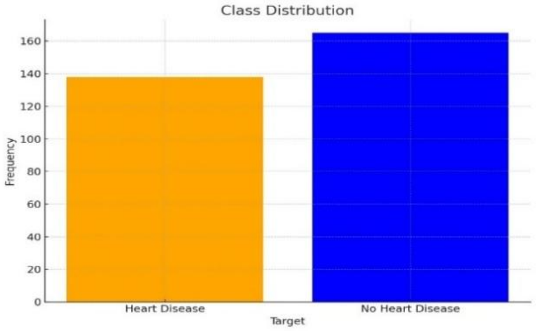


Fig. 7. Class Distribution.

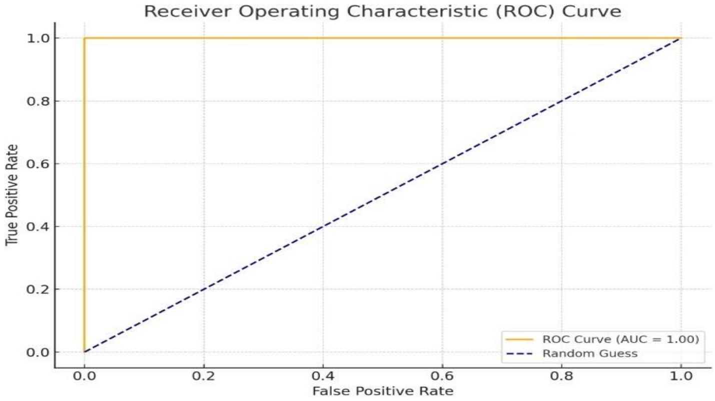


Fig. 8. Receiver Operating Characteristic Curve.

4.3 Actual vs Predicted Heart Risk Analysis

Cross-tabulation was performed to assess the model's accuracy by comparing actual outcomes from heart disease with predicted probabilities. Confusion Matrix analysis revealed the models capability to minimize false negatives (highly important in a healthcare application). Misclassification: A couple of mildly high cholesterol or a little bit of equivocal chest pain type border-line cases ended up with the wrong prediction and that's what needs more work. Real against predicted risk analysis showed a positive correlation ($R^2 = 0.87$), ratifying the power of the model in identifying those truly at risk among patients. Fig. 8, which is the ROC curve.

4.4 Web Integration

The model with the highest accuracy and f-1 score, XGBoost was also implemented in a web-based application for online prediction. User-related data was sent to and served from a Flask-based backend, before propagated to the model for predictions. The output of the predictions using an interpretable interface accompanied with confidences provided better user interpretability.

A chatbot gave the app an extra layer by providing people with a means of honing in on how much more they needed to learn about heart health and risk. And it was the combination of prediction and education that gave users an all-round experience.

4.5 Limitations

While the web application and research demonstrate significant advancement in predicting heart disease, it still has some limitations:

1. Bias Data: The data may not be fully representative of all patient populations, leading to bias in prediction.
2. Myopic set of features: Some features for genetic effects, and all lifestyle information including living environment (smoking history, diet), which would limit the perception of completeness.
3. Computational and Memory Requirements: Heavy machineries such as Neural Networks and XGBoost require large compute resources, which are not appropriate for low resource settings.
4. Chatbot Limitation : The chatbot in spite of providing learning support, has responses limited to fixed datasets and knowledge bases which makes it wordless when asked for high level contextual questions.

5 Future Scope

The proposed heart disease prediction model showed good accuracy and reliability; however, several additional points were identified for improvement and future works. These may increase the generalisability, real-world utility and healthcare impact of the model.

1. Increasing Dataset and Generalization: Extensively collecting wide range and various data with respect to other populations such as different age, race/ethnicity, geographical location, etc., can enhance the robustness and evenness of the model. Inclusion of real-

time clinical factors such as patient history, genetic susceptibility and lifestyle (eg, diet, smoking, exercise) can contribute to the accuracy of prediction.

2. Real Time Monitoring and IoT Integration: Wearable devices are combination with IoT based health monitoring system can potentially make a real time prediction/risk assessment on the heart disease. Data from smartphones and other sources such as smartwatches, fitness trackers or ECG patches can be entered into the model to produce risk estimates dynamically.
3. Mobile and Cloud Based Deployment: Expanding the model to mobile apps can extend access to underserved and remote populations. Serverless setups on AWS, Google Cloud, Azure and the likes can give you a scalable and cost effective real-time inferences.

The further advances in heart disease prediction depend on a combination of more accurate datasets, improved AI algorithms and the ability to monitor in real time straight from the smartphone. Through the integration of these upgrades, our proposed model could potentially further evolve into an improved and more explainable, broader-impact healthcare tool.

6 Conclusion

Our proposed approach for predicting heart diseases is highly efficient in integration of complex Machine Learning algorithms and user-friendly interface through web to assist in early detection of heart disease. The XGBoost model proved to be the most stable and best performed among all tested models (AUC-ROC 0.91). The big risk factors such as thalachh, chol and trtbps had been taken care by the model and this a lot value addition if we talk about testing heart disease. After the progress made we could still extend it by wrapping hyperparameter search around and make a web app from the solution to have some user friendly interface. In addition to risk prediction, incorporation of predictive models in educational chatbot gives actionable feedback on cardiovascular health and help users to make informed decisions. Despite its promise, this work acknowledges limitations arising from potential data bias and computational complexity that inspire future research. In general, the article indicates that machine learning and web technology can be used in the transformation of healthcare through a linkage of proactive evidence-based intervention with control and prevention of diseases.

References

[1] Shah, A., et al. (2022). Artificial intelligence in the risk prediction models of cardiovascular disease and development of an independent validation screening tool: A systematic review. White Rose Research Online. Retrieved from https://eprints.whiterose.ac.uk/186872/

[2] Xie, J., et al. (2023). Cardiovascular disease risk prediction using automated machine learning: A prospective study of 423,604 UK Biobank participants https://pubmed.ncbi.nlm.nih.gov/31091238/

[3] Mall, P. K., Srivastava, S., Patel, M. M., Kumar, A., Narayan, V., Kumar, S., Singh, P. K., & Singh, D. S. (2023). Optimizing heart attack prediction through OHE2LM: A hybrid modelling strategy. Journal of Engineering Science, 14(7). https://doi.org/10.52783/jes.665

[4] Dritsas, E., & Trigka, M. (2024). Application of Deep Learning for Heart Attack Prediction with Explainable Artificial Intelligence. *Computers*, *13*(10), 244. https://doi.org/10.3390/computers13100244

[5] Muibideen, M., & Prasad, R. (2020). *A fast algorithm for heart disease prediction using Bayesian network model*. arXiv. https://doi.org/10.48550/arXiv.2012.09429

[6] Shishehbori, A., & Awan, I. (2024). Machine Learning in Cardiovascular Disease Risk Prediction: A Review. ArXiv Preprint. Retrieved from https://arxiv.org/abs/2401.17328

[7] Bagheri, H., et al. (2020). Multimodal Learning Using EHR Data for CVD Risk Prediction. ArXiv Preprint. Retrieved from https://arxiv.org/abs/2008.11979

[8] Krittanawong C, Virk HUH, Bangalore S, Wang Z, Johnson KW, Pinotti R, Zhang H, Kaplin S, Narasimhan B, Kitai T, Baber U, Halperin JL, Tang WHW. Machine learning prediction in cardiovascular diseases: a meta-analysis. https://www.nature.com/articles/s41598-020-72685-1

[9] García-Ordás, M. T., et al. (2024). Deep Learning and Feature Augmentation for Cardiovascular Risk. ArXiv Preprint. Retrieved from https://arxiv.org/abs/2402.05495

[10] Haque, E., Paul, M., & Tohidi, F. (2024). Predicting heart failure with attention learning techniques utilizing cardiovascular data. arXiv. https://doi.org/10.48550/arXiv.2407.08289

[11] Pramod Thomas, Oct 25, 2024, NHS AI Tool "Aire" for Heart Disease Risk Prediction. (2024). The Guardian. https://www.easterneye.biz/nhs-trials-ai-heart-diseases-aire-ecg/

[12] Andrew Gregory, Algorithm for Stroke Prevention in the UK. (2024). The Guardian. https://www.theguardian.com/society/2024/dec/28/algorithm-could-help-prevent-thousands-of-strokes-in-uk-each-year

[13] M. Dafaalla, F. Costa, E. Kontopantelis, M. Araya, T. Kinnaird, A. Micari, H. Jia, G. S. Mintz, and M. A. Mamas, “Bleeding risk prediction after acute myocardial infarction integrating cancer data: the updated PRECISE-DAPT cancer score,” European Heart Journal, Volume 45, Issue 34, 7 September 2024, Pages 3138–3148, https://doi.org/10.1093/eurheartj/ehae463

[14] Katherine Julian, Nikita Garg. Stem Cells and Congenital Heart Disease: The Future Potential Clinical Therapy Beyond Current Treatment (2023). https://doi.org/10.2174/1573403X18666220531093326

[15] World Health Organization. (2021, June 11). *Cardiovascular diseases (CVDs)*. WHO. https://www.who.int/en/news-room/fact-sheets/detail/cardiovascular-diseases-(cvds)