Towards Proactive Heart Health: A Machine Learning-Powered Approach for Chronic Heart Failure Detection

Mawuli Agboklu
Joint School of Nanoscience and Nanoengineering
University of North Carolina at Greensboro.

Benjamin Lartey
Department of Electrical and Computer Engineering
North Carolina A&T State University, Greensboro

Frederick Adrah
Joint School of Nanoscience and Nanoengineering.
University of North Carolina at Greensboro.

Dennis LaJeunesse
Joint School of Nanoscience and Nanoengineering.
University of North Carolina at Greensboro.

ABSTRACT
Myocardial infarction, more commonly known as "heart attack" is one of the most dangerous diseases worldwide. Timely detection and intervention are crucial for saving the lives of patients and reducing mortality rates. Beside traditional clinical interventions, machine learning (ML) techniques have garnered considerable attention for their potential in aiding the early detection of heart disease in recent years. In this study, we will use ML algorithms such as Random Forum (RF), Support Vector Machine (SVM), Extreme Gradient Boosting (XGBoost) and Logistic regression (LR) to develop models to predict the possibility of chronic heart failure in patients hospitalized with myocardial infarction 72 hours after their hospitalization. Varied optimization techniques were applied to these models to improve their predictive outcomes. The models were evaluated using metrics such as accuracy, precision, recall, f1, mcc and confusion matrix and compared against each other to determine which of them generated better results. The XGBoost algorithm demonstrated superior performance compared to the other models. The dataset was collected from UCI machine learning repository with the database containing 1700 patient records and 111 input features.

General Terms
Myocardial infarction, Machine Learning, Heart disease, Algorithms

Keywords
Logistic regression, Support Vector Machine

Abbreviations: True Positive (TP), True Negative (TN), False Positive (FP), False Negative (FN).

1. INTRODUCTION
According to the World Health Organization, cardiovascular disease claims over 17.9 million lives annually globally. This high mortality rate is particularly prevalent in low and middle-income nations [1]. Cardiovascular disease stands as a foremost contributor to global mortality rates [2]. Forecasting such ailments poses a significant challenge within clinical data analytics [3]. According to the Centers for Disease Control and Prevention (CDC), one person dies every 33 seconds from heart disease, and it costed the United States approximately $239.9 billion each year from 2018 to 2019. This figure encompasses the expenses related to healthcare services, medications and lost productivity resulting from fatalities [4].

However, Machine learning (ML) has emerged as an effective tool, helping in decision-making and predictive analytics amidst the large quantity of data generated by the healthcare industry [5] and/or ML repositories [6]. Identifying heart disease poses challenges due to numerous contributing risk factors; including diabetes, high blood pressure, elevated cholesterol levels and abnormal pulse rates, among others [6][7]. Physicians aiming to effectively prevent major cardiovascular events must recognize that cardiovascular diseases, such as myocardial infarction is influenced by various interconnected factors [8][9]. They should inquire about specific, readily identifiable risk factors that elevate the likelihood of such events in patients [9].

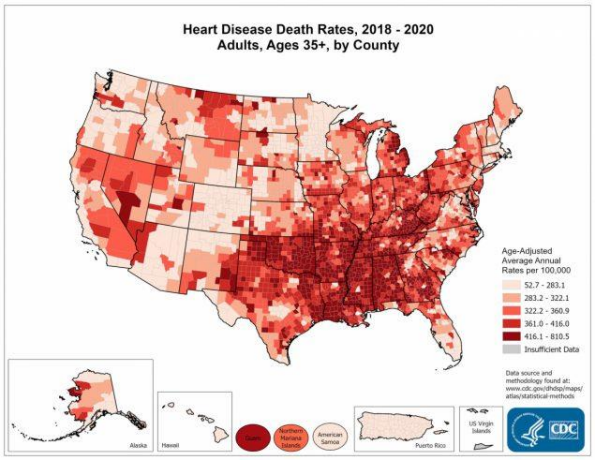

Figure 1: Heart disease death rates in the United States by county, 2018-2020 (CDC – heart disease fact sheets, https://www.cdc.gov/heartdisease/facts.htm) [4]

Diagnosing heart disease is critical, as it can be life-threatening, and timely detection of this condition can mitigate its severity and ultimately save lives [10][11][12]. Recently, data mining and neural network methodologies have been harnessed to assess the severity of heart disease in human population [1] [13].

This research paper focuses on an approach to detect heart attack within 72 hours after a patient has been hospitalized with cardiovascular disease. This research paper is particularly significant due to the fact that approximately 1 in 5 heart attacks

are silent—meaning the damage occurs without the individual's awareness [4]. Therefore, it is crucial to utilize machine learning to predict heart attacks

2. RELATED WORK

With the emergence of AI and machine learning in modern times, numerous studies have enriched the field by utilizing machine learning, artificial intelligence and deep learning methods to predict the occurrence of heart disease.

In one study within this domain, ML algorithms were trained using data extracted from the Cleveland heart disease dataset. Researchers utilized a variety of ML algorithms, encompassing Decision Tree (DT), Discriminant Analysis (DA), Logistic Regression (LR), Naïve Bayes (NB), Support Vector Machines (SVM), K-Nearest Neighbors (KNN), and Ensemble techniques [14]. The effectiveness of these algorithms was evaluated through 10-fold cross-validation, both with and without the application of Principal Component Analysis (PCA). Logistic Regression achieved the highest accuracy of 85.8% with PCA, while retaining 9 components. Additionally, Ensemble classifiers demonstrated an accuracy of 83.8% with PCA, maintaining 10 components [14].

Previous research in this field employed data exploration and mining techniques to uncover hidden patterns using the Python programming language. Machine learning algorithms such as the logistic regression, decision tree classifier and Gaussian Naïve Bayes models (GNB) were developed to forecast the occurrence of heart diseases in patients. Both Logistic Regression and GNB demonstrated the highest predictive accuracy, scoring 82.75% [15]. This method also utilized data from the UCI ML repository with a dataset containing 14 attributes sampled from 303 patients. Expanding upon this approach from a different perspective, other researchers employed six machine learning algorithms (random forest, K-nearest neighbor, logistic regression, Naïve Bayes, gradient boosting, and AdaBoost classifier) using datasets obtained from the Cleveland and IEEE Dataport. In this methodology, multiple ML models were integrated using ensemble techniques to generate a collective outcome anticipated to exhibit higher accuracy than any individual algorithm, achieving accuracies as high as 95% [16].

With these methods, it demonstrates the potential of ML as a potent tool for heart attack and myocardial infarction prediction. The increasing incidence of heart attacks among younger individuals, coupled with the financial burdens and limitations of existing medical tools, underscores the urgency for innovative solutions in cardiovascular health [17]. In response, computerized systems have emerged as promising alternatives to traditional methods, providing quicker and more efficient heart disease risk predictions [10][11]. Leveraging machine learning holds the promise of enhancing early detection and diagnosis, thereby offering a more effective approach to addressing the challenges posed by heart disease.

Another method employed in predicting heart disease utilizes deep learning, specifically feature extraction from a Convolutional Neural Network (CNN). This approach involved designing an ensemble model, where the CNN model was employed to augment the feature set for training linear models. These linear models, including the stochastic gradient descent classifier, logistic regression, and support vector machine, were integrated into a soft-voting based ensemble model [18]. In this approach, four distinct datasets and their outcomes were contrasted with recent methods employed in heart disease research. The findings demonstrate the superior performance of the proposed model, achieving an accuracy of 0.93 and scores of 0.92 each for precision, recall and F1 score. These results not only affirm the effectiveness of the proposed methodology but also underscore the ability of the ensemble model to generalize across multiple datasets [18].

A recurring theme in the research conducted in this field has unveiled a predominant focus on prediction accuracy [19][20], suggesting that the majority of researchers have strived to optimize ML-based methods to enhance predictive accuracy.

This research acknowledges existing studies in this domain and delves further into an ML-driven methodology aimed at detecting the likelihood of chronic heart failure in patients admitted with myocardial infarction within 72 hours of hospitalization, employing ML models.

3. METHODOLOGY

This section delves into the methodology process used in building and evaluating the machine learning models used in this prediction analysis.

Dataset Description

The dataset used in this predictive analysis is entitled Myocardial Infarction Complications and was obtained from the UCI machine learning repository. It is a real multivariate dataset which contained 122 features and 1700 samples.

3.1 Data Preprocessing

The dataset had 122 hot – encoded variables including possible complications. Each of the 1700 instances had unique IDs which eliminated the possibilities of duplicate rows however, there were missing values from the dataset hence the data preprocessing process was commenced by fixing the missing values using K – Nearest Neighbor (KNN) Imputation. Feature importance analysis of individual models ie. Random forest, SVM, XGBoost and Logistic Regression was performed to eliminate noise, control overfitting and to generally improve performance of the models.

3.2 Background of Prediction Models

This section details the different ML algorithms used in accomplishing the task of this study and their comparative advantages in solving the task at hand. The algorithms deployed were Random Forest (RF), Support Vector Machine (SVM), Extreme Gradient Boost (XGB) and Logistic Regression (LR).

3.2.1 Random Forest

Random forest is an ensemble (i.e., a collection) of unpruned decision trees. Random forests are often used when we have very large training datasets and a very large number of input variables (hundreds or even thousands of input variables). A random forest model is a classifier that consists of many decision trees and outputs the class that is the mode of the class output by individual trees [21]. Random Forest algorithm is appropriate for high dimensional data modeling because it is able to handle continuous, categorical, and binary data. It is robust to overfitting due to its aggregation of predictions from multiple decision trees and has high accuracies. These characteristics make it one of the favorable algorithms to deploy for classification tasks in medical diagnosis. In applying random forest to this research, hyperparameter tuning

optimization function with random search was deployed to effectively customize and optimize the efficiency of the model.

3.2.2 Support Vector Machine (SVM)

Support Vector Machine is a machine learning algorithm that learns data points by example to allow it to classify new samples correctly into their respective sets. It uses the principles of a hyperplane, maximum – margin hyperplane, soft margin and kernel functions in its implementation [22]. There are different implementations of SVMs i.e. linear and non-linear implementations with a wide array of different kernels to accompany it [23]. SVMs are mostly applied in classification tasks but can also be applied to regression tasks. It is preferred due to its effectiveness with high dimensional datasets, versatility due to the different kernel functions and relative computational (memory) efficiency. In deploying SVM for this research, the linear kernel function was selected because of the nature of the dataset and the objective of the task.

3.2.3 Extreme Gradient Boosting (XGBOOST)

XGBoost is a variant of the Gradient Boosting Machine algorithm. This algorithm creates new models to correct the error of previous models and aggregate them to make a prediction. Essentially, it trains weak leaners and combines them into a strong predictive model[24]. Eventhough XGBoost is considered most effective with classification tasks[25], it can also be applied to regression tasks[26] and the algorithm can be optimized with different optimization techniques in both situations. Xgboost is scalable and has high predictive performance. For this research, no optimization technique was deployed to the XGB classifier.

3.2.4 Logistic Regression (LR)

Logistic regression algorithm is a classification algorithm used for binary tasks. It is used to formulate predictive models where the expected outcome can only be one of two options such as a disease state[27]. Logistic regression is widely preferred in medical classification tasks due to its simplicity and easy to interpretation. Although LR predicts probabilities of the default class, the inputs are transformed using the logistic function[28]. Different optimization techniques such as Maximum – likelihood Estimation and Regularization can be applied to improve the algorithm depending on the size of the dataset. For this research, no optimization techniques were applied.

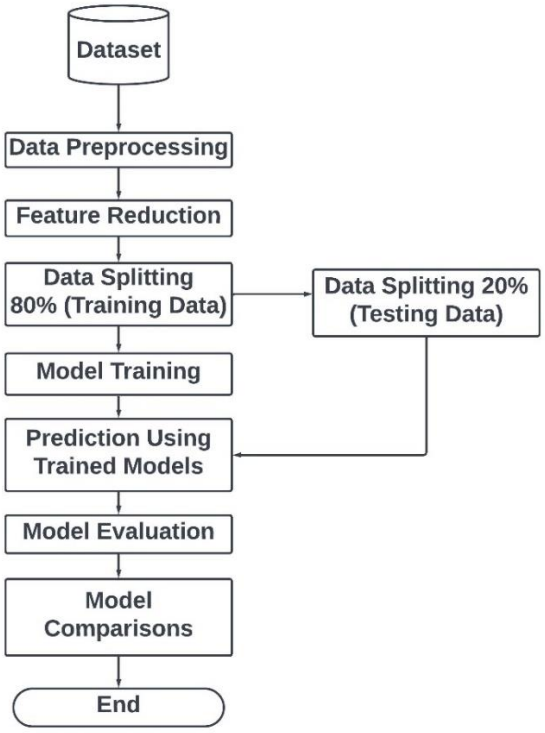

Figure 2. A Workflow diagram of the methodology adopted in this paper.

3.3 Evaluation Metrics accuracy score is an evaluation metric which measures the accuracy of a model in making prediction.

Accuracy Score: The accuracy score is an evaluation metric which measures the accuracy of a model in making predictions. It serves as a baseline for measuring the accuracy of models. It is calculated mathematically as the number of correct predictions out of total number of predictions.

$$Accuracy = \frac{\mathrm{TP} + \mathrm{TN}}{\mathrm{TP} + \mathrm{TN} + \mathrm{FP} + \mathrm{FN}}$$

Precision Score: Precision Score is an evaluation metric which measures the frequency of positive predictions of a model. It is the ratio of the true positive predictions to the total number of predicted positives. It is calculated mathematically as true positives out of the total number of positive instances predicted by the model.

$$Precision = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FP}}$$

Recall Score: Recall score is a metric that measures the correctness of a model in predicting positive instances from all the actual positive instances in the dataset. It can be referred to as the sensitivity of a model to the task at hand. It is calculated mathematically by dividing the true positives by the number of all positive instances.

$$Recall = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}}$$

F1 Score: F1 score is the harmonic mean of the precision and recall score. It measures the general performance of the model. It is mathematically calculated as multiplying 2 by the

product of precision and recall and diving it by the sum of precision and recall.

$$F1 = \frac{2 \times (Precision \times Recall)}{(Precision + Recall)}$$

MCC Score: Matthews Correlation Coefficient is an evaluation metric for binary classification tasks. It considers all components of the confusion matrix and produces a result between +1 to -1 to indicate the performance of the model where +1 shows a good performance 0 indicates a random performance and -1 indicates a poor performance.

$$MCC = \frac{(TP \times TN - FP \times FN)}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}}$$

Confusion Matrix: Confusion Matrix is an evaluation matrix which shows the prediction of a model versus the actual values in the dataset. It gives a numerical interpretation to the performance of the model. Confusion Matrix has four (4) components i.e. True Positive, True Negative, False Positive and False Negative. Depending on the classification task and its impact, weight is given to any of the four components.

4. RESULTS

	Random Forest	SVM	XGBoost	LR
Accuracy	0.80 (80%)	0.76 (76%)	0.82 (82%)	0.79 (79%)
Precision	0.88 (88%)	0.375 (37.5%)	0.74 (74%)	0.7 (70%)
Recall	0.18 (18%)	0.0375 (3.75%)	0.35 (35%)	0.175 (17.5%)
F1 Score	0.31 (31%)	0.068 (6.8%)	0.47 (47%)	0.28 (28%)
MCC	0.34 (34%).	0.051 (5.1%)	0.42 (42%)	0.27 (27%)

Fig 3. Presents a tabular representation of the performance of various models across different evaluation metrics.

The results section presents the performance outcomes of various models evaluated using multiple metrics. Figure 2 offers a tabular summary of these performances, displayed in both absolute figures and percentage terms. Figure 3 provides a graphical comparison, plotting the models against each other for a clear comparative analysis. Figures 5 through 8 illustrate the individual confusion matrices for each model, offering a detailed look at their classification accuracies and errors.

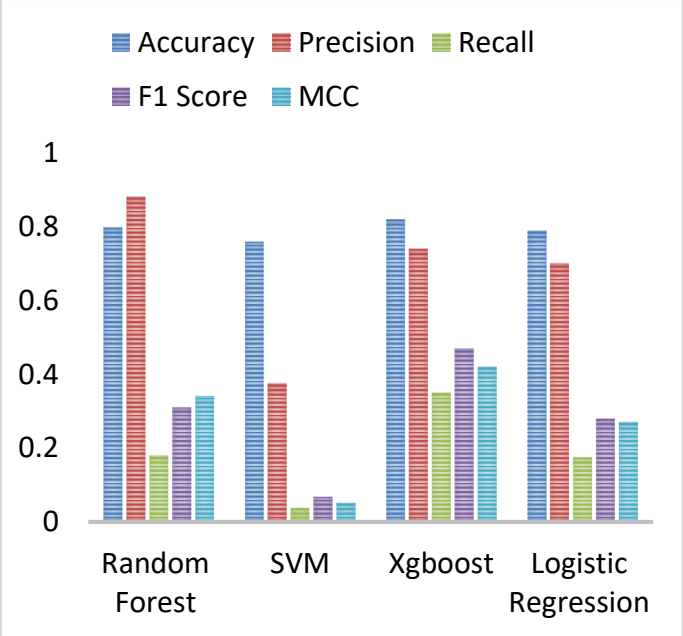

Figure 4 provides a graphical representation of the models, plotted against each other to compare their performance across various evaluation metrics.

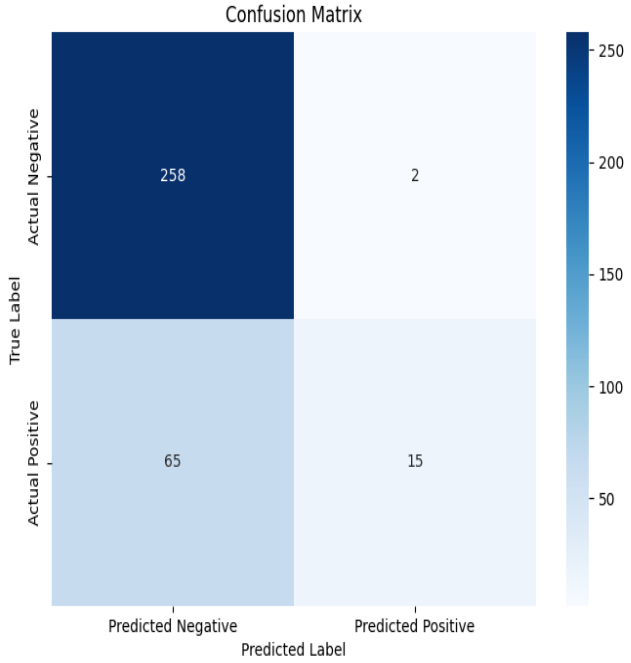

Figure 5. Random Forest Confusion Matrix

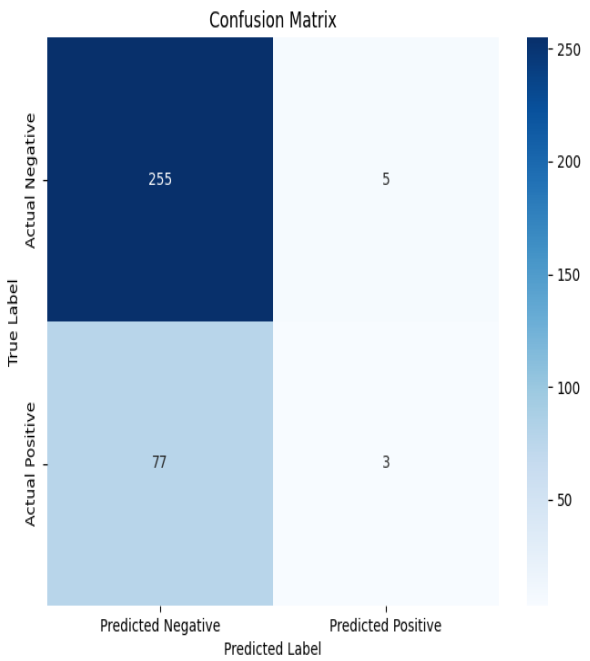

Figure 7. SVM Confusion Matrix

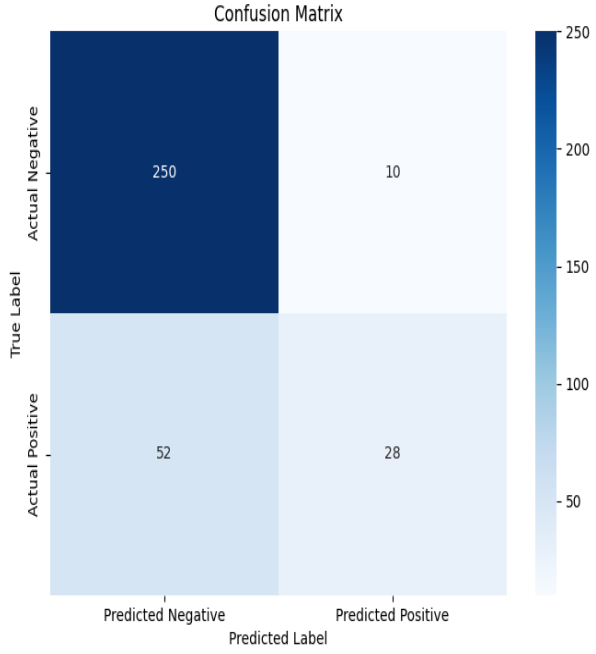

Figure 6. XGBoost Confusion Matrix

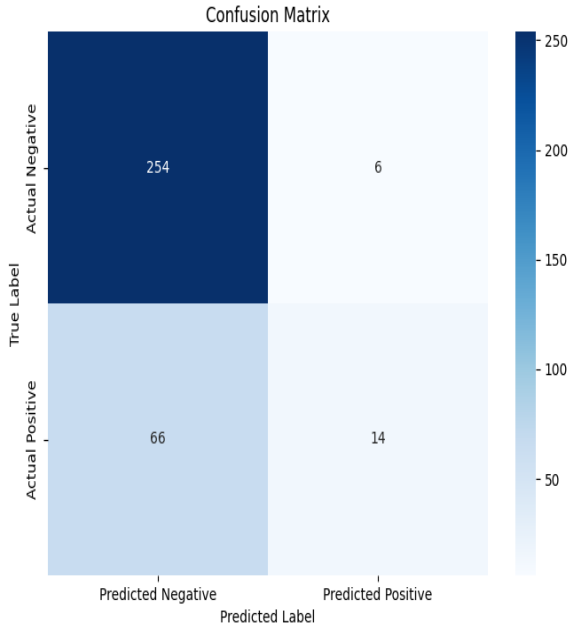

Figure 8. Logistic Regression Confusion Matrix

4. DISCUSSION

From the results, XGBoost algorithm exhibited the best predictive behavior with an accuracy score of 82% followed by Random Forest with 80%, LR with 79% and SVM with 76%. The trend continued throughout the other evaluation categories except precision where random forest outperformed XGBoost. Nonetheless overall, the results suggest XGBoost as the most efficient and better model to predict chronic heart failure.

6. CONCLUSION

Heart health is a global health concern, and it is important to develop innovative techniques and systems which augment the service delivery of health practitioners towards heart health issues. Machine Learning provides a robust platform to unravel and analyze datasets, and suggesting models which are efficient in predicting biological instances accurately. Predicting the possibility of chronic heart failure in patients with any form of myocardial infarction 72 hours after their hospitalization is a groundbreaking contribution towards heart healthcare. The models developed in this research can form the basis for further research work, driving the development of more advanced and accurate predictive tools. By leveraging larger and more diverse datasets, these models can be refined to improve their predictive accuracy and applicability across different populations and settings. The future scope of this project can include integrating these predictive models with Electronic Health Records (EHR) systems to facilitate real-time monitoring and early detection of heart failure risk in hospitalized patients, automated alerts and decision support systems can also be developed based on this research to assist healthcare providers in making timely and informed interventions. Finally, these predictive models can inform public health strategies and policies aimed at reducing the burden of heart disease, ultimately improving patient outcomes, and reducing healthcare costs.

7. REFERENCES

[1] D. Shah, S. Patel, and S. K. Bharti, “Heart Disease Prediction using Machine Learning Techniques,” *SN COMPUT. SCI.*, vol. 1, no. 6, p. 345, Oct. 2020, doi: 10.1007/s42979-020-00365-y.

[2] B. R. Smith and E. R. Edelman, "Nanomedicines for cardiovascular disease," *Nature Cardiovascular Research*, vol. 2, no. 4, pp. 351–367, 2023.

[3] J. Soni, U. Ansari, D. Sharma, and S. Soni, "Predictive data mining for medical diagnosis: An overview of heart disease prediction," *International Journal of Computer Applications*, vol. 17, no. 8, pp. 43–48, 2011.

[4] CDC, "Heart Disease Facts | cdc.gov," Centers for Disease Control and Prevention. Accessed: May 07, 2024. [Online]. Available: https://www.cdc.gov/heartdisease/facts.htm

[5] I. E. Agbehadji, B. O. Awuzie, A. B. Ngowi, and R. C. Millham, "Review of big data analytics, artificial intelligence and nature-inspired computing models towards accurate detection of COVID-19 pandemic cases and contact tracing," *International journal of environmental research and public health*, vol. 17, no. 15, p. 5330, 2020.

[6] S. Mohan, C. Thirumalai, and G. Srivastava, "Effective Heart Disease Prediction Using Hybrid Machine Learning Techniques," *IEEE Access*, vol. 7, pp. 81542–81554, 2019, doi: 10.1109/ACCESS.2019.2923707.

[7] S. C. Smith Jr, "Multiple risk factors for cardiovascular disease and diabetes mellitus," *The American journal of medicine*, vol. 120, no. 3, pp. S3–S11, 2007.

[8] F. Jyotsna *et al.*, "Exploring the complex connection between diabetes and cardiovascular disease: analyzing approaches to mitigate cardiovascular risk in patients with diabetes," *Cureus*, vol. 15, no. 8, 2023.

[9] T. A. Haffey, "How To Avoid A Heart Attack: Putting It All Together," *Journal of Osteopathic Medicine*, vol. 109, no. s51, pp. 14–20, May 2009, doi: 10.7556/jaoa.2009.20004.

[10] P. Rani *et al.*, "An Extensive Review of Machine Learning and Deep Learning Techniques on Heart Disease Classification and Prediction," *Arch Computat Methods Eng*, Mar. 2024, doi: 10.1007/s11831-024-10075-w.

[11] Z. Keshavarz-Motamed, "A diagnostic, monitoring, and predictive tool for patients with complex valvular, vascular and ventricular diseases," *Scientific Reports*, vol. 10, no. 1, p. 6905, 2020.

[12] J. Mistry and A. Ganesh, "An Analysis of IoT-Based Solutions for Congenital Heart Disease Monitoring and Prevention," *Journal of Xidian University*, vol. 17, no. 7, pp. 325–334, 2023.

[13] D. Yewale and S. Vijayaragavan, "Data-Driven Insights: A Genetic Algorithm Feature Optimization Approach to Heart Disease Prediction," presented at the 2024 International Conference on Intelligent and Innovative Technologies in Computing, Electrical and Electronics (IITCEE), IEEE, 2024, pp. 1–6.

[14] K. V. V. Reddy, I. Elamvazuthi, A. A. Aziz, S. Paramasivam, and H. N. Chua, "Heart Disease Risk Prediction using Machine Learning with Principal Component Analysis," in *2020 8th International Conference on Intelligent and Advanced Systems (ICIAS)*, Jul. 2021, pp. 1–6. doi: 10.1109/ICIAS49414.2021.9642676.

[15] D. Ananey-Obiri and E. Sarku, "Predicting the presence of heart diseases using comparative data mining and machine learning algorithms," *International Journal of Computer Applications*, vol. 176, no. 11, pp. 17–21, 2020.

[16] N. Chandrasekhar and S. Peddakrishna, "Enhancing Heart Disease Prediction Accuracy through Machine Learning Techniques and Optimization," *Processes*, vol. 11, no. 4, Art. no. 4, Apr. 2023, doi: 10.3390/pr11041210.

[17] P. Ponikowski *et al.*, "Heart failure: preventing disease and death worldwide," *ESC heart failure*, vol. 1, no. 1, pp. 4–25, 2014.

[18] F. Rustam, A. Ishaq, K. Munir, M. Almutairi, N. Aslam, and I. Ashraf, "Incorporating CNN Features for Optimizing Performance of Ensemble Classifier for Cardiovascular Disease Prediction," *Diagnostics*, vol. 12, no. 6, Art. no. 6, Jun. 2022, doi: 10.3390/diagnostics12061474.

[19] S. H. B. Hani and M. M. Ahmad, "Machine-learning algorithms for ischemic heart disease prediction: a systematic review," *Current Cardiology Reviews*, vol. 19, no. 1, 2023.

[20] S. M. S. Shah, S. Batool, I. Khan, M. U. Ashraf, S. H. Abbas, and S. A. Hussain, "Feature extraction through parallel probabilistic principal component analysis for heart disease diagnosis," *Physica A: Statistical Mechanics and its Applications*, vol. 482, pp. 796–807, 2017.

[21] D. Gao, Y.-X. Zhang, and Y.-H. Zhao, "Random forest algorithm for classification of multiwavelength data," *Research in Astronomy and Astrophysics*, vol. 9, no. 2, p. 220, 2009.

[22] W. S. Noble, "What is a support vector machine?," *Nature biotechnology*, vol. 24, no. 12, pp. 1565–1567, 2006.

[23] S. Suthaharan and S. Suthaharan, "Support vector machine," *Machine learning models and algorithms for big data classification: thinking with examples for effective learning*, pp. 207–235, 2016.

[24] A. Asselman, M. Khaldi, and S. Aammou, "Enhancing the prediction of student performance based on the machine learning XGBoost algorithm," *Interactive Learning Environments*, vol. 31, no. 6, pp. 3360–3379, 2023.

[25] H. Tan, "Machine learning algorithm for classification," in *Journal of Physics: Conference Series*, IOP Publishing, 2021, p. 012016.

[26] B. Lartey, A. Homaifar, A. Girma, A. Karimoddini, and D. Opoku, "XGBoost: a tree-based approach for traffic volume prediction," *2021 IEEE International Conference on Systems, Man, and Cybernetics (SMC)*, pp. 1280–1286.

[27] A. B. Musa, "Comparative study on classification performance between support vector machine and logistic regression," *International Journal of Machine Learning and Cybernetics*, vol. 4, pp. 13–24, 2013.

[28] E. A. DiGangi and J. T. Hefner, "Ancestry estimation," *Research methods in human skeletal biology*, pp. 117–149, 2013.