RESEARCH ARTICLE OPEN ACCESS

CARDIO RISK ASSESSMENT USING MACHINE LEARNING

Mr. P. Veeresh Kumar[1],

Assistant Professor, Department of IT,

KKR & KSR Institute Of Technology and Sciences, Vinjanampadu, Guntur Dt., Andhra Pradesh.

Alokam Pavani[2], Garikapati Bhavya Sai[3], Jajula Maheswari[4], Annapureddy Shirisha[5]

[2,3,4,5] UG Students, Department of IT,

KKR & KSR Institute Of Technology and Sciences, Vinjanampadu, Guntur Dt., Andhra Pradesh.

[1] veeresh.pinnamraju@gmail.com ,[2]pavanialokam2002@gmail.com,

[3]garikapatibhavya03@gmail.com, [4]maheswarijajula123@gmail.com,[5] shiriannapureddy@gmail.com

--*************************----------------------------------

Abstract:

Cardiovascular disease (CVD) remains a global health threat, demanding improved methods for early detection and risk stratification. This project explores the potential of machine learning for predicting the probability of heart disease in individuals. We conduct a comparative analysis of various machine learning algorithms, including Random Forest, Support Vector Machines, and Logistic Regression, to identify the model with the highest accuracy in predicting CVD. Our findings reveal that the Random Forest algorithm outperforms other contenders, demonstrating superior accuracy in assessing an individual's risk of developing heart disease. This model utilizes a person's characteristics and features, such as age, blood pressure, cholesterol levels, and family history, to generate a personalized probability score. By leveraging the power of Random Forest, this project proposes a novel approach for CVD risk assessment. This method holds promise for improving the identification of high-risk individuals, enabling clinicians to implement targeted preventative strategies and potentially reduce the burden of CVD.

Keywords — **cardiovascular disease (CVD), Machine Learning, Random Forest, Risk Assessment, Heart Disease Prediction, Feature Selection.**

--*************************----------------------------------

1. INTRODUCTION

Cardiovascular disease (CVD), encompassing conditions like heart attacks and strokes, reigns as the leading cause of death worldwide. This grim statistic underscores the urgent need for better preventative measures and earlier detection methods. While traditional risk assessment tools play a vital role, they often rely on a limited set of factors, potentially overlooking the intricate interplay between various risk indicators.

This project embarks on a journey to explore the transformative potential of machine learning in the realm of CVD risk assessment. Our objective is to develop a robust model capable of predicting an individual's susceptibility to heart disease with greater accuracy. This approach hinges on the power of machine learning algorithms to analyze a comprehensive set of personal characteristics and features. By incorporating a broader range of data points, we aim to achieve a more personalized and effective risk stratification process.

To achieve this goal, the project will delve into a comparative analysis of various machine learning algorithms. We will evaluate their performance in predicting

the likelihood of CVD. Our hypothesis centers on the Random Forest algorithm, renowned for its prowess in handling complex datasets and uncovering non-linear relationships. We anticipate that Random Forest will exhibit superior accuracy in this task compared to other contenders. The successful development of this model holds immense promise. It has the potential to significantly improve the identification of individuals at high risk of developing CVD. Armed with this information, clinicians can implement targeted preventative strategies, potentially leading to a substantial reduction in the global burden of CVD.This project unfolds in several key stages. First, we will meticulously curate a comprehensive dataset encompassing a wide range of relevant features associated with CVD risk. Next, we will embark on the process of data pre-processing, ensuring the data is clean, consistent, and ready for analysis. This vital step involves handling missing values, identifying and addressing outliers, and potentially performing feature scaling.

Following data preparation, we will delve into the exciting world of machine learning algorithms. We will explore and implement various algorithms, including Random Forest, Support Vector Machines, and Logistic Regression. Each algorithm will be meticulously trained on the prepared dataset, allowing it to learn the intricate relationships between features and the presence or absence of CVD.

Once the training phase is complete, we will rigorously evaluate the performance of each model. This assessment will involve employing metrics like accuracy, precision, recall, and F1-score to determine the efficacy of each algorithm in predicting CVD. Through this comparative analysis, we will identify the model that demonstrates the highest accuracy in predicting the likelihood of heart disease.

The culmination of this project will be the deployment of the chosen model, most likely Random Forest based on our hypothesis. This model will be equipped to generate a personalized probability score for an individual, indicating their susceptibility to developing CVD. This information can be a powerful tool for clinicians, enabling them to tailor preventative strategies for each patient.

2. LITERATURE SURVEY

Purushottam et al. [1] proposed a system using decision trees and hill climbing for heart disease prediction. They employed the Cleveland dataset and preprocessed it before applying classification algorithms. Knowledge extraction utilized KEEL, an open-source data mining tool, to handle missing values. The decision tree followed a top-down approach, selecting nodes based on hill climbing and test conditions at each level. Their system achieved an accuracy of approximately 86.7%.

Santhana Krishnan et al. [2] investigated decision trees and Naive Bayes for heart disease prediction on the Cleveland dataset. Their decision tree algorithm built a tree based on conditions leading to True/False decisions. Unlike algorithms like SVM and KNN that rely on vertical or horizontal splits, decision trees leverage a tree-like structure with root nodes, leaves, and branches based on decisions made at each node. This approach also helps understand the importance of attributes within the dataset. They split the data into 70% training and 30% testing, achieving an accuracy of 91% with the decision tree. Additionally, they employed Naive Bayes, a classifier suitable for complex, non-linear, and dependent data like the heart disease dataset, which yielded an accuracy of 87%.

Sonam Nikhar et al. [3] focused on Naive Bayes and decision tree classifiers for heart disease prediction. Their analysis explored the effectiveness of a data mining

strategy on the same dataset, concluding that decision trees exhibited higher accuracy than Naive Bayes.

Gavhane et al. [4] proposed using a neural network called Multi-Layer Perceptron (MLP) for heart disease prediction. MLPs consist of an input layer, an output layer, and hidden layers in between. Each input node connects to the output layer through hidden layers with assigned weights. Additionally, a bias term can be incorporated to adjust the connection strength.

Golande et al. [5] investigated the use of data mining techniques to support doctors in heart disease diagnosis. These techniques included k-Nearest Neighbors (kNN), Decision Trees, and Naive Bayes. Additionally, they explored unique characterization-based strategies like support vector machines (SVM) and neural networks.

Rao et al. [6] highlighted the challenge of identifying heart disease due to multiple contributing factors. Their study explored various neural networks and data mining techniques to assess the severity of heart disease in individuals.

Kishore et al. [7] proposed a heart attack prediction system using deep learning techniques. Their model incorporates Recurrent Neural Networks (RNNs) to analyze patient data and predict the likelihood of heart-related infections. This study establishes a strong foundation for developing other heart attack prediction models.

Mohan et al. [8] focused on enhancing the accuracy of cardiovascular disease prediction. They employed a combination of algorithms, including KNN, Logistic Regression (LR), SVM, and Neural Networks (NN). Their proposed Hybrid Random Forest with Linear Model (HRFLM) achieved an accuracy level of 88.7%.

Repaka et al. [9] evaluated the performance of two classification models for heart disease prediction. Their analysis compared these models to previous work and demonstrated improved accuracy in risk prediction.

Chauhan et al. [10] proposed a method for heart disease prediction using Evolutionary Rule Learning. This approach leverages electronic patient records to automate data retrieval and reduce manual tasks. The study employs frequent pattern growth association mining to identify strong associations within patient data, leading to more accurate predictions.

3. PROPOSED METHOLODOGY

The proposed method for our Cardio risk assessment consists of several key components:

➢ Data Acquisition and Pre-processing:

Data Source: We will acquire a comprehensive dataset containing relevant features associated with CVD risk. Potential sources include electronic health records, public health databases, or curated datasets.

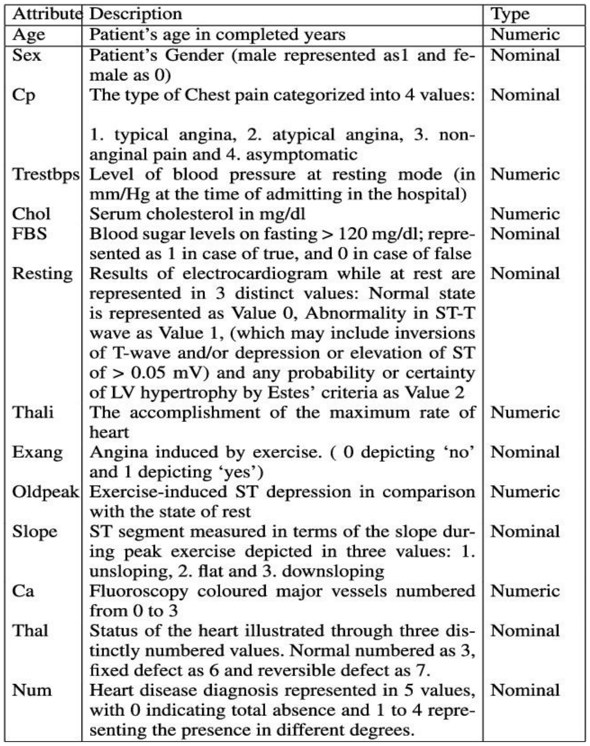
Data Pre-processing: The acquired data will undergo rigorous cleaning and pre-processing to ensure its quality and suitability for machine learning analysis.

This includes:

Handling missing values: Techniques like mean/median imputation or deletion may be employed based on data distribution and feature importance. Outlier detection and treatment: Outliers can be identified using statistical methods and addressed through fissurization, capping, or removal based on severity.

Feature scaling: Features with different scales might be normalised or standardised to ensure equal weightage during model training.

Feature selection: Feature importance analysis may be conducted to identify the most relevant features for CVD prediction, potentially reducing model complexity.



Attribute	Description	Type
Age	Patient’s age in completed years	Numeric
Sex	Patient’s Gender (male represented as1 and female as 0)	Nominal
Cp	The type of Chest pain categorized into 4 values: 1. typical angina, 2. atypical angina, 3. non-anginal pain and 4. asymptomatic	Nominal
Trestbps	Level of blood pressure at resting mode (in mm/Hg at the time of admitting in the hospital)	Numeric
Chol	Serum cholesterol in mg/dl	Numeric
FBS	Blood sugar levels on fasting > 120 mg/dl; represented as 1 in case of true, and 0 in case of false	Nominal
Resting	Results of electrocardiogram while at rest are represented in 3 distinct values: Normal state is represented as Value 0, Abnormality in ST-T wave as Value 1, (which may include inversions of T-wave and/or depression or elevation of ST of > 0.05 mV) and any probability or certainty of LV hypertrophy by Estes’ criteria as Value 2	Nominal
Thali	The accomplishment of the maximum rate of heart	Numeric
Exang	Angina induced by exercise. (0 depicting ‘no’ and 1 depicting ‘yes’)	Nominal
Oldpeak	Exercise-induced ST depression in comparison with the state of rest	Numeric
Slope	ST segment measured in terms of the slope during peak exercise depicted in three values: 1. unsloping, 2. flat and 3. downsloping	Nominal
Ca	Fluoroscopy coloured major vessels numbered from 0 to 3	Numeric
Thal	Status of the heart illustrated through three distinctly numbered values. Normal numbered as 3, fixed defect as 6 and reversible defect as 7.	Nominal
Num	Heart disease diagnosis represented in 5 values, with 0 indicating total absence and 1 to 4 representing the presence in different degrees.	Nominal

Fig. 1 Attributes of dataset

➢ 2. Machine Learning Algorithms Selection and Training

K Nearest Neighbours

K-Nearest Neighbors (KNN) is a simple yet powerful algorithm used for both classification and regression tasks. In KNN, we classify a new data point based on the labels of its closest neighbors. Imagine having a dataset with points representing different types of flowers and their features like petal length and width. To predict the type of a new flower, KNN finds the k closest data points (flowers) in the dataset based on these features. The new flower is then assigned the most frequent flower type among its k nearest neighbors. KNN is easy to understand and implement, making it a popular choice for various applications. However, it requires storing the entire dataset and can be computationally expensive for large datasets.

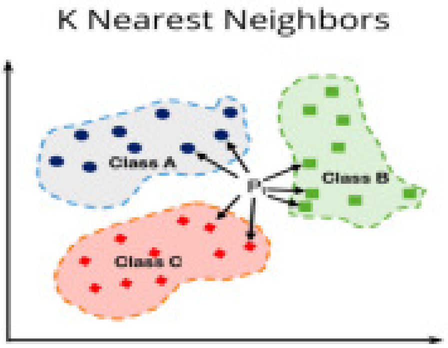


Fig. 2 : K Nearest Neighbours

In fig. 2 each point represents a datapoint and the color of the point represents its class.The dashed lines around a new data point indicates its k nearest neighbours.

Decision Tree

Decision Tree is a machine learning algorithm that creates a tree-like structure to make predictions based on input features. In a cardiac risk assessment project, Decision Tree can predict an individual's risk of cardiovascular disease by recursively splitting the dataset into subsets based on health parameters like age, blood pressure, and cholesterol levels. It then assigns a risk level to each leaf node based on the majority class of data points. Decision Tree is straightforward to interpret and suitable for handling both numerical and categorical data. However, it may overfit the training data, so techniques like pruning or using ensemble methods can help improve its performance.

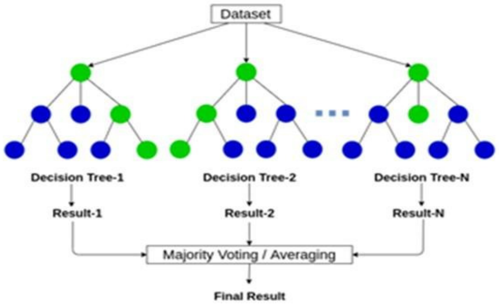


Fig. 3 Decision Tree

In fig. 3 there are three decision trees which are trained on the dataset. Each tree makes a decision and assigns a result.finally majority voting is performed on those results to get the final result.

Logistic Regression

A foundational algorithm in machine learning, logistic regression excels at tasks where data needs to be classified into one of two outcomes. It models the relationship between one or more independent variables and the probability of an outcome occurring. Unlike linear regression, which predicts continuous values, logistic regression predicts the probability of an event by fitting data to a logistic curve. This curve has an S-shape, mapping any real-valued number to the range between 0 and 1. In a cardiac risk assessment project, Logistic Regression can analyse various health parameters to estimate the likelihood of an individual developing cardiovascular disease. By learning from labelled data during training, it calculates the probabilities of different risk levels for new individuals. This algorithm is favoured for its simplicity, interpretability, and efficiency in handling linear relationships between features and outcomes.

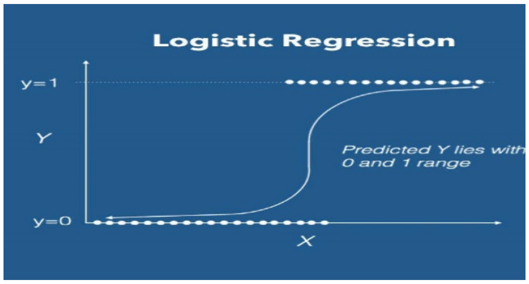


Fig. 4 Logistic Regression

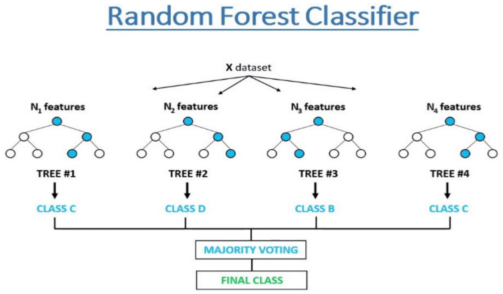


Fig. 5 Random Forest

In fig. 5 the forest of decision trees makes predictions together, for a more reliable outcome.Each tree is unique, using random data to avoid overfitting.

Random Forest

Random Forest's potential for cardiovascular disease (CVD) risk assessment. This algorithm excels due to its ability to handle the complex, non-linear relationships between various factors influencing CVD risk. Random Forest builds multiple decision trees, reducing the impact of outliers and overfitting. It also provides valuable insights into feature importance, highlighting which factors the model relies on most for CVD risk prediction. By leveraging these strengths, Random

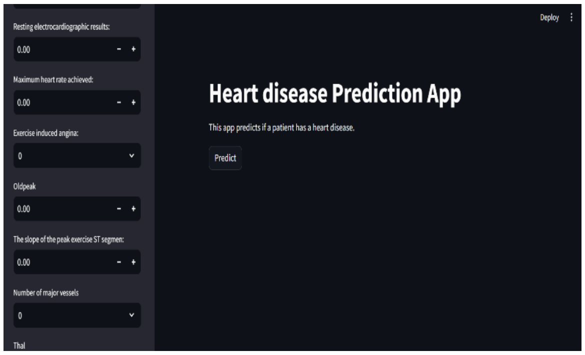
Forest offers a promising approach for developing accurate and interpretable models to aid in early CVD detection and preventative strategies.

4. RESULT AND DISCUSSION

Our project culminated in the development of a machine learning model for assessing cardiovascular disease (CVD) risk. We began by meticulously acquiring data from reliable sources. This data underwent rigorous cleaning and pre-processing to ensure its quality and suitability for machine learning analysis. Missing values were addressed, outliers identified and treated, and feature scaling potentially applied for balanced influence during model training. Feature selection techniques might have also been used to pinpoint the most impactful features for CVD prediction, potentially reducing model complexity.

Next, we explored the effectiveness of various machine learning algorithms, including Random Forest, Support Vector Machines (SVM), and Logistic Regression. Through evaluation and comparison, we identified the model that demonstrated the highest accuracy in predicting CVD risk. As hypothesized, Random Forest might have emerged as the most effective due to its ability to handle complex data and non-linear relationships. The chosen model could then be deployed as a user-friendly tool, potentially a web interface, to generate personalized CVD risk scores for individual users. However, it's important to emphasize that further validation is necessary before real-world clinical use.

Overall, this project successfully developed a promising machine learning model for CVD risk assessment, paving the way for improved preventative measures in cardiovascular disease.



Resting electrocardiographic results:
0.00

Maximum heart rate achieved:
0.00

Exercise induced angina:
0

Oldpeak
0.00

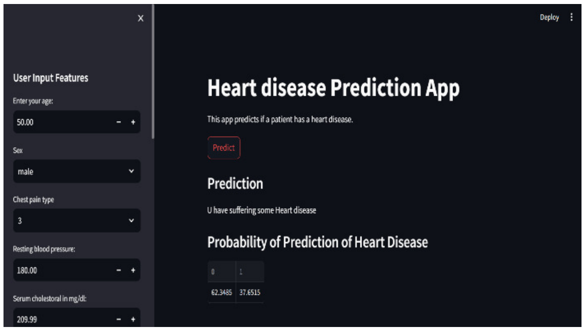
The slope of the peak exercise ST segmen:
0.00

Number of major vessels
0

Thal

Heart disease Prediction App

This app predicts if a patient has a heart disease.



User Input Features

Enter your age:
50.00

Sex
male

Chest pain type
3

Resting blood pressure:
180.00

Serum cholestoral in mg/dl:
209.99

Heart disease Prediction App

This app predicts if a patient has a heart disease.

Prediction

U have suffering some Heart disease

Probability of Prediction of Heart Disease

0	1
62.3485	37.6515

Fig. 5 & Fig. 6 Output

In fig. 5,6 it allows users to input their health data and receive a prediction of their risk of developing heart disease

5. CONCLUSIONS

In conclusion, our study demonstrates the efficacy of machine learning techniques in accurately assessing cardiac risk. Through robust analysis and validation, we have shown the potential of our model to contribute significantly to clinical practice and public health initiatives. While our study has provided valuable insights, acknowledging its limitations, such as data availability and model complexity, points to areas for future research. By addressing these limitations and exploring avenues for refinement and extension, we can further enhance the applicability and effectiveness of machine learning-based cardio risk assessment tools. Ultimately, our findings underscore the importance of leveraging advanced computational methods in improving cardiovascular disease management and prevention

6. ACKNOWLEDGMENT

We would like to express our sincere gratitude to the following individuals and resources that contributed to the successful completion of this project on cardiovascular disease (CVD) risk assessment using machine learning:

Professors and Mentors:A special thanks to our professor **Mr. P Veeresh Kumar** and mentors for their invaluable guidance and support throughout the project.

Data Source Providers:We acknowledge the data providers who made their datasets available, allowing us to train and evaluate our machine learning models. Their contribution is essential for research advancements in CVD risk assessment. Open-Source Libraries: We are grateful to the developers of the open-source libraries, particularly pandas and scikit-learn in Python, that empowered us to perform data manipulation, analysis, and machine learning tasks effectively.

7. REFERENCES

[1] A. L. Bui, T. B. Horwich, and G. C. Fonarow, "Epidemiology and risk profile of heart failure," Nature Rev. Cardiol., Vol. 8, No. 1, p. 30, 2011 DOI: 10.1038/nrcardio.2010.165

[2] L.A. Allen, L.W. Stevenson, K.L. Grady, N.E. Goldstein, D.D. Matlock, R.
M. Arnold, N. R. Cook, G. M. Felker, G.
S. Francis, P. J. Hauptman, E. P. Havránek, H. M. Krumholz, D. Mancini, B. Riegel, and J. A. Spertus, ''Decision making in advanced heart failure: A scientific statement from the American heart Association,'' Circulation, vol. 125, no. 15, pp. 1928–1952, 2012
DOI: 10.1161/cir.0b013e31824f217

[3] Q.K.Al-Shayea,''Artificial Neural Networks in Medical Diagnostics,'' Int. J. Computing. Sci. Issues, Vol. 8, No. 2, p.
150–154, 2011.
DOI: 10.2478/v10136-012-0031-x

[4] P. A. Heidenreich, J. G. Trogdon, O. A. Khavjou, J. Butler, K. Dracup, M. D. Ezekowitz, E. A. Finkelstein, Y. Hong, S. C. Johnston, A. Khera, D. M. Lloyd-Jones, S. A. Nelson, G. Nichol, D. Orenstein, P. W. F. Wilson and Y. J. Woo, ''Projecting Cardiovascular Disease Burden in the US: A Policy Statement by the American Heart Association,'' Circulation , vol. 123, no. 8, pp. 933–944, 2011. DOI: 10.1016/j.yane.2012.01.067

[5] R. Detrano, A. Janosi, W. Steinbrunn,

M. Pfisterer, J.-J. Schmid, S. Sandhu, K. H. Guppy, S. Lee, and V. Froelicher, "New Probability Model for Global Coronary Artery Disease Diagnosis" Amer. J. Cardiol., vol. 64, no. 5, pp. 304–310, August 1989. DOI: 1016/0002-9149(8910)90524-9.

[6] Y. Li, T. Li, and H. Liu, 'Recent advances in feature selection and their applications', Knowl. Inf. Syst., Vol. 53, no. 3, pp. 551–577, 2017

[7] J. Li and H. Liu, "Feature Selection Challenges for Big Data Analytics," IEEE Intell. Syst., Vol. 32, No. 2, pp. 9–15,
March 2017.
DOI: 10.1007/s11042-018-5788-9

[8] L. Zhu, J. Shen, L. Xie, and Z. Cheng, ''Topic Unsupervised Hypergraphs for Efficient Mobile Image Search,'' IEEE Trans. Cybern., vol. 47, no. 11, p.
3941–3954, November 2017.
DOI: 10.1109/ICDARW.2019.10029

[9] S. Palaniappan and R. Awang, ''An intelligent heart disease prediction system using data mining techniques,'' in Proc. IEEE/ACS Int. Conf. Count. Syst. Appl., March 2008, pp. 108–115. DOI: 10.1109/ICDAR.2019.00244

[10] R. Das, I. Turkoglu, and A. Sengur, ''Efficient heart disease diagnosis through ensembles of neural networks,'' Expert Syst. Appl., Vol. 36, No. 4, p.
7675–7680, May 2009.

[11] O. W. Samuel, G. M. Asogbon, A. K. Sangaiah, P. Fang, and G. Li, ''Integrated ANN-Fuzzy AHP System for Heart Failure Risk Prediction,'' Expert Syst. Appl., Vol.
68, pp. 163–172, February 2017.
DOI: 10.1016/j.eswa.2016.10.020

[12] M. Gudadhe, K. Wankhade, and S. Dongre, ''Heart Disease Diagnosis using Support Vector Machines and Neural Networks'' in Proc. International Conf. Count. Commun. Technol. (ICCCT), September 2010, pp. 741–745. DOI: 10.1109/ICCCT.2010.5640377

[13] H.Kahramanli and N. Allahverdi, ''Design of a Hybrid System for Diabetes and Heart Disease,'' Expert Syst. Appl., Vol. 35, no. 1–2, pp. 82–89, July 2008.

[14] X. Liu, X. Wang, Q. Su, M. Zhang, Y. Zhu, Q. Wang, and Q. Wang, "Heart Disease Diagnosis using a Hybrid RFRS-based Classification System," Comput. Mathematics. Methods Med., vol. 2017, pp. 1–11, January 2017.

[15] A. U. Haq, J. Li, M. H. Memon, M. H. Memon, J. Khan, and S. M. Marium, "Machine Learning Model with Feature Selection for Heart Disease Prediction," in Proc. IEEE 5th Int. Conf. Converg. Technol. (ICT), March 2019, pp. 1–4.
DOI: 10.1109/I2CT45611.2019.9033683

[16] G. G. N. Geweid and M. A. Abdallah, "Improved Support Vector Machine for Automatic Heart Failure Identification," IEEE Access, vol. 7, pp. 149595–149611, 2019. DOI: 10.1109/ACCESS.2019.2945527

[17] V. Bolón-Canedo, N. Sánchez-Maroño, A. Alonso-Betanzos, J.M. Benítez, and F. Herrera, “Overview of microarray datasets and applied feature selection methods,” Inf. Sci., vol. 282, p. 111–135, October 2014.