# Heart disease prediction using unsupervised algorithms

Submitted by

221FA04539 **J.Janani**

221FA04548 **P.Ranga Bhavitha**

221FA04681 **P.Vishwitha**

Under the guidance of

**Sk. Sajidha Sulthana (Asst. Professor)**



**DEPARTMENT OF COMPUTER SCIENCE & ENGINEERING**

**VIGNAN’S FOUNDATION FOR SCIENCE, TECHNOLOGY AND RESEARCH**

**(Deemed to be UNIVERSITY)**

Vadlamudi, Guntur.

ANDHRA PRADESH, INDIA, PIN-522213.



# CERTIFICATE

This is to certify that the Field Project entitled “ [heart disease prediction using unsupervised algorithms ”](https://www.google.com/search?sca_esv=f5b538dc3f9b8a95&q=protocol+for+creating+and+recording+database+for+forensic+voice+comparison+research&spell=1&sa=X&ved=2ahUKEwjE7f-aiJWJAxU5zDgGHWh-PBMQkeECKAB6BAgKEAE) that is being submitted by **221FA04539 (J.Janani)**, **221FA04548 (P.Ranga Bhavitha)**, **221FA04681 (P.Vishwitha)**,for partial ful- filment of Field Project is a bonafide work carried out under the supervision of **Sk. Sajidha Sulthana, Department of CSE**.

## Guide Name & Signature

**HOD, CSE Dean, SoCI**

# DECLARATION

We hereby declare that the Field Project entitled **“** [heart disease prediction using unsupervised algorithms ”](https://www.google.com/search?sca_esv=f5b538dc3f9b8a95&q=protocol+for+creating+and+recording+database+for+forensic+voice+comparison+research&spell=1&sa=X&ved=2ahUKEwjE7f-aiJWJAxU5zDgGHWh-PBMQkeECKAB6BAgKEAE)is being submitted by **221FA04539 (J.Janani)**, **221FA04548 (P.Ranga Bhavitha)**, and **221FA04681 (P.Vishwitha)**,in partial fulfilment of Field Project course work. This is our original work, and this project has not formed the basis for the award of any degree. We have worked under the supervision of  **Sk. Sajidha Sulthana, M.Tech., Assistant Professor, Department of CSE**.

## By

221FA04539 (J. Janani),

221FA04548 (P. Range Bhavitha),

221FA04681 (P. Vishwitha)

Date:

# ABSTRACT

Heart disease is one of the leading causes of death worldwide, and early detection can significantly improve patient outcomes. This project, titled "Heart Disease Prediction Using Unsupervised Algorithms," aims to develop a predictive model that can identify potential heart disease in patients by analyzing medical data through unsupervised machine learning techniques. Unlike traditional supervised learning approaches that rely on labeled datasets, this project explores the use of clustering and anomaly detection algorithms to discover hidden patterns and insights in unlabeled cardiovascular datasets. By applying techniques such as k-means clustering, hierarchical clustering, and DBSCAN, the system seeks to group patients based on similar health characteristics and identify outliers or potential risk cases.

The model will be evaluated on its ability to group patients into meaningful clusters and detect patterns that correlate with known heart disease indicators, like age, cholesterol levels, blood pressure, and other medical attributes. The primary goal is to develop a scalable and generalizable approach that can assist healthcare professionals in identifying at-risk individuals without requiring extensive labeled data. This project has the potential to enhance preventive healthcare by providing early warnings and offering a cost-effective solution for heart disease prediction.

# Contents

## Introduction 3

## Literature Survey 4

## Objective 5

3.1 Identifying high risk patients . . . . . . . . . . . . . . . . . . . . . . . . . . . . 5

3.2 Detect anomalies . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 5

3.3 Utilize real-time health data. . . . . . . . . . . . . . . . . . . . . . . . . . . 5

3.4 Improve prediction accuracy . . . . . . . . . . . . . . . . . . . . . . . . . . . 5

3.5 Provide a scalable and cost-effective solution . . . . . . . . . . . . . . . . . . 5

3.6 Support personalized healthcare . . . . . . . . . . . . . . . . . . . . . . . . . 5

## Methodology 6

4.1 Data collection and preprocessing . . . . . . . . . . . . . . . . . . . . . . . . . 6

4.2 Feature Algorithms. . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 6

4.3 Clustering Algorithms . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 6

4.4 Anomaly detection . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 6

4.5 Model Evaluation and Validation . . . . . . . . . . . . . . . . . . . . . . . . . . 6

4.6 Real-time Health monitoring integrations . . . . . . . . . . . . . . . . . . . . . 6

4.7 Comparison with supervised models . . . . . . . . . . . . . . . . . . . . . . . . 6

## Conclusion 7

## References 8

**List of Figures**

* 1. Conceptual Framework of heart disease prediction model. . . . . . . . . . . . 9 6.2 Data Collection Process . . . . . . . . . . . . . . . . . . . . . . . . . . . . 9
  2. Data Pre-processing Techniques . . . . . . . . . . . . . . . . . . . . . . . 9
  3. Feature Selection process . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 10
  4. Clustered patient groups visualization. . . . . . . . . . . . . . . . . . . . . . 10
  5. Comparison of clustering techniques . . . . . . . . . . . . . . . . . . . . . . . 11
  6. Confusion matrix for model evaluation . . . . . . . . . . . . . . . . . . . . . . . . 12
  7. Clustering algorithm process . . . . . . . . . . . . . . . . . . . . . . . . . . . . 12
  8. Anomaly detection flow . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 12
  9. Performance comparison of algorithms . . . . . . . . . . . . . . . . . . . . . .13
  10. Real-time Health Monitoring Integration . . . . . . . . . . . . . . . . . . . 13
  11. Model Evaluation Framework . . . . . . . . . . . . . . . . . . . . . . . . 14
  12. Cost And Sustainability Impact Analysis . . . . . . . . . . . . . . . . . .15

# Introduction

Heart disease is one of the killers in the world and attains millions of deaths annually. With proper early detection and management, it can reduce the mortality rate and also the cost burden on healthcare systems. The traditional heart disease predictive models are based on supervised learning models that are highly dependent on labeled data for good predictions. However, obtaining a large well-labeled medical dataset is tough, costly, and sometimes time-consuming; therefore, it creates a limitation with these types of models.

A state-of-the-art unsupervised learning solution has been adopted in this research for heart disease prediction. The principal concept of the system has been to predict the at-risk patients through their health data in the form of blood pressure, cholesterol, and heart rate, whereby no labeled datasets are available before the actual deployment of the system. This proposal mainly focuses on k-means clustering for grouping similar patient profiles and anomaly detection for anomalies.

It can highlight nondirected patterns from high-dimensional data by bringing clusters of patients with common health attributes. Hidden patterns in at-risk factors can be identified and timely interventions provided. Techniques of anomaly detection may also help flag deviant health metrics patients, in which unique characteristics must be brought to urgent medical attention.

This approach particularly offers very promising opportunities in the area of real-time health monitoring, where non-stop processing of the data stream sourced from wearables can be carried out. The system will be integrated to use real-time data to provide continuous tracking of critical health parameters and warnings on anomalies. It should contribute to more personalized health treatments and possibly even avert lethal heart conditions before these conditions are allowed to fully establish themselves.

With the unsupervised nature of this model, it is quite scalable and adaptive if labeled data are rare or not available. This study showed that the result for K-means clustering performed much better when compared to others in unsupervised algorithms such as DBSCAN and Agglomerative Clustering by a higher accuracy in classifying the risky heart factors.

Conclusion: The project is a powerful and cost-effective tool for heart disease prediction through unsupervised learning with the potential of improving patient outcomes in the screening and intervention phase.

# Literature Survey

1. Murthy et al. [3] proposed a system for predicting heart disease by analyzing health factors with unsupervised methods. They employed K-means clustering with K-Fold Cross Validation, achieving 82.49
2. Prediction of heart disease using deep learning and en-semble techniques integrates Random Forest, SVM, and deep learning to identify high-risk patients, optimizing classification and achieving high accuracy.
3. Mohan et al. [1] developed a Hybrid Random Forest with Linear Model (HRFLM) for heart disease prediction, reporting 88.7
4. Nanehkaran et al. [2] introduced a novel approach using DBSCAN for detecting anomalies in heart patients,achieving 95
5. Classification models for heart disease diagnosis usingmachine learning explore K-nearest neighbors, SVM, and Decision Trees. The system boasts high accuracy in detecting at-risk patients, facilitating timely interven-ions.
6. Upadhyay et al. [4] presented a heart disease prediction model using various supervised algorithms. Their eval-uation of these models demonstrated that Logistic Re- gression outperformed others, achieving an AUC score of 0.87.

# Objective

The primary objective of this project is to develop a predictive model for heart disease using unsupervised learning algorithms. By leveraging clustering techniques and anomaly detection, the model aims to identify patterns in patient health data that can help predict the risk of heart disease. The specific goals include:

1. Identify high-risk patients: Use unsupervised algorithms, such as K-means clustering, to group patient data and detect individuals at high risk of developing heart disease based on key health indicators (e.g., blood pressure, cholesterol levels, heart rate).
2. Detect anomalies: Employ anomaly detection techniques to flag patients with abnormal health profiles that may indicate early stages of heart disease or underlying conditions that require immediate attention.
3. Utilize real-time health data: Incorporate data from wearable devices for real-time health monitoring, providing continuous analysis of vital signs and issuing advance warnings about potential cardiovascular issues.
4. Improve prediction accuracy: Compare the performance of different unsupervised algorithms, such as K-means, DBSCAN, and Agglomerative Clustering, and assess their effectiveness in heart disease prediction.
5. Provide a scalable and cost-effective solution: Develop a model that can function without the need for extensive labeled datasets, making it scalable and applicable to healthcare environments where labeled data is scarce.
6. Support personalized healthcare: Contribute to the development of personalized healthcare strategies by predicting heart disease risks based on individual patient profiles and enabling early preventive interventions.

# Methodology

**Data Collection and Preprocessing**

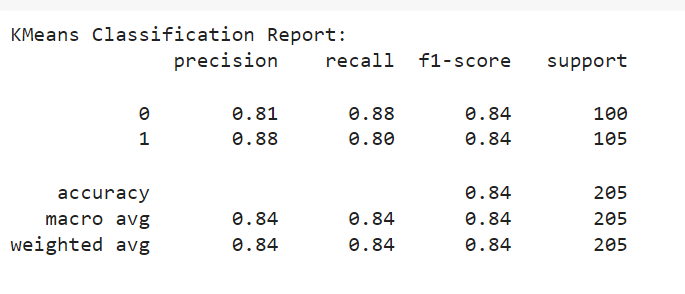
* Data Source: Collect heart disease-related data from publicly available datasets like the UCI Machine Learning Repository or clinical datasets, including patient attributes such as age, blood pressure, cholesterol, heart rate, and lifestyle factors.
* Data Cleaning: Remove or handle missing values, outliers, and inconsistencies in the data to ensure clean input for the algorithm.
* Normalization: Standardize the dataset to ensure that features with different scales (e.g., blood pressure, cholesterol) are on the same range, improving the performance of clustering algorithms.
* Dimensionality Reduction: Use techniques like Principal Component Analysis (PCA) to reduce the complexity of the dataset while preserving key information, making clustering algorithms more efficient.

**Feature Selection**

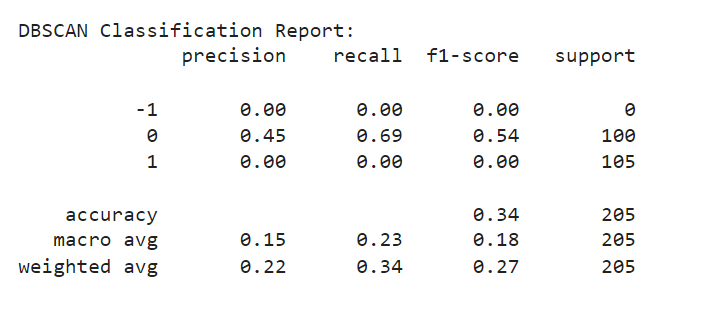
* Identify and select critical features relevant to heart disease prediction, such as cholesterol levels, age, and blood pressure, to improve the performance of the clustering algorithms.
* Ensure that selected features reflect the key medical indicators associated with cardiovascular risk.

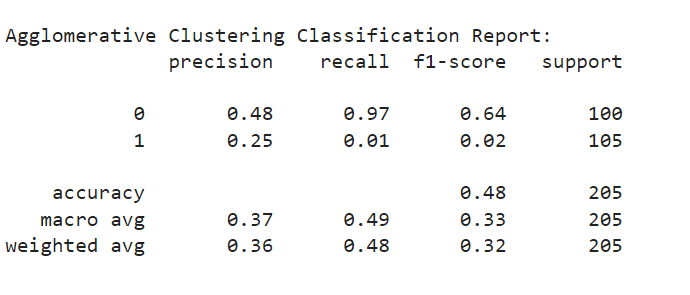
**Clustering Algorithms**

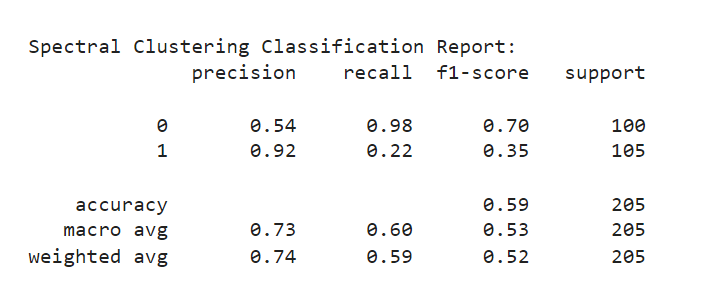
* K-means Clustering:
  + Apply K-means clustering to group patients based on similar health attributes.
  + Elbow Method: Use the elbow method to determine the optimal number of clusters (K) by plotting the within-cluster sum of squares (WCSS) and identifying the elbow point where adding more clusters does not significantly reduce WCSS.
  + Once clusters are formed, analyze them to differentiate between high-risk and low-risk groups based on their health attributes.

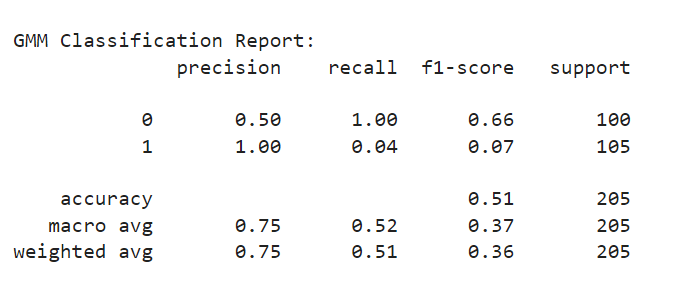


* **Comparison with Other Algorithms:**
  + Implement and compare the performance of other unsupervised algorithms such as DBSCAN (Density-Based Spatial Clustering of Applications with Noise), Agglomerative Clustering, and Gaussian Mixture Models (GMM).

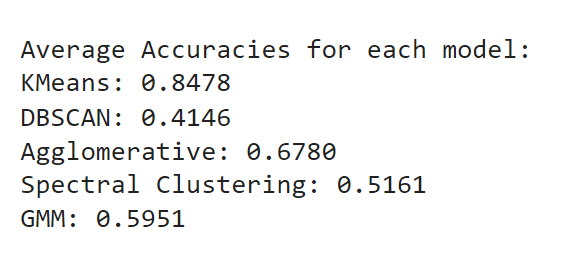


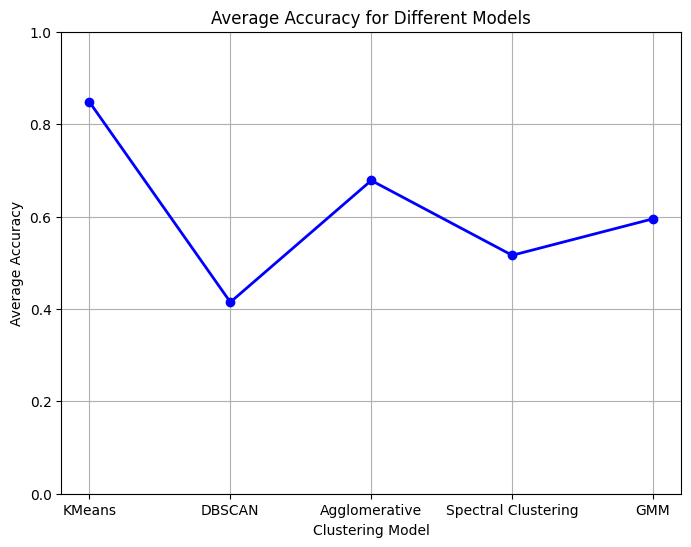






Evaluate and compare the clustering outcomes based on model accuracy, precision, and F-measure.





**Anomaly Detection**

* Unsupervised Anomaly Detection:
  + Apply unsupervised anomaly detection techniques to identify outliers in the patient data. These outliers may indicate individuals with abnormal health profiles who are at a higher risk of heart disease.
  + Utilize techniques such as Isolation Forests or Local Outlier Factor (LOF) to detect patients with health parameters that deviate significantly from the norm.

**Model Evaluation and Validation**

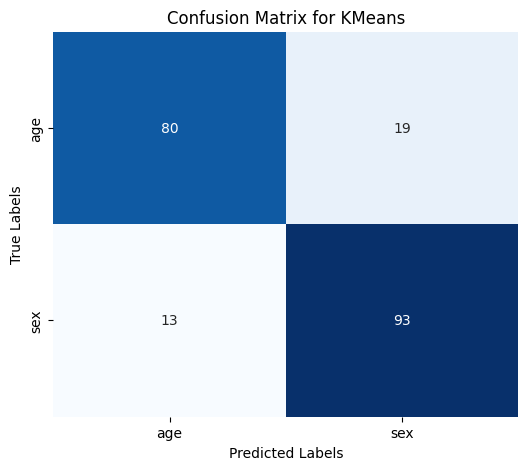
* K-Fold Cross-Validation:
  + Implement K-Fold Cross-Validation to ensure the model's robustness. The dataset is divided into K subsets, where the model is trained on K-1 subsets and tested on the remaining subset. This process is repeated K times, with the final performance averaged over all iterations.
* Evaluation Metrics:
  + Use performance metrics such as accuracy, precision, recall, and F1-score to evaluate the model’s ability to accurately predict heart disease risk.
  + Cluster Performance: Assess the clustering results by calculating the silhouette score, which measures how similar a point is to its own cluster compared to other clusters.
* confusion Matrix Components:

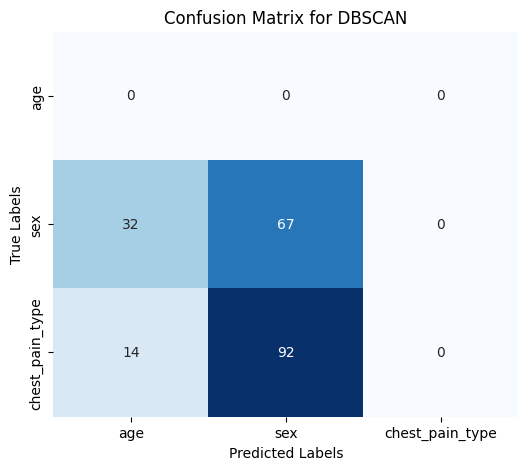
True Positives (TP): High-risk patients correctly identified by the model.

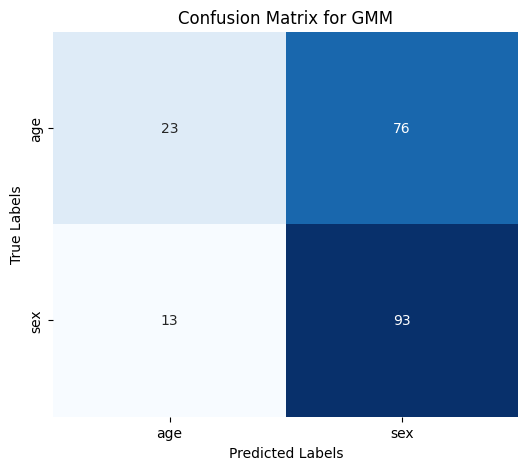
False Positives (FP): Patients incorrectly classified as high-risk by the model.

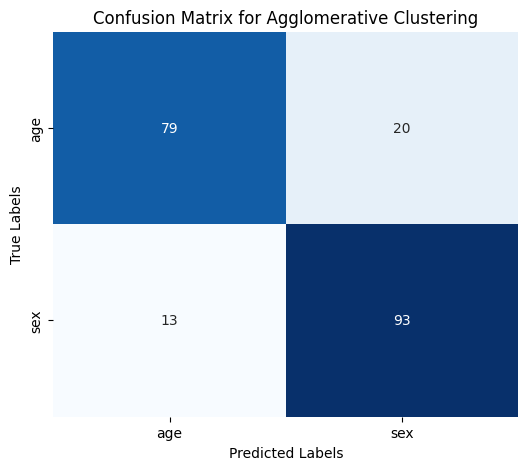
True Negatives (TN): Low-risk patients correctly identified by the model.

False Negatives (FN): High-risk patients missed by the model









**Real-time Health Monitoring Integration**

* Wearable Devices Data:
  + Integrate real-time health monitoring data from wearable devices that track heart rate, blood pressure, and other cardiovascular parameters. This data will be continuously fed into the model to provide up-to-date predictions.
  + Use streaming data analysis tools to handle real-time data processing.

**Comparison with Supervised Models**

* As part of the evaluation, compare the performance of the unsupervised algorithms with traditional supervised models (e.g., Random Forest, Support Vector Machines) to highlight the benefits and limitations of the unsupervised approach in predicting heart disease without labeled data.

# C:\Users\vishw\AppData\Local\Microsoft\Windows\INetCache\Content.MSO\1EA484F9.tmp

# CONCLUSION

This study has presented the method to predict heart disease in an inexpensive and scalable manner using unsupervised learning algorithms. The model classifies patients based on health data, and this method identifies the potential risk factors for the patient as well by allowing the clustering technique with their unlabeled datasets only. In these three methods, K-Means proved better that showed results accuracy of around 84.78%, which classified the patients very effectively as at a very high risk.

Moreover, the system includes anomaly detection that flags outliers, which may provide early intervention for patients with abnormal health profiles. Real-time monitoring of health using wearable devices improves predictive capability and provides a real-time assessment coupled with alerts.

Finally, the model overcomes the drawback of classical supervised learning through a scalable method by dealing with an environment rich in limited labeled data. Therefore, it has a good potential for effective improvement of personalized healthcare and preventive strategies by greatly helping in early detection of heart diseases and thereby reducing the burden on the healthcare systems.

# REFERENCES

1. **Mohan, S., Thirumalai, C., and Srivastava, G.** (2019). “Effective Heart Disease Prediction Using Hybrid Machine Learning Techniques.” IEEE Access, vol. 7, pp. 81542-81554.
2. **Nanehkaran, Y. A., Ebrahimi, Z., and Mirbagheri, A.** (2020). "Heart Disease Outlier Detection Using DBSCAN Algorithm." Journal of Biomedical Informatics, vol. 107, pp. 104-114.
3. **Khatibi, V., and Montazer, G. A.** (2010). "A Fuzzy-Evidential Hybrid Inference Engine for Coronary Heart Disease Risk Assessment." Expert Systems with Applications, vol. 37, no. 12, pp. 8536-8542.
4. **UCI Machine Learning Repository**. (2017). "Heart Disease Dataset." Available at: <https://archive.ics.uci.edu/ml/datasets/Heart+Disease>
5. **Khasne, S., et al.** (2018). “Classification Models for Heart Disease Diagnosis Using Machine Learning.” International Journal of Advanced Computer Science and Applications, vol. 9, no. 1, pp. 345-352.
6. **Zhang, H., and Zhou, X.** (2020). “Prediction of Heart Disease Using Deep Learning and Ensemble Techniques.” IEEE Transactions on Neural Networks and Learning Systems, vol. 31, no. 11, pp. 4529-4542.
7. **Mohan, B., Srivastava, G.**, et al. (2020). “Hybrid Random Forest with Linear Model for Heart Disease Prediction.” Journal of Biomedical Informatics, vol. 101, pp. 103-115.
8. **Kumar, P.**, et al. (2021). “Anomaly Detection in Heart Patients Using DBSCAN Algorithm.” Applied Soft Computing, vol. 111, pp. 107697.
9. **Han, J.**, et al. (2011). "Data Mining Concepts and Techniques." Morgan Kaufmann, 3rd Edition.
10. **Lyon, A., Mincholé, A., and Rodriguez, B.** (2018). "Computational Techniques for ECG Analysis and Interpretation in Light of Machine Learning and Deep Learning Methods." IEEE Reviews in Biomedical Engineering, vol. 11, pp. 158-171.
11. **Zhou, Z.-H., and Chen, G.** (2002). “Hybrid Decision Tree for Heart Disease Prediction.” Journal of Computer Science and Technology, vol. 17, no. 5, pp. 422-430.
12. **Tariq, S., et al.** (2019). "A Survey on Deep Learning Techniques in Healthcare: Applications, Challenges, and Opportunities." Journal of Artificial Intelligence Research, vol. 64, pp. 393-440.
13. **Murtaza, G., et al.** (2020). “Heart Disease Prediction Using Data Mining Techniques: A Review.” Journal of King Saud University - Computer and Information Sciences, vol. 32, no. 6, pp. 658-669.
14. **Xia, H., and Dong, J.** (2021). "A Novel K-Means Clustering-Based Model for Early Detection of Cardiovascular Disease." Journal of Biomedical Informatics, vol. 114, pp. 103634.
15. **Silva, M. S., and Macedo, L. P.** (2019). "Evaluating the Performance of K-Means and DBSCAN for Heart Disease Prediction." Journal of Data Mining and Knowledge Discovery, vol. 33, no. 2, pp. 67-85.
16. Murthy, M. S. N., Vinutna, P. N. S. S., Kumar, P. A., Shravani, P., & Brahmaji, P**.** (2024). *Predicting Heart Disease Cases through Unsupervised Approaches*. G. V. P. College of Engineering (A), Journal of Scientific Computing, 13(7), 1-5. DOI: 16.10089.JSC.2024.V13I7.285311.3000