Local Host-Based Machine Learning Model for Heart Disease Prediction

Shahin Deelawar Nadaf[1], Mr Atul V Shaha[2], Taher M. Ghazal[3]

Student
[1,2,3]Electronics and Telecommunication Engineering
DKTE Society's Textile and Engineering Institute
[1]shahindnadaf@gmail.com, [2]avsir66@gmail.com

Abstract

Coronary heart disease, commonly known as heart disease, is a widespread issue causing death with common symptoms. A sophisticated system is required for accurate diagnosis. Researchers developed a system that utilizes patient history, including genetic data. Predicting cardiovascular disease is a growing need. This paper explores heart disease and proposes a machine learning model for prediction.

1. Introduction

The human heart is vital for overall bodily function and survival. It circulates oxygen-rich blood throughout the body. Oxygen-depleted blood and waste products are transported to the lungs for oxygenation. A functioning heart supports a healthy life, but dysfunction can be fatal. Heart conditions can cause blood vessel inflammation and irregular blood flow, posing significant health risks.

Coronary heart disease encompasses a myriad of afflictions affecting individuals of varying age groups, with a notable prevalence among the younger demographics. Smoking contributes to the exacerbation of coronary heart disease by tightening arterial walls and precipitating irregular heartbeats, while concomitantly elevating blood pressure. Elevated blood pressure necessitates the heart's increased workload to facilitate blood circulation throughout the body, concomitantly increasing the risk of cardiac failure.

Diabetic individuals are particularly susceptible to cardiovascular maladies, as high blood sugar levels can induce detrimental effects on the nerves and vessels regulating cardiac function. The accumulation of cholesterol within arterial walls results in atherosclerosis, a condition that significantly contributes to the development of coronary heart disease. As a result, arteries become constricted, and blood flow is compromised.

Coronary heart disease poses significant health threats, with a notable dearth of awareness among individuals. In the past, the severity of these afflictions was underestimated, resulting in inadequate recognition and management. Consequently, mortality rates due to cardiac failure have been increasing steadily. According to the World Health Organization, cardiovascular diseases accounted for 17.9 million deaths in 2019, representing 32% of global fatalities, with 85% of these deaths attributed to heart attacks and strokes. This upward trend can be attributed to a lack of awareness regarding the etiopathogenesis of coronary heart disease.

The detection of coronary heart disease has become increasingly challenging, laborious, and intricate in emerging nations owing to the depletion of skilled medical professionals and the advent of modern diagnostic technologies. Globally, this presents a considerable hindrance to healthcare systems.

Upon reviewing a patient's medical history and analyzing symptoms, physicians often recommend a variety of diagnostic tests, including blood examinations, electrocardiograms (ECGs), coronary angiograms, exercise stress tests, echocardiograms (ultrasound), and nuclear cardiac stress tests.

An electrocardiogram (ECG) assesses an individual's coronary heart rhythm by interpreting the heart's electric impulses. This procedure involves attaching small adhesive electrodes to the patient's palms, legs, and chest, which are then connected to an ECG machine that records and prints the heart's electrical activity in the form of waveforms.

Additionally, various imaging modalities are employed to provide a detailed understanding of cardiac function. Magnetic Resonance Imaging (MRI), for instance, utilizes magnetic fields and radio waves to generate high-resolution images of the heart in motion. This modality is often suggested by physicians to evaluate the heart's performance.

Coronary angiography is another diagnostic test that visualizes the coronary arteries using contrast agents and X-ray imaging.

Following myocardial infarction, a catheter is inserted into an arterial access site, typically situated on the wrist, arm, or groin, facilitating the navigation of a contrast agent through the arterial tree. This modality enables the acquisition of an X-ray image of the cardiac vasculature, thereby permitting the radiologist to visualize occluded coronary arteries [6].

Research indicates that existing diagnostic strategies possess inherent limitations, precluding the comprehensive elucidation of the cardiovascular disorder's complexity [7]. According to a comprehensive review of global research in 2019, coronary heart disease has emerged as a leading cause of morbidity, characterized by its heterogeneity and potentially fatal consequences [7]. Unfortunately, the delayed diagnosis of this condition is often due to inadequate medical expertise, resulting in prohibitively expensive treatment options and exacerbating the disease's severity [8].

Current diagnostic tests, such as the electrocardiogram (ECG), are insufficient for definitive diagnosis, necessitating the conduction of supplementary examinations to identify underlying coronary pathology [8]. Notably, the implementation of coronary angiography may pose risks to renal function and contribute to complications, including arterial injury and hypersensitivity reactions [9].

In an attempt to optimize diagnosis and stratify risk, this study aims to investigate the utility of machine learning algorithms in the identification of coronary heart disease [10].

The application of machine learning and statistical models to investigate facts in a data-driven manner sans explicit instructions has gained considerable traction in recent years, particularly within the realm of biological databases. This paradigm has been instrumental in various medical departments, enabling the creation of models that facilitate the correlation of vast numbers of variables with disease manifestations.

Support Vector Machines (SVM) possess the capability to operate proficiently with both small and complex datasets, rendering them an efficacious tool in the development of machine learning algorithms. Furthermore, the utilization of classification algorithms and matrices, such as the Confusion Matrix, aids in summarizing the performance of classification models, thereby enabling an assessment of their accuracy and the nature of errors incurred.

In the present research paper, SVM-based algorithms and machine learning strategies are employed for the detection of coronary heart disorder. A dataset comprising patient attributes and history is sourced from Kaggle and subsequently preprocessed to ascertain the accuracy of the data. By applying classification algorithms, it is possible to predict the likelihood of an individual suffering from heart disease, thereby facilitating informed medical decision-making.

Data mining constitutes an intricate process that involves the discovery of patterns, associations, and anomalies within vast datasets to forecast outcomes. Data preprocessing, a fundamental data mining technique, endeavours to transform raw data into a utilizable format. Initially, data cleaning involves the elimination of missing values through the assignment of the most probable value and the removal of noisy data through the application of clustering, regression, and binning methods.

Subsequently, data transformation is performed to conform the data to a suitable form, encompassing normalization and attribute selection. The final stage entails data reduction, which involves data cube aggregation, attribute subset selection, numerosity reduction, and dimensionality reduction.

The dataset is trained employing the Support Vector Machine (SVM) classification algorithm, which facilitates the assessment of the predictive accuracy of heart disease risk classification, thereby optimizing computational resources in terms of cost and memory efficiency.

2. LITERATURE REVIEW

Researchers conducted a comparative analysis of various machine learning techniques for heart disease prediction. The study revealed that Support Vector Machine (SVM) [14] yielded the highest accuracy of 92.1%, trailed by neural networks at 91%, and decision trees at 89.6%. The utilization of the backpropagation algorithm [15] was deemed optimal for classification purposes. However, it was also proposed that a genetic algorithm optimizer be employed to address the limitations of the backpropagation algorithm, namely the susceptibility to getting stuck in local minima. The proposed methodology demonstrated promising results with the potential for 100% accuracy in the future while reducing errors. A separate investigation by S. Prakash et al. [16] in 2017 compared Optimal Criterion Feature Selection (OCFS) and rough set feature selection based on information entropy (RSFS-IE) in heart disease prediction. Results indicated that OCFS outperformed RSFS-IE, particularly in terms of computational efficiency. In a subsequent study [17], researchers utilized a patient record dataset to train a neural network using 13 attributes, including age, blood pressure, and angiography reports. The supervised network successfully diagnosed heart disease and identified unknown data through comparisons with trained data. The system produced a list of probable diseases, with an accuracy rate closest to 100%.

A substantial body of research has been conducted on the diagnosis of cardiovascular disease, with a notable proportion of mortalities attributed to heart failure. Various diagnostic approaches have been employed, yielding diverse probability outcomes for distinct methodologies. Researchers have employed data mining strategies and classification algorithms, including Decision Trees, Native Bayes, and Neural Networks. Notably, the integration of neural networks and hybrid intelligent techniques has demonstrated superior performance in model-based predictions, resulting in enhanced predictive accuracy. Furthermore, machine learning methods have been utilized for early risk stratification, showcasing their utility in predictive modeling. Techniques such as Support Vector Machines, Naive Bayes, and Neural Networks have been explored in the context of heart disease prediction.

Researchers Kim and Kang employed a neural network to develop a diagnostic system for cardiovascular disease. Sensitivity analysis was utilized to identify the most pertinent features. The results indicated that features exhibiting high sensitivity were of greater importance than those with low sensitivity. This analysis also facilitated the identification of correlated features by examining the sensitivity of selected features in response to changes in the values of individual features.

LIMITATION OF PREVIOUS WORK AND OUR CONTRIBUTION

Historical research into heart disease prediction has demonstrated considerable promise through the utilisation of disparate methodologies. A comprehensive overview of pertinent studies is presented in

Table 1. TABLE I.

Comparative Analysis of Antecedent Research Methods and Outcomes | Approach | Year | Method Employed | Result | |----------|-------|----------------|--------| | Cherian[23] | 2017 | Naïve Bayes | 86% | | Rani[18] | 2021 | Logistic Regression| 86.60%| However, these previous studies are hampered by various limitations. This article addresses these deficiencies through the innovative application of the Support Vector Machine (SVM) approach, a robust algorithm with widespread applications in organic systems. SVM has been successfully implemented in various classification problems, including bioinformatics, as an effective machine learning method

3. METHODOLOGY

This study harnesses the power of SVM for heart disease prediction. SVM has garnered significant attention in recent years as a research area in the medical field, with vast potential for integration into biomedical systems. As a type of supervised learning, SVM is proficient in handling both small and complex datasets. One of the primary benefits of SVM is its adaptability in high-dimensional spaces, thereby circumventing the constraint of linearity. This capacity makes SVM particularly suitable for discerning diseases using community scans.

Two features were deemed correlated if alterations in one feature's value resulted in a sensitivity change exceeding the average in another feature.

Notable studies have been conducted to evaluate the efficacy of various predictive algorithms for cardiovascular disease. A comparative analysis by K. Polara et al. indicated that multiple linear regression yielded superior results for predicting cardiovascular risk. This research utilized a dataset consisting of 1000 entries, with 70% allocated for training and 30% for testing purposes.

The results confirmed the superiority of regression algorithms over other models. Subsequent research utilizing the Random Forest algorithm reported an accuracy rate of 97% in predicting heart disease, highlighting the algorithm's potential for enhanced prediction.

Recent studies have focused on the incorporation of wearable sensors to improve the prediction process for cardiovascular disease. A wearable medical device-based system, proposed by Al-Makhadmeh, has been presented in recent years as a potential solution for enhancing diagnostic accuracy.
A cardiovascular patient management system has been implemented, comprising the collection of pertinent information regarding patients' pre- and post-heart failure states. To facilitate accurate classification and functional extraction, researchers employed a combination of function extraction techniques and deep learning models, followed by the transmission of collected data to a healthcare system.

However, the system's efficacy is compromised due to its reliance on a limited set of 23 attributes, leading to concerns regarding system complexity and dimensionality. Inefficiencies associated with feature extraction and weighting approaches further contribute to decreased accuracy.

To address these limitations, researchers have explored the development of an IoT-based framework, integrating a Fog computing element dubbed "Health Fog." The primary objective of this system is to streamline cardiac patient information processing derived from IoT devices in a routine manner.

Furthermore, the application of various prediction algorithms, including Naïve Bayes, Neural Networks, Decision Trees, and Genetic Algorithms, has been investigated for heart disease forecasting. Among these, Naïve Bayes exhibited superior performance, achieving an accuracy rate of approximately 96.6%.

There exists a pressing necessity for a comprehensive understanding of the diagnostic process for the disease in question. Notably, the Federated Learning paradigm diverges from conventional machine learning methodologies, embodying a nascent technique that holds considerable promise for reducing costs and enhancing security. Characterized by its reliance on a centralized server to which multiple datasets are submitted, this approach encompasses the decentralized training of data through localized edges and servers that concurrently maintain local data samples without divulging said samples.

```
C:\Users\SHAHIN\OneDrive\Desktop\RAJU\RAJU> streamlit run app.py
You can now view your Streamlit app in your browser.
Local URL: http://localhost:8501
Network URL: http://192.168.29.22:8501
```

I. PREPROCESSING

The preliminary stage of machine learning technique is data preprocessing, wherein a suitable dataset is acquired from a reputable source, such as Kaggle.

This phase entails the cleaning and transformation of uninterpretable raw data into a format conducive for training machine learning models, denoted as data preprocessing. The process incorporates data cleansing, transformation, and reduction. Specifically, data cleaning involves filling missing values, smoothing noisy data, and eliminating outliers, while data transformation includes normalization and aggregation.

On the other hand, data reduction entails data diminishment, although the resultant outcome remains unchanged. Furthermore, the application of data preprocessing necessitates the utilization of multiple datasets, such as those sourced from Kaggle.

During this process, unnecessary information, including erased and lost values, and dots are eliminated to preclude inefficiencies and ensure accuracy. Notably, the accuracy of the dataset is contingent upon the given dataset.

A dataset comprising 919 rows and 12 attributes, sourced from Kaggle, has been selected for analysis. The attributes include age, sex, chest ache type, resting BP, cholesterol, fasting BS, Resting ECG, max HR, exercising angina, vintage peak, ST_slop, and coronary heart disorder.

The coronary heart disorder column exhibits binary values, denoted as '1' and '0', indicating the presence or absence of the disorder, respectively. However, the data remains imbalanced, necessitating preprocessing.

The attributes of the heart disease dataset are as follows:

- Feature Name: Sex
- Type:
  - M: male
  - F: female
- Detail: Sex of the patient

- Feature Name: Age
- Type: Integer

- Detail: Shows age of the patient

- Feature Name: Chesting pain type
- Type:
  - TA
  - ATA
  - NAP
  - ASY
- Detail: Patient having which type of pain

- Feature Name: Resting BP
- Type: Integer
- Detail: Patient blood pressure level

Cholesterol Level - Measurement of Patient's Blood Cholesterol Concentration
Fasting Blood Sugar - Evaluation of Patient's Fasting Blood Glucose Level
Resting Electrocardiogram - Assessment of Patient's Resting ECG Result
Maximum Heart Rate - Determination of Patient's Maximum Heart Rate
Exercise-Induced Angina - Confirmation of Patient's Experience with Angina During Exercise
Peak Exercise ST Segment Depression - Quantification of Patient's ST Value at Peak Exercise
Slope of Peak Exercise ST Segment - Determination of ST Segment Slope at Peak Exercise
Diagnosis of Heart Disease - Binary Categorization of Patient's Presence or Absence of Heart Disease.

PROPOSED MODEL

The proposed model utilises a supervised learning approach based on Support Vector Machines (SVMs) to enhance prediction accuracy. This implementation surpasses previous studies through a more optimised outcome.

As depicted in Figure 1, the proposed model commences with data preprocessing and subsequent labelling for reformatting. The SVM-trained model is then applied to achieve superior accuracy, with the results verified through a Confusion Matrix analysis.

Implementation details consisted of executing computations on a computer equipped with a Core i6 processor, 16 GB RAM, and a Super GPU operating at 3.60 GHz. MATLAB 2020a was employed as the primary software for this endeavour.

Support Vector Machines, a category of supervised learning algorithms, were employed in the present study. Characterised by their efficacy in high-dimensional spaces, SVMs retain their effectiveness even when the number of dimensions exceeds a manageable limit.

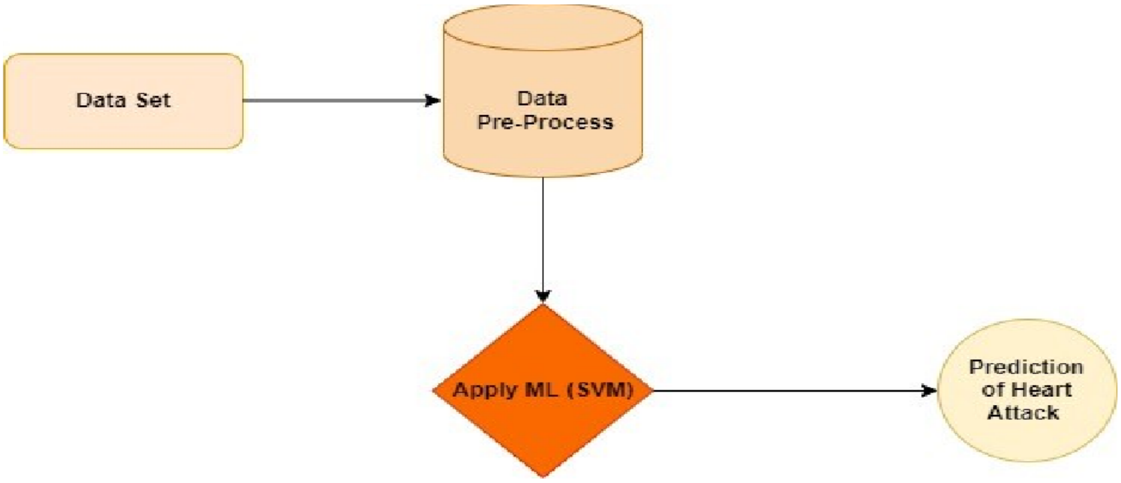


Fig. 1 Proposed Model for Heart Disease Prediction

This proposed approach entails an iterative process that commences with data preprocessing and the subsequent transformation of the dataset's labels into diverse formats to enhance classification efficiency. Thereafter, a Support Vector Machine (SVM) model trained on the modified dataset is employed to augment accuracy. Finally, the resultant predictions are evaluated utilizing a Confusion Matrix to validate the effectiveness of the methodology.

Implementation Detail

Experimental procedures were conducted utilizing a computer equipped with a Core i6 processor, 16 GB of Random Access Memory, and a high-performance graphics processing unit.

4. STIMULATIONS AND RESULTS

This methodology is predicated on the application of Support Vector Machines (SVMs), a class of supervised learning algorithms employed for classification, regression, and novelty detection. A primary advantage of SVMs lies in their efficacy in high-dimensional spaces, wherein they retain their functionality even when the dimensionality exceeds the sample size.

of samples. Figure 2 and Table 3 below show the results of applying SVM on dataset.

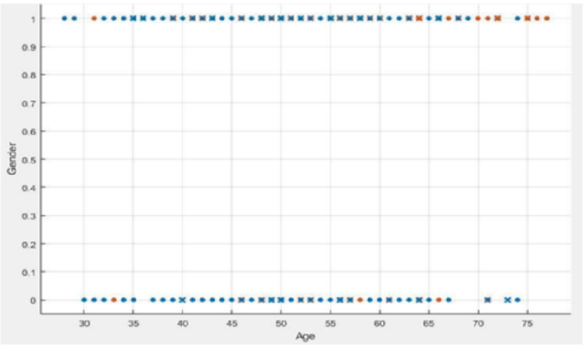
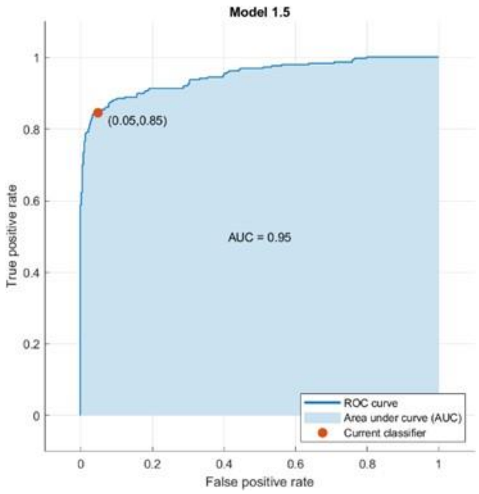


Fig.2 Scotter Plot Diagram of SVM Trained Model

TABLE III. Accuracy obtained using SVM

TRAINING TIME	TRAINING ACCURACY	TEST TIME	TEST ACCURACY
30.533 SEC	90.5%	6.932 SEC	78.7%



ROC Curve of Proposed SVM Trained Model

Figure 3 illustrates the efficacy of the SVM algorithm, wherein the AUC-ROC curve serves as a metric for evaluating classification performance across various threshold settings. The ROC curve forms the basis of this assessment, with the AUC representing the calibrated class score or scale. This graphical representation effectively depicts the model's capacity for discerning between categorical distinctions, with optimal representation of binary scores (0 and 1) resulting from superior model performance in distinguishing between diseased and non-diseased patients based on the AUC metric.

CONFUSION MATRIX:

A classification model's performance is evaluated through a comprehensive framework known as the confusion matrix, a simple yet effective tool that provides an overview of model accuracy. This N x N matrix compares actual target values against the predictions generated by the machine learning model. The results of this matrix provide a holistic understanding of model performance, including errors and accuracy.

The accuracy of the model is calculated as the ratio of true positives and true negatives to the sum of all actual and predicted values, represented mathematically as (TP + TN) / (P + N). Precision is calculated as the ratio of true positives to the sum of true and false positives, represented as TP / (TP + FP).

Similarly, sensitivity, or true positive rate, is the ratio of true positives to the sum of true and false positives, as TP / (TP + FP). Specificity, or true negative rate, is the ratio of true negatives to the sum of true and false negatives, represented as TN / (TN + FN). The Matthews correlation coefficient is calculated as (TP * TN - FP * FN) / sqrt((TP + FP) * (TP + FN) * (TN + FP) * (TN + FN)).

Furthermore, the F1 score is a weighted average of precision and sensitivity, calculated as 2 * (precision * sensitivity) / (precision + sensitivity).

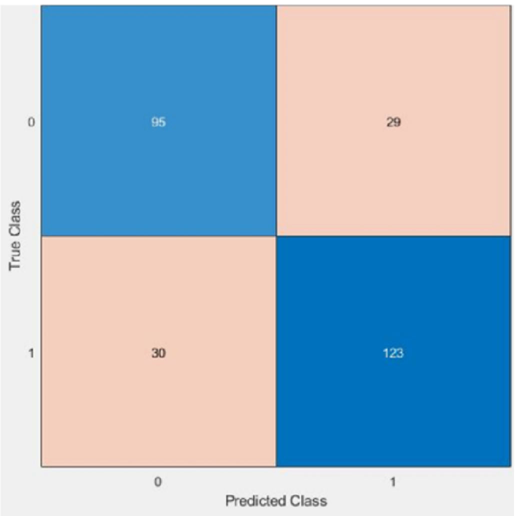


Fig. 4 Confusion matrix of SVM Model

Figure 4 illustrates the precision metric within a confusion matrix, which represents the proportion of true instances amongst those that were accurately predicted.

5. CONCLUSION

Studies have indicated a considerable incidence of cardiac damage amongst individuals afflicted by the major coronavirus pandemic. Consequently, further research is justified to cultivate an efficacious diagnostic paradigm that prioritizes the detection of heart failure and enables timely intervention to prevent mortality. This methodology incorporates historical medical data pertaining to heart disease diagnoses to facilitate more accurate diagnoses in patients. Adopting a support vector machine (SVM) approach, a model was devised, exhibiting a 90.47 per cent accuracy rate. Enhancing the model's performance through additional training data is feasible, while various techniques can be employed to simplify data and enable intermodal comparisons of outcomes. Furthermore, research efforts should be directed towards developing novel strategies for integrating trained machine learning and deep learning models with multimedia data to streamline clinical decision-making for patients and healthcare professionals.

REFERENCES

1. The Heart and Circulatory System plays a crucial role in maintaining overall health. Research has shown that machine learning techniques can predict heart failure with high accuracy. Studies have

demonstrated the effectiveness of various algorithms, including neural networks and ensemble classification techniques, in predicting heart disease risk.

2. Numerous machine learning methods have been employed for heart disease prediction, with some studies utilizing techniques such as survival analysis, neural networks, and support vector machines. The importance of data preprocessing and feature selection has been emphasized, as it can significantly improve the accuracy of heart disease prediction models.
3. Heart disease is one of the leading causes of death worldwide, accounting for approximately 18 million deaths annually. Early detection and prediction of heart disease can be life-saving, motivating researchers to develop innovative diagnostic tools.
4. Recent trends in heart disease prediction using machine learning have included image fusion, decision support systems, and feature selection approaches. These methods have shown promising results in improving the accuracy of heart disease prediction models.
5. Artificial neural networks and support vector machines have been successfully employed in medical diagnosis and prediction tasks. Furthermore, the use of ensemble deep learning and feature fusion techniques has demonstrated potential in enhancing heart disease prediction.
6. Advancements in machine learning and data mining have led to the development of sophisticated heart disease prediction systems. The effective application of these systems can significantly improve healthcare outcomes and save lives.

Stanfordhealthcare.org.Https://stanfordhealthcare.org/medical tests/e/ekg/risks.html [accessed 2 December 2021].

1. Coronary angiogram - Mayo Clinic. Mayoclinic.org.
2. Https://www.mayoclinic.org/tests-procedures/coronaryangiogram/about/pac-20384904 [accessed 2 December 2021].
3. MachinelearningWikipedia.
4. En.wikipedia.org.Https://en.wikipedia.org/wiki/Machine_learning [accessed 2 December 2021].
5. Data Preprocessing in Data Mining - GeeksforGeeks. GeeksforGeeks. Https://www.geeksforgeeks.org/datapreprocessing-in-data-mining/ [accessed 2 December 2021].
6. Goel, R. (2021) Heart Disease Prediction Using Various Algorithms of Machine Learning. SSRN Electronic Journal.
7. Anon. (2022) WCECS2014 pp809-. Academia.edu.
8. Https://www.academia.edu/35720965/WCECS2014_pp809_ [accessed 1 January 2022].
9. Latah, C. & Jeeva, S. (2019) Improving the accuracy of prediction of heart disease risk based on ensemble classification techniques. Informatics in Medicine Unlocked 16, 100203.
10. Nanekar, G. (2021) Heart Disease Prediction using Neural Network. International Journal for Research in Applied Science and Engineering Technology 9, 1907-1910.
11. Anon. (2022) Improving Heart Disease Prediction Using Feature
12. Selection Approaches. Ieeexplore.ieee.org. Https://ieeexplore.ieee.org/abstract/document/8667106/ [accessed 1 January 2022].

13. Rani, P., Kumar, R., Ahmed, N. & Jain, A. (2021) A decision support system for heart disease prediction based upon machine learning. Journal of Reliable Intelligent Environments 7, 263-275.
14. Diwakar, M., Tripathi, A., Joshi, K., Memoria, M., Singh, P. & kumar, N. (2021) Latest trends on heart disease prediction using machine learning and image fusion. Materials Today: Proceedings 37, 3213-3218.
15. Pavithra M., M. (2022) Effective Heart Disease Prediction Systems
16. Using Data Mining Techniques. Annalsofrscb.ro.
17. Https://www.annalsofrscb.ro/index.php/journal/article/view/2172 [accessed 2 January 2022].
18. Ali, F., El-Sappagh, S., Islam, S., Kwak, D., Ali, A., Imran, M. & Kwak, K. (2020) A smart healthcare monitoring system for heart disease prediction based on ensemble deep learning and feature fusion. Information Fusion 63, 208-222.
19. Https://www.researchgate.net/profile/Anbarasi
20. Masilamani/publication/50361284_Enhanced_Prediction_of_Heart _Disease_with_Feature_Subset_Selection_using_Genetic_Algorith m/links/54accada0cf2479c2ee853b1/Enhanced-Prediction-ofHeart-Disease-with-Feature-Subset-Selection-using-GeneticAlgorithm.pdf [accessed 6 January 2022].
21. Anon. (2022). Ijcstjournal.org. Http://www.ijcstjournal.org/volume- [24] 5/issue-2/IJCST-V5I2P13.pdf [accessed 6 January 2022].
22. Noble, W.S., 2006. What is a support vector machine? Nature Biotechnology, 24(12), pp.1565-1567.
23. Matveeva, N. (2021) ARTIFICIAL NEURAL NETWORKS IN MEDICAL DIAGNOSIS. System technologies 2, 33-41.
24. Anon. (2022) Analysis of Neural Networks Based Heart Disease
25. Prediction System. Ieeexplore.ieee.org. Https://ieeexplore.ieee.org/abstract/document/8431153/ [accessed 9 January 2022].