Prediction of Heart Disease using AI–Enhanced Naive Ensemble Decision Classifier

S. Sathyapriya[1]*, Arul Anitha[2] and V. A. Jane[3]*

[1]*Assistant Professor, Department of Computer Science, St. Joseph's College, Tiruchirappalli, Tamil Nadu*

[2]*Assistant Professor, Department of Computer Application, Jayaraj Annapackiam College for Women, Periyakulam, Theni, Tamil Nadu*

[3]*Assistant Professor, Department of B. Voc. (SD & SA), St. Joseph's College, Tiruchirappalli, Tamil Nadu*

******Corresponding Author e-mail id:* **withdifferentjane07@gmail.com**

Abstract

Data mining is an advanced technology that involves extracting valuable information from large datasets. This paper explores the application of data mining techniques in the medical sector, specifically focusing on the diagnosis of cardiac diseases through the analysis of extensive datasets. The study aims to identify significant patterns and information that can enhance clinical diagnostics. A novel naïve ensemble decision classifier is developed to forecast risk variables associated with heart disease. Additionally, the research utilizes the Authorized Cipher Text Encryption Standard (ACTES) method to ensure secure data transmission. The experimentation and testing are conducted using the Python programming language. The findings demonstrate the efficacy of the proposed diagnostic approach in predicting risk factors related to cardiovascular illnesses, thus highlighting its potential implications for improving medical diagnostics and data security.

Keywords: *Data mining, Heart disease, naïve ensemble decision classifier, Authorized cipher text Encryption Standard*

Introduction

Data mining is a crucial technology in modern healthcare, enabling the extraction of valuable insights from vast datasets [1]. The healthcare industry is inundated with extensive data encompassing patient records, diagnoses, and treatment information [2]. Despite the voluminous data generated by hospital its potential for informing therapeutic decisions remains largely untapped [3]. Recognizing this gap, hospitals are increasingly turning to Information Management System (IMS) for efficient data management. These systems generate a substantial amount of data, which is conveyed via charts, statistics, text, and graphics. However, this kind of data is seldom used to make therapeutic decisions [4].

This study focuses on the diagnosis of cardiac disease, a critical area where data-driven insights can significantly improve patient outcomes. Various data mining approaches have been explored to diagnose cardiac conditions and predict probabilities of disease occurrence [5] However, the diversity in predictive systems and methodologies poses challenges for healthcare institutions aiming to provide high-quality services at reasonable costs [6].

Various data mining methods have been employed to diagnose conditions and assess probabilities, with different systems proposed for predicting cardiac disease [7]. The healthcare sector grapples with the challenge of delivering high-quality services affordably, necessitating accurate diagnoses and proper medication dosages to avoid unfavorable outcomes. Computer-generated data and decision support systems (DSS) offer cost-saving solutions [8].

To address these challenges, this research proposes the use of a novel ensemble decision classifier for predicting heart disease [9]. Additionally, data security is paramount in healthcare, and the study incorporates the use of the Authorized Cipher Text Encryption Standard (ACTES) to safeguard sensitive patient information [10]. Hence, the major contribution of the paper are as follows:

1. To develop a flexible and secure classifier for accurate heart disease identification
2. To enhance data security through the implementation of the ACTES algorithm

The structure of this paper is as follows: the next section reviews relevant literature, followed by a detailed explanation of the methodology employed in this study. Subsequently, the findings are presented, analyzed, and discussed, leading to conclusions and implications for future research.

Related Works

Kaan Uyar et.al. proposed RFNN (recurrent fuzzy neural networks) and Genetic algorithms as computational tools for analyzing cardiac illnesses. The study included 297 patient data instances, with 45 used for testing and 252 for training. The testing phase yielded an accuracy rate of 97.78%. This heart disease dataset proved valuable for conducting thorough investigations. The study calculated several parameters such as accuracy, RMSE (root mean square error), misclassification error rate, specificity, sensitivity, precision, and F-score. The inclusion of specific metrics like accuracy and error rates provided a quantitative assessment of the proposed approach's effectiveness. However, further discussion could delve into the limitations of using computational models alone versus combined expertise from medical specialists in diagnosis and treatment decisions [11].

Anuradha Lamgunde et.al. proposed using a genetic algorithm alongside the backpropagation technique as an efficient method for predicting cardiac disease. The study utilized a variety of input characteristics to analyze the system's capability to predict heart disease. The approach involved a total of 13 medical variables, including gender, blood pressure, and cholesterol, to estimate the likelihood of a patient developing heart disease. The use of multiple medical variables demonstrates a comprehensive approach to assessing the risk of heart disease. However, further discussion could include insights into the accuracy and reliability of the predictive model compared to other existing methods, as well as potential limitations or areas for future research [12].

Theresa Princy and her colleagues conducted a study on various categorization approaches aimed at accurately identifying risk factors in individuals based on characteristics like gender, age, blood pressure, cholesterol, and pulse rate. They employed diverse data mining classification methods such as NB-Naïve Bayes, Decision Tree Algorithm, KNN, and NN-Neural Network to classify patients' risk levels. This comprehensive consideration of multiple factors led to a high level of accuracy in determining the risk level [13].

S. Indhumathi and colleagues suggested using the Naïve Bayes algorithm to predict the increased likelihood of cardiovascular illness in patients. They utilized pre-processed data for the training set, following the main steps of data mining, which are classification and prediction. In the classification step, they conducted pre-processing activities like data cleansing, normalization, and data reduction. The prediction phase involved categorizing and forecasting types of illnesses. As a result, the training set included illness types, while the testing set was created using questions. The output generated was then sent to doctors or experts [14].

In their study, S. Dangare et al. described the structure and functioning of a system's layers, particularly in the context of using the backpropagation algorithm. The description of the input, hidden, and output layers, along with the backpropagation process, is clear and straightforward. Input data is fed into the input layer, and the output layer generates the resulting output. They then compared the actual output with the predicted output. The backpropagation algorithm was used to calculate errors and adjust weights between the output and preceding hidden layers. Once the backpropagation process concluded, the forward procedure started and continued until the error was minimized [15]

K. Pramanik et al. proposed a Hybrid Algorithm that combined the ID3 and KNN algorithms to forecast cardiac disease. The data was pre-processed with the KNN algorithm. This pre-processed data formed the training set and was organized using a tree structure. The ID3 method predicted cardiac disease through a classifier, while the KNN Algorithm classified inaccurate values [16].

Rishabh Saxena and colleagues asserted that cardiovascular disorders had become increasingly prevalent in modern society. They utilized the Cleveland Heart Disease Dataset (HDD) to explore complexities and analyze individuals' heart disease-related conditions efficiently. Cardiovascular Disease (CVD) encompasses a range of medical conditions affecting the heart. The dataset comprised medical test results from 303 angiography patients at the Cleveland Clinic in Ohio. This dataset was also applied to 425 patients at the Hungarian Institute of Cardiology in Budapest, Hungary, with a frequency of 38% [17].

Ashok Kumar Dwivedi introduced a notable model that efficiently detected the presence of heart disease in a large number of samples. The study evaluated the effectiveness of six machine learning algorithms in predicting cardiac disease. The performance of these approaches was measured using eight different classification performance metrics. The assessment was conducted using the receiver operating characteristic curve. By employing logistic regression, the model achieved a classification accuracy of 85%, with a sensitivity of 89% and a specificity of 81% [18].

Animesh Hazra et.al. argued that the healthcare industry generated a substantial amount of data daily, much of which remained unexamined and underutilized. They noted a lack of adequate methods for extracting meaningful information from this data for tasks like clinical illness detection. The objective of their study was to review existing research on predicting cardiac disease using data mining methods. Their aim was to analyze a combination of mining algorithms and determine the most successful technique(s) [19].

Ashish Chhabbi et al. conducted research on different data mining methods to uncover and analyze hidden patterns in databases. These methods could prove beneficial in tackling challenging questions regarding the prediction of cardiac disease. Data was gathered from the UCI repository, and the study implemented the Naive Bayes and enhanced k-means algorithm. The results revealed that the advanced k-means algorithm achieved greater accuracy compared to the standard k-means method, which relied on a predetermined number of clusters [20].

Kamal Kant and his colleagues suggested a heart disease prediction model using the Naïve Bayes data mining method. This method is a statistical classifier that assumes properties are independent of each other. To determine the class, the posterior probability needs to be notably increased. In terms of predicting illness and real-time expert systems, the Naïve Bayes classifier shows strong performance and efficiency. Additionally, the study also employed Neural Networks and Decision trees [21].

Sharan Monica L et.al conducted a study on knowledge discovery in databases (KDD) using mining techniques such as J48, NB Tree, and basic CART. Their goal was to accurately predict heart disease using the fewest variables in the WEKA tool. J48, based on the open-source C4.5 algorithm, collects information to make judgments. The Naive Bayes (NB) classifier constructs models for continuous datasets based on predictive abilities. CART (Classification and Regression Trees) can swiftly predict important data connections. These methods were implemented using the WEKA software. CART achieved a precision rate of 92.2%, while J48 was the most efficient, completing in just 0.08 seconds [22].

Sumitra Sangwan and her colleagues developed a hybrid algorithm that merges the k-means and A-priori algorithms to efficiently extract large amounts of data and derive valuable information. Initially, the k-means clustering method is used for the clustering process. Next, the A-priori technique identifies frequent item-sets and extracts frequent term-sets for Boolean association rules. This approach employs a "bottom-up" strategy, systematically increasing frequent subgroups by adding one item at a time and then testing them against the entire dataset. The results indicate that combining clustering with A-priori significantly enhances heart disease prediction capability [23].

Rishi Dubey et.al conducted a study on predicting heart disease using different data mining methods. Several research studies have demonstrated that hybrid approaches outperform individual classification techniques in terms of accuracy. It has been observed that neural networks are highly effective in prediction tasks. The system achieves reliable results when trained well using genetic algorithms. Additionally, patients may soon have access to targeted therapy options, going beyond merely predicting the risk of developing heart disease in individuals [24].

Monika Gandhi et.al conducted research to explore how data mining approaches can effectively uncover hidden patterns in large datasets. The study aimed to assess how combining these techniques with other methods could help healthcare organizations make informed decisions. Data mining classification methods, including decision trees, neural networks, and Naive Bayes, were utilized to discover, extract, and categorize data [25].

Problem Statement

The rising prevalence of cardiovascular disease presents a critical challenge for healthcare systems globally. Timely and accurate identification of patients at risk is paramount, yet current diagnostic methods may be prone to errors and financial burdens for patients. In this context, there is a need for a cost-effective and reliable predictive model that utilizes the Naïve Ensemble Decision Classifier to enhance early detection and management of heart disease, thereby addressing key challenges faced by healthcare providers and patients alike.

Proposed Work

This study aims to develop a robust prediction model for heart disease using measurements obtained from physical examinations. To enhance both prediction accuracy and data security during transmission, we employ the Naïve Ensemble Decision Classifier for accurate prediction and the Authorized Cypher Text Encryption Standard to ensure secure data transmission. By leveraging these methods, we aim to create a reliable and secure prediction model that can assist healthcare providers in early detection and management of heart disease.

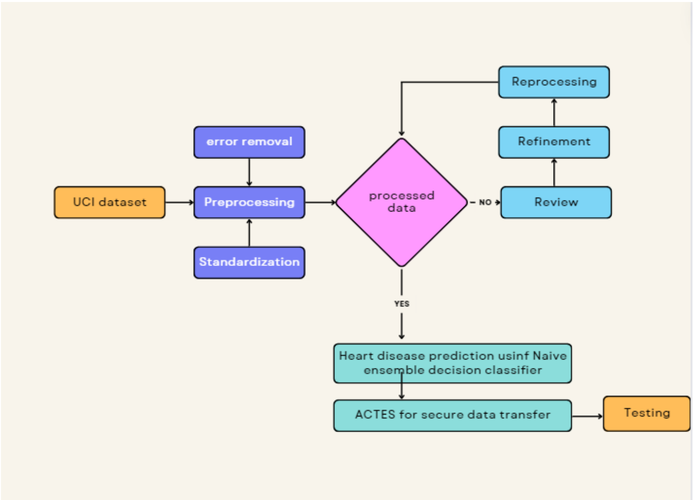

Figure 1 Schematic representation of the suggested methodology

a. Dataset

The UCI dataset is used to collect medical data from people diagnosed with cardiac problems. CAD-related concerns are often evaluated and documented for angiography. The patient's attributes are being collected, including demographic information, medical history, and laboratory data. These attributes encompass sex, age, presence of hypertension, smoking history, diabetes mellitus, type of chest pain, dyslipidemia, random blood sugar levels, levels of low and high density lipoprotein cholesterol, triglycerides, systolic and diastolic blood pressure, weight, height, BMI (body mass index), central obesity, waist circumference, ankle-brachial index, duration of exercise, METS obtained, rate pressure product, recovery duration with persistent ST changes, results of the Duke treadmill test, and angiography findings.

b. Preprocessing

Feature selection is a crucial concept for enhancing the performance of the model. By selecting the most salient attributes, it aids in reducing overfitting, enhancing accuracy, and reducing training duration. In order to choose the top 25 important attributes for prediction in both datasets we used the information gain ratio. This ratio is calculated by dividing the information gain by the intrinsic information. An increased ratio of a certain attribute might be advantageous in distinguishing between secure and hazardous software. Only features with an information gain of 0.5 or 0.85 were included in the dataset.

Categorising the features of a dataset is the first stage in feature engineering. Vectors are created from the categorised datasets. We used the 'Label Encoding' technique to transform the category data into vectors, however there are other viable options, such as the 'One Hot Encoding' approach.

c. Prediction

The Naive Ensemble Decision Classifier (NEDC) is a basic probabilistic classifier that relies on the Bayesian theorem. It assumes strong independence between two class features, regardless of whether one is present or not. NEDC classifiers perform well in supervised learning scenarios. They are based on conditional independence, meaning the value of one variable for a given class is not influenced by other variables. NEDC works best with inputs that have many dimensions. Using Naïve Bayesian techniques, NEDC can help create predictive models.

Step 1: With each record represented by an n-dimensional attribute vector and D standing for the training set, this would imply $X = (x_1, x_2 \ldots, x_n)$, predicting n measurements from n attributes (say A_1 to A_n.)

Step 2: Consider m no: of classes for prediction (say $C_1, C_2 \ldots \ldots C_m$) By Bayes' theorem:

$$P(C_i \mid X) = \frac{P(X|Ci)*P(Ci)}{P(X)} \quad (1)$$

Step 3: Since $P(X)$ being a constant for every class, hence $P(X \mid C_i)^* \, P(C_i)$ must be maximized.

Step 4: Thereafter class conditional independence is presumed. Thus,

$$P(X \mid C_i) = P(x_1 \mid C_i) * P(x_2 \mid C_i) \ldots \ldots P(x_m \mid C_i) \quad (2)$$

Step 5: For predicting class of $X, P(X \mid C_i)P(C_i)$ is computed for every class C_i.

The suggested classifier predict that disease class label of $X = C_i$ class if

$$P(X \mid C_i)P(C_i) > P(X \mid C_j)P(C_j) \quad \text{for } 1 \leq j \leq m, j \neq i \quad (3)$$

d. Data security

In this case, we assume that the ACTES method of random selection was followed consistently. Equation (1.14), which accepts a seed as input, produces cypher text by encoding it using the starting key and secret key in a conventional way.

The cryptographic hash function is denoted by "F", the encrypted seed is denoted as "signifies", and a random string is represented by the symbol "ρ". The last stage in deciphering ciphertext from a seed and an encrypted seed is described by the equation.

$$\hat{\delta} = F(q, g) \quad (4)$$

"Here $\hat{\delta}$ ", $Y(Q_n) = (\hat{\delta} \oplus \delta) + g_s (5)$ Herein, $Y(Q_n)$ signify the DP Q_n encoded communication. Some parts of an encrypted message, such as the seed, initial key, secret key, random string, and so on, can only be deciphered by authorized persons on the transmission side. The initial data may be securely restored using the encrypted seed, protected random string, beginning key, and secret key.

$$\text{linear encryt}_l \, (I, J) = i^p j + a. \quad (6)$$

If i, j are still the linear internal data components, while (a) is a constant

$$\text{quad}_l \, (I, J) = 1 - \frac{\|i-j\|^2}{\|i-j\|^2 + a}. \quad (7)$$

If i, j is the given input variables of a polynomial crypted kernel function

$$\text{poly}_l \, (I, J) = (\lambda i^p j + a)^d, \lambda > 0,$$
$$\text{sig}_l \, (I, J) = \text{tang} \, (\lambda i^p j + a), \lambda > 0. \quad (8)$$

It has been believed that improving data encryption should primarily focus on the kernel's singularity. A radial basis function kernel is required to convert the feature difference into a multidimensional discrepancy if it is always visible.

The last stage prior to data transmission is encryption.

Performance Analysis

The tests were conducted on an HP ELITEBOOK 64-bit Windows 10 Education laptop using a number of libraries, including Pandas, Numpy, Matplotlib, sklearn, and Keras. The laptop had a Python-written random-access memory (RAM) and an Intel(R) Core(TM) i5-8350U central processing unit (CPU).

For the purpose of assessing a model's efficacy in cardiac disease detection applications, provides a variety of well-known performance matrices. Figure 2 displayed the overall results of the

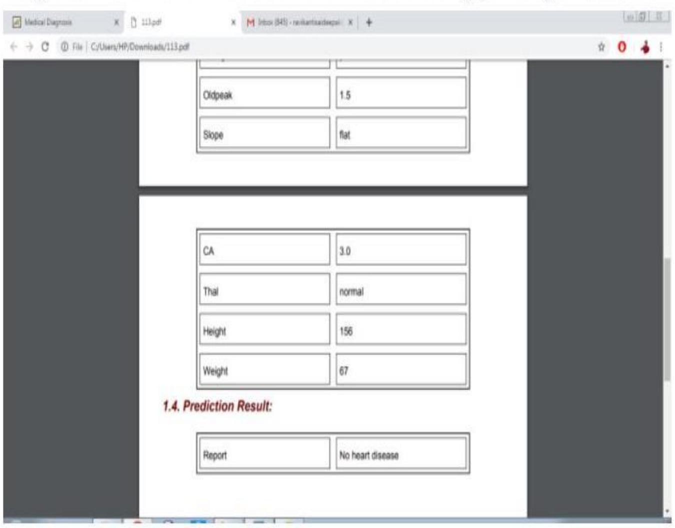

Figure 2 simulation output

In order to demonstrate the efficacy of the proposed technique, the widely used machine learning methodology was employed. These criteria are used to establish these performance metrics.:

- t_p = true positive
- t_n = true negative
- f_p = false positive
- f_n = false negative
- p = total positive $= t_p + f_n$
- n = total negative $= t_n + f_p$

Accuracy is defined as the degree to which a model correctly identifies illness and non-disease classifications:

$$\text{Accuracy} = \frac{t\_p + t\_n}{p+n}$$

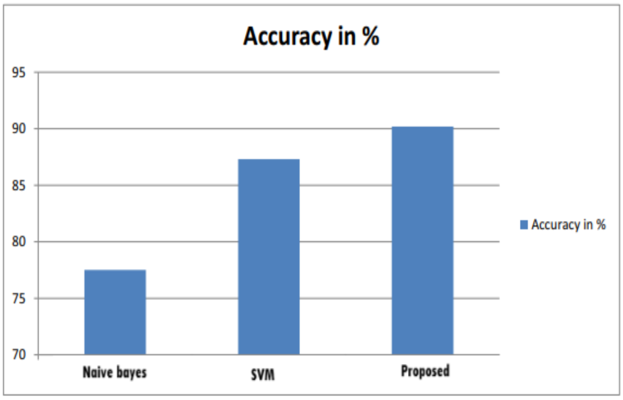

Figure 3 Accuracy prediction

The overall accuracy of the suggested and the existing mechanism was illustrated in figure 3.

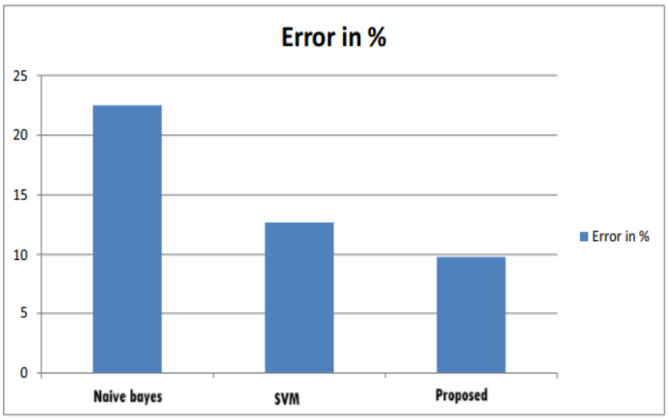

Figure 4 Error prediction

The overall error rate of the suggested and the existing mechanism was illustrated in figure 3.

Table 1 Security level analysis

Security level (%)				
"File size (mb)	**DES [26]**	**RSA [27]**	**AES [28]**	**Proposed**
20	78.0	77.23	73.0	85.0
40	80.0	85.0	76.0	90.0
60	83.0	88.0	80.0	92.0
80	85.0	80.0	79.0	98.0
100	90.0	90.0	85.0	96.0"

In Table 1 we can see a thorough comparison of the various encoding systems and the level of security they provide. With success rates of 78%, 77.23%, and 73%, respectively, DES, RSA, and AES provide the most strong security mechanisms for a 20 MB file. Data sizes 40–100 GB also receive a thorough review of security measures. In the graph up there, you can see how the proposed technique outperforms the state-of-the-art encryption methods.

The results demonstrate that the suggested method is both accurate and error-free, making it an adequate safeguard for the confidentiality and integrity of encrypted data transmissions.

Conclusion

The primary goal of this study is to develop a method for predicting the occurrence of cardiovascular disease that integrates the NED classification strategy with the AETES algorithm. It has been shown that the proposed method outperforms the current approach in terms of accuracy, with a yield of 90.77 percent accuracy even after decreasing the characteristics. Evaluations of AETES's security performance show that it outperforms DES, RSA, and AES. To make sure that application-based research on heart disease prevention doesn't fall completely behind technological advancements, a research process overhaul to cut down on delays is necessary. The primary goal of this study is to develop a method for predicting the occurrence of cardiovascular disease that integrates the NED classification strategy with the AETES algorithm. It has been shown that the proposed method outperforms the current approach in terms of accuracy, with a yield of 90.77 percent accuracy even after decreasing the characteristics. Evaluations of AETES's security performance show that it outperforms DES, RSA, and AES. To make sure that application-based research on heart disease prevention doesn't fall completely behind technological advancements, a research process overhaul to cut down on delays is necessary.

References

1) Purushottama. C, Kanak Saxenab, Richa Sharma (2016), "Efficient Heart Disease Prediction System", Elsevier, Procedia Computer Science, No. 85, pp. 962 – 969.
2) Jane, V. A., and L. Arockiam. "CID: Central Tendency Based Noise Removal Technique for Iot Data." Webology (ISSN: 1735-188X) 18, no. 6 (2021).
3) Arockiam, L., S. Sathyapriya, V. A. Jane, and A. Dalvin Vinoth Kumar. "Prevalence of Diabetes Mellitus in Tiruchirappalli District using Machine Learning." [3]
4) Jane, V. A., and L. Arockiam. "Iot Data Preprocessing-A Survey." Webology (ISSN: 1735-188X) 18, no. 6 (2021).

5) Arockiam, L., S. Sathyapriya, V. A. Jane, and A. Dalvin Vinoth Kumar. "Prevalence of Type-II Diabetics Association with PM 2.5 and PM 10 in Central Region of Tamil Nadu, India.
6) Priyaa, P. Kamakshi, S. Sathyapriya, and L. Arockiam. "Nutrition monitoring and calorie estimation using internet of things (IoT)." Int J Innov Technol Explor Eng 8, no. 11 (2019): 2669-72.
7) Ramanathan, K., and V. A. Jane. "Prediction of The Diabetics Using The Deep Learning Based Artificial Memory Vector Network." In 2022 Fourth International Conference on Emerging Research in Electronics, Computer Science and Technology (ICERECT), pp. 1-9. IEEE, 2022.
8) Shalet K.S, V. Sabarinathan, V. Sugumaran, V. J. Sarath Kumar (2015), "Diagnosis of Heart Disease Using Decision Tree and SVM Classifier", International Journal of Applied Engineering Research, Vol. 10, No.68, pp. 598-602.
9) Kaan Uyar, Ahmet İlhan (2017), "Diagnosis of heart disease using genetic algorithm based trained recurrent fuzzy neural networks", Elsevier B.V, ICTASC, pp. 588-593.
10) Abhishek Rairikar, Vedant Kulkarni, Vikas Sabale, Harshavardhan Kale, Anuradha Lamgunde "Heart Disease Prediction Using Data Mining Techniques" (C) IEEE, ICCC, 2017, p.p. 1-8.
11) Theresa Princy. R, J. Thomas "Human Heart Disease Prediction System using Data Mining Techniques", (C) IEEE, ICCPCT, 2016.
12) Miss. Chaitrali S. Dangare, Dr. Mrs. Sulabha S. Apte, "A Data mining approach for prediction of heart disease using neural network's", IJCET, Volume 3, Issue 3, 2012, pp. 30-40.
13) S. U. Amin, K. Agarwal, and R. Beg, "Genetic Neural Network Based Data Mining in Prediction of Heart Disease Using Risk Factors", IEEE, ICT, 2013.
14) Beant Kaur h, Williamjeet Singh, "Review on Heart Disease Prediction System using Data Mining Techniques", (c) IJRITCC, Vol.2, Issue: 10, p.p.3003-08, 2014.
15) Rishabh Saxena, Aakriti Johri, Vikas Deep and Purushottam Sharma "Heart Diseases Prediction System Using CHC-TSS Evolutionary, KNN, and Decision Tree Classification Algorithm", (c) Springer, Advances in Intelligent Systems and Computing, 2019, p.p.809-819.
16) Ashok Kumar Dwivedi "Performance evaluation of different machine learning techniques for prediction of heart disease", (c) Springer, Computer Applications and Mathematics, 2016.
17) Animesh Hazra, Subrata Kumar Mandal, Amit Gupta, Arkomita Mukherjee and Asmita Mukherjee "Heart Disease Diagnosis and Prediction Using Machine Learning and Data Mining Techniques: A Review", Advances in Computational Sciences and Technology, Volume 10, Number 7 (2017) pp. 2137-2159.

18) Ashish Chhabbi,Lakhan Ahuja,Sahil Ahir, and Y. K. Sharma,19 March 2016,"Heart Disease Prediction Using Data Mining Techniques", (C) IJRAT, Special Issue National Conference "NCPC-2016", pp. 104-106.

19) Kamal Kant, and Dr. Kanwal Garg, 2014, "Review of Heart Disease Prediction using Data Mining Classifications”, IJSRD, Vol. 2, Issue 04, pp. 109-111.

20) Sharan Monica L, and Sathees Kumar B,February 2016, "Analysis of Cardiovascular Heart Disease Prediction Using Data Mining Techniques, (C) IJMCS, Vol. 4, Issue 1, pp. 55-58.

21) Sumitra Sangwan, and Tazeem Ahmad Khan, Mar. 2015,"Review Paper Automatic Console for Disease Prediction using Integrated Module of Apriori and k-mean through ECG Signal", (c) IJTRE, Vol. 2, Issue 7, pp. 1368-1372.

22) Rishi Dubey, and Santosh Chandrakar, Aug. 2015, "Review on Hybrid Data Mining Techniques for The Diagnosis of Heart Diseases in Medical Ground", Vol. 5, Issue 8, pp. 715-718.

23) Monika Gandhi, Shailendra Narayan Singh "Predictions in Heart Disease Using Techniques of Data Mining", (c) IEEE, ICFTCAKM, 2015, p.p. 520 – 525.

24) Shivhare.R, Shrivastava.R, and Gupta.C, An Enhanced Image Encryption Technique using DES Algorithm with Random Image overlapping and Random key Generation, In: International Conference on Advanced Computation and Telecommunication (ICACAT),Bhopal, Vol.7, pp. 1-9,2018.

25) Rao.V, Sandeep.N, Rao.A.R, and Niharika.N, FPGA Implementation of Digital Data using RSA Algorithm, Journal of Innovation in Electronics and Communication Engineering, vol. 9, no. 1, pp. 34-37, 2019.

26) Dong.A, Randolph,D.A and Rajanna.S.K, Enabling privacy preserving record linkage systems using asymmetric key cryptography, In: AMIA Annual Symposium Proceedings,America, p. 380, 2019.