Journal of Statistics & Management Systems
ISSN 0972-0510 (Print), ISSN 2169-0014 (Online)
Vol. 27 (2024), No. 2, pp. 441–453
DOI : 10.47974/JSMS-1278

An experiment-based investigation into machine learning for predicting coronary heart disease

Lamiaa Mohammed Salem Akoosh [†]
Farheen Siddiqui *
Sherin Zafar [§]
Sameena Naaz [‡]
M. Afshar Alam [@]
Department of Computer Science and Engineering
School of Science & Technology
Jamia Hamdard
New Delhi
India

Abstract

Extensive inquiry has been conducted to explore potential applications of machine learning methodologies in the realm of cardiovascular disease management. To facilitate a more comprehensive investigation This study explores machine learning algorithms, specifically Support Vector Machines (SVM) and Artificial Neural Networks (ANN), for disease identification, focusing on cardiovascular diseases. Utilizing a Kaggle dataset of around seventy thousand medical records, the research aims to refine methodology and assess performance variations. SVM and ANN techniques are applied to the Kaggle dataset, revealing SVM accuracies of 0.9997 (default), 0.9998 (RBF kernel, C=100.0), and 1.000 (linear kernel, C=1000.0). The Feedforward neural network, using Adam optimization across 50 batches and 10 epochs, achieved perfect accuracy of 1.000.

Subject Classification: *Primary 60G25, Secondary 97RXX.*

Keywords: *Heart disease, Machine learning, SVM, ANN.*

This research was conducted during the NREUP at Michigan State University and it was sponsored by NSF grants DMS-1156582 and DMS-1359016.

[†] *E-mail:* **lamiaaakoosh@gmail.com**
* *E-mail:* **fsiddiqui@jamiahamdard.ac.in (Corresponding Author)**
[§] *E-mail:* **zafarsherin@gmail.com**
[‡] *E-mail:* **snaaz@jamiahamdard.ac.in**
[@] *E-mail:* **aalam@jamiahamdard.ac.in**

1. Introduction

Chronic diseases, particularly cardiovascular diseases (CVD), pose a global health challenge with significant morbidity and mortality. Factors like unhealthy lifestyles, stress, and limited healthcare access contribute to CVD's prevalence. [1,2]

This study addresses gaps in early detection by exploring the role of machine learning in cardiovascular risk prediction in community-based primary care. Leveraging advanced methodologies, the research focuses on support vector machines and artificial neural networks. Statistical analysis of input datasets and correlation examinations inform the study's findings, contributing to the broader landscape of machine learning applications in cardiovascular risk assessment.

The structure of the paper is outlined as follows: Section 2 provides an overview of other related work, while section 3 details the methods and materials employed. The results are thoroughly discussed in section 4. Finally, section 5 presents conclusions and considerations for future work.

2. Literature Review

(Bhatt et al., 2023) This study aims to build a robust predictive model for cardiovascular diseases using k-modes clustering and Huang initialization. Various machine learning models, including Decision Tree, XGBoost, Random Forest, and Multilayer Perceptron, were tested on a Kaggle dataset with 70,000 cases. The Multilayer Perceptron model, with cross-validation, exhibited the highest accuracy of 87.28%. AUC metrics also favored the Multilayer Perceptron, indicating its superior performance in cardiovascular disease prediction [3].

(AbdElminaam et al., 2023) This study enhances cardiac disease prediction accuracy using Machine Learning. Logistic Regression consistently outperforms other algorithms, reaching accuracy rates of 91.6% and 90.8%. Notably, Random Forest achieves the highest accuracy of 98.6% on the final dataset. The findings emphasize Machine Learning's effectiveness and propose collaboration with medical professionals for improved predictions in cardiac disease [4].

(Dalal et al., 2023) The study aimed to develop a machine learning algorithm for cardiovascular disease prediction using Kaggle's comprehensive dataset. Various models achieved a remarkable accuracy of 99.1%, representing a significant improvement over previous techniques.

These findings highlight the potential of machine learning for early disease diagnosis and improved treatment outcomes, contributing to healthcare advancements [5].

(Trigka & Δρίτσας, 2023) In this study, multiple machine learning models were tested, evaluating the effect of the synthetic minority oversampling method (SMOTE). The stacking ensemble model, coupled with SMOTE and 10-fold cross-validation, outperformed other models with an accuracy of 90.9%, precision of 96.7%, recall rate of 87.6%, and an impressive AUC of 96.1% [7].

(Khan et al., 2023This study applied machine learning techniques to create a robust framework for accurate prediction and decision-making in Cardiovascular Disease (CVD) patients. Conducted at Khyber Teaching Hospital and Lady Reading Hospital in Pakistan, the Random Forest algorithm emerged as the most successful, achieving an accuracy of 85.01%, a sensitivity of 92.11%, and an outstanding ROC curve of 87.73% for CVD. Despite lower specificity (43.48%) and minimal misclassification (8.70%), the findings support the Random Forest algorithm's suitability for classifying and predicting CVD, indicating potential widespread use in healthcare for improved disease categorization and prognostic capabilities [8].

(Rani & Masood, 2020) Research has examined many machine learning methods, such as Naïve Bayes (NB) and Feedforward Artificial Neural Networks (ANN). Weighted Support Vector Machine (WSVM) was used to study this dataset. The neural network model in this paper beat state-of-the-art Random Forest (WRF) CHD diagnosis algorithms. On the chosen dataset, the ANN model has 99.6% accuracy and 90.9% weighted accuracy. Thus, a noninvasive, nonclinical artificial neural network (ANN) model for unborn infant coronary heart disease (CHD) diagnosis is recommended [9].

(Δρίτσας & Trigka, 2023) This study utilizes supervised machine learning, highlighting the effectiveness of SMOTE for accurate Cardiovascular Disease (CVD) prediction. Analyzing various risk factors, it employs a Stacking ensemble model with SMOTE and achieves outstanding performance, with an 87.8% Accuracy rate, 88.3% Recall rate, 88% Precision rate, and an exceptional AUC score of 98.2%. [10].

3. Materials and Methods

This study aimed to assess the likelihood of Cardiovascular Disease (CVD) based on patient clinical data, utilizing various classification

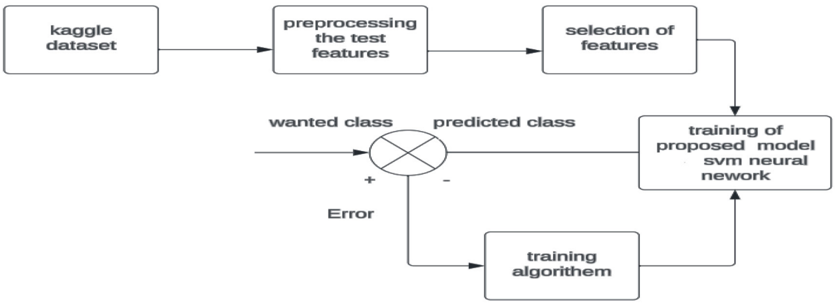

Figure 1

Proposed Model Methodology Flowchart

techniques such as Support Vector Machine and Neural Network methodologies. The goal is to establish an automated predictive model for CVD risk applicable in healthcare practice. Publicly available standard datasets, like Kaggle Dataset, were employed for CVD identification. The models were trained and validated using relevant patient data, and subsequent quantitative performance metrics were analyzed to outline the approach for predicting cardiovascular disease the mythology of proposed Model Methodology Flowchart shown in the Figure 1.

1. Obtain cardiovascular disease (CVD) data from the Kaggle Dataset.
2. Cleanse data by removing outliers, normalizing, and handling null values.
3. Reduce dataset dimensionality by eliminating highly correlated elements.
4. Choose two user-friendly machine learning algorithms for feature categorization.
5. Conduct a comprehensive evaluation and comparative analysis of performance metrics to identify the most effective approach.

3.1 *Data source*

The Cardiovascular Disease detection dataset, sourced from Kaggle via https://www.kaggle.com/datasets/sulianova/cardiovascular-disease-dataset, consists of 70,000 patient records with 11 attributes and a target variable.

3.2. *Data description*

The study classifies input features into three categories: Objective (empirical data), Medical Evaluation (comprehensive examination conclusions), and Subjective (therapeutic intervention data). Demographic and physiological factors in the dataset include age, height, weight, gender, blood pressure, cholesterol, glucose levels, smoking, alcohol, physical activity, and cardiovascular disease presence. Key risk factors include abnormal blood pressure, elevated cholesterol, and prolonged hyperglycemia. Diabetes increases susceptibility to comorbidities, and the dataset includes a mix of categorical and continuous variables detailed in accompanying Figures 2 and 3.

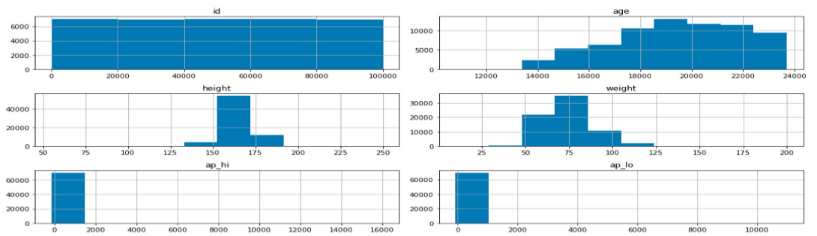

Figure 2

Histogram of CVD data set continuous characteristics

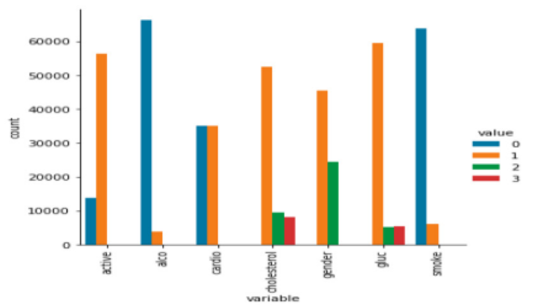

Figure 3

Factor plot of the Categorical Features

3.3. *Dataset Analysis*

The dataset analysis reveals key insights:

- Individuals without cardiovascular disease outnumber those with the condition.

- Females are more prevalent than males.
- Ages range from 29 to 65 years.
- Heights span from 25 to 250 centimeters.
- Weights vary from 10 kg to 200 kg.
- Non-smokers predominate over smokers.
- Abstainers from alcohol outnumber consumers.
- Regular physical activity correlates with better well-being.
- Normal glucose levels prevail over abnormal levels.
- Normal cholesterol levels surpass abnormal levels.
- Diastolic blood pressure ranges from 70 to 110.
- Systolic blood pressure varies from 150 to 160,020.
- Cardiovascular disease is more prevalent in females.
- Non-smokers and non-drinkers show elevated disease incidence.
- Physical activity associates with increased disease prevalence.
- Disease rates don't significantly rise with glucose or cholesterol levels.
- Diastolic blood pressure correlates with reduced disease susceptibility.
- Higher systolic blood pressure associates with decreased susceptibility.
- Heights from 150 to 199 cm show increased susceptibility.
- Cardiovascular disease is more common in those aged 54 to 65.
- Weights from 50 to 150 pounds exhibit increased disease susceptibility.
- Segmentation based on target values reveals variations in cholesterol and glucose levels.

The splitting of the dataset feature by the target variable shown in the Figure 4.

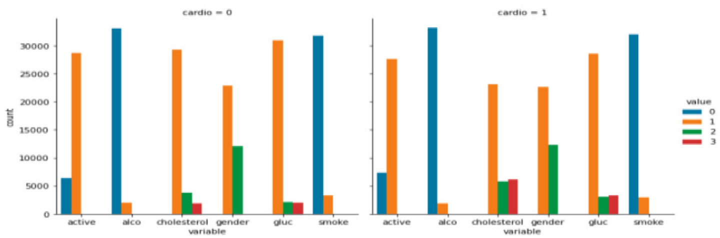

Figure 4

Splitting the dataset by target values

The correlation matrix indicates significant variations in age, weight, cholesterol, and target variable values as shown in the Figure 5.

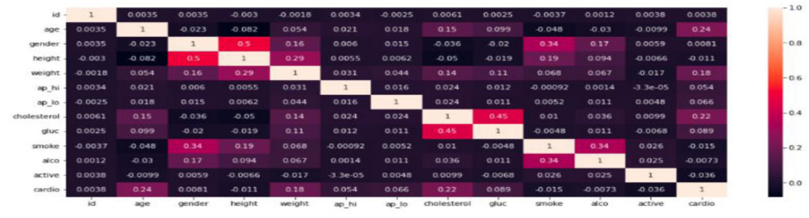

Figure 5

Correlation matrix between the Features and target value

Individuals with cardiovascular disease often exhibit high levels of cholesterol and glucose in their bloodstream, coupled with lower levels of physical activity [17-19].

Pre-Processing Of Cardiovascular Disease Data Cleaning Data:

Addressing data quality concerns, the authors enhanced precision by eliminating erroneous readings. Meticulous examination revealed potential typos, such as a minimum age of 29 years, and impractical values for height and weight. Additionally, inaccuracies in blood pressure readings, including negative values, were identified. To mitigate these issues, data points deviating significantly from the pattern were excluded.

- **Normalization**

In the initial step of pre-processing, the dataset is normalized using the StandardScaler() function. This transformative process involves subtracting the mean and dividing by the standard deviation, resulting in rescaled values within the range of 0 to 1. The formula for standardized values is expressed as follows:

$$z = \frac{x - \mu}{\sigma} \tag{1}$$

Herein, the symbol μ denotes the mean value within the specified distribution. The symbol σ represents the standard deviation inherent in the provided distribution. The standard score, denoted by Z, serves as a statistical metric quantifying the number of standard deviations by which an individual observation deviates from the mean.

4. Result and discussion

4.1 *Machine learning*

To evaluate classification models on test data, statisticians use a confusion matrix. Accuracy, a widely used metric, represents the proportion of correct predictions and is typically expressed as a percentage. The formula for accuracy is as follows:

$$Accurcy \frac{TP+TN}{TP+FP+FN+TN} \tag{2}$$

cm, columns= [‘Actual Positive: 1’, ‘Actual Negative: 0’],

Index= [‘Predict Positive: 1’, ‘Predict Negative: 0’])

ANN

A Feedforward Neural Network, using the Adam optimizer for 10 epochs with a batch size of 50, demonstrated effective performance on the dataset, converging the training and validation loss to around 1.744, and stabilizing the training loss at 1.755 as shown in the Figure 6.

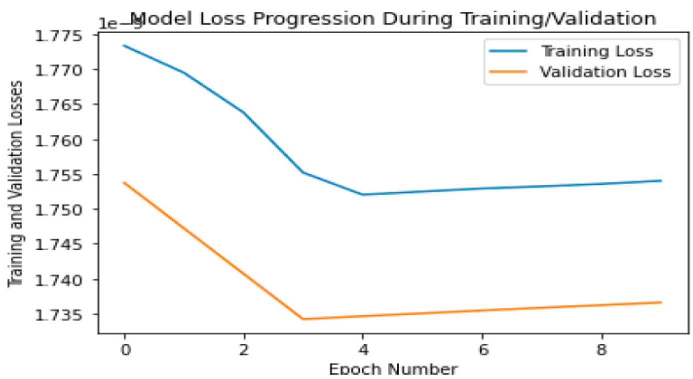

Figure 6

Loss for batch size 50 for 100 epoch

The accuracy level of ANN shown in Figure 7 and the confusion matrix of the same model shown in Figure 8.

The model’s accuracy scores are as follows:

- Default hyperparameters: 0.9997
- Linear kernel with C=1.0: 1.0000

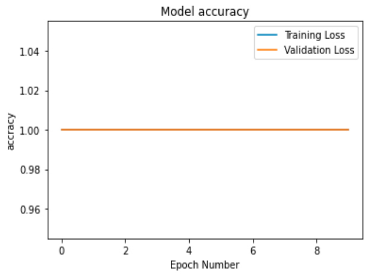

Figure 7

ANN accuracy

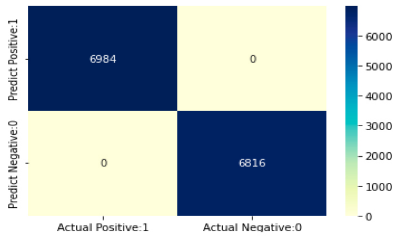

Figure 8

ANN Confusion Matrix

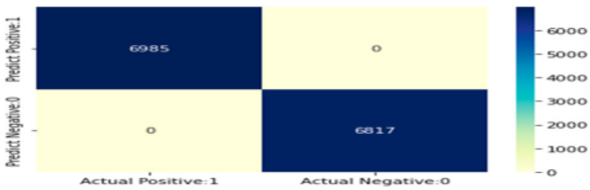

Figure 9

SVM Confusion Matrix

- Linear kernel with C=1000.0: 1.0000
- RBF kernel with C=100.0: 0.9998

Additionally, the training-set accuracy score is 1.0000

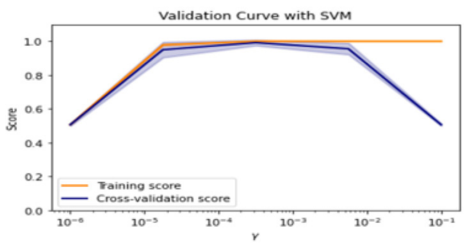

Figure 10

SVM Validation curve

Support Victor Model

This model employed both the linear kernel with varying C values and the Radial Basis Function (RBF) kernel the confusion matrix of the model displayed in Figure 9.

The model's accuracy scores are as follows:

- Default hyper-parameters: 0.9997
- Linear kernel with C=1.0: 1.0000
- Linear kernel with C=1000.0: 1.0000
- RBF kernel with C=100.0: 0.9998

Additionally, the training-set accuracy score is 1.0000
The Figure 10 displayed the SVM validation curve.

Figure 11 displayed the accuracy level of all the models.

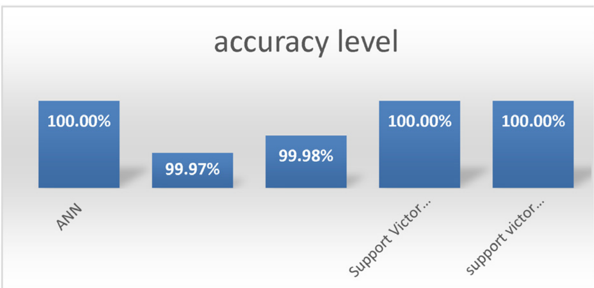

Figure 11

Accuracy level chart

The Table 1 illustrate accuracy level of each classifier.

Table 1

Accuracy Level For Each Classifier

Classifier model	Accuracy level
ANN	accuracy=1.000
SVM	with default hyper parameters: accuracy= 0.9997 with linear kernel and C=1.0: accuracy= 1.0000 with linear kernel and C=1000.0: accuracy= 1.0000 with RBF kernel and C=100.0 accuracy: 0.9998

The Table 2 provide a comparative analysis with current methods.

Table 2

Proposed model comparative analysis with current state-of-the-art method

Study	Author	Dataset	Classifier	Maximum accuracy
10	(Abdulsalam et al., 2023)	Cleveland dataset	SVM, ANN, QSVC, QNN, VQC Bagging QSVC	90.16%
11	(Nandy et al., 2021)	Cleveland dataset	Proposed Swarm-ANN	95.78%
12	(Salman Shukur & Mohsin Mijwil, 2023)	Cleveland dataset	Support vector machines	96%
13	(Özcan & Peker, 2023)	Original data set of 1190	CART mode	87.25%
14	(Biswas et al., 2023)	UCI	SVM and LR	94.51 %
15	(Özbilgin et al., 2023)	198 volunteers	SVM, KNN, Naive Bayes, SVM, DT NN	93%
13	(AlArfaj & Mahmoud, 2023)	KAGGLE	DEEP LEARNING	97,11%
14	(Abdulsalam ET AL, 2023)	Cleveland	Bagging-QSVC	90.16%
15	(Dileep, P. ET AL, 2023)	UCI dataset real time dataset	BiLSTM	94.78% 92.84%
16	(Shukur, & Mijwil, 2023)	Cleveland	SVM	96%
	THIS WORK	Cardiovascular disease, kaggle	KNN, ANN, SVM	100%

5. Conclusion

Recent research delves into the potential of machine learning for disease detection and prognosis, emphasizing the need for precision in real-world healthcare applications. The study evaluates traditional and deep learning methods for predicting heart diseases using Kaggle's Heart Disease Dataset (70,000 data points). Utilizing Support Vector Machine (SVM) and Artificial Neural Network (ANN) models, the SVM model with an RBF kernel and C=100.0 achieves 0.9998 accuracy, while the linear kernel SVM with C=1000.0 achieves 1.000 accuracy. The Feedforward neural network, with Adam optimization over 10 epochs and a batch size of 50, attains perfect accuracy. The study advocates for the superiority of deep learning, emphasizing its effectiveness in handling unstructured data without feature engineering. Future work aims to incorporate a hyper model using a deep neural network to further enhance accuracy.

Reference

[1] World Health Organization. (2022). About Us. [https://www.who.int/about].

[2] World Heart Federation. (2021). Cardiovascular Disease: A Global Crisis. [https://www.world-heart-federation.org/resources/cardiovascular-disease-a-global-crisis/].

[3] Bhatt, C., Patel, P., Ghetia, T., & Mazzeo, P. L. (2023, February 6). Effective Heart Disease Prediction Using Machine Learning Techniques. Algorithms; Multidisciplinary Digital Publishing Institute. https://doi.org/10.3390/a16020088.

[4] AbdElminaam, D. S., Radwan, M. A., Abdelrahman, N. M., Kamal, H. W., Elewa, A. K. A., & Mohamed, A. M. (2023, January 25). MLHeartDisPrediction: Heart Disease Prediction using Machine Learning. *Journal of Computing and Communication.* https://doi.org/10.21608/jocc.2023.282098

[5] Dalal, S., Goel, P., Onyema, E. M., Alharbi, A., Mahmoud, A., Algarni, M. A., & Halifa, A. (2023, March 1). Application of Machine Learning for Cardiovascular Disease Risk Prediction. Computational Intelligence and Neuroscience; Hindawi Publishing Corporation. https://doi.org/10.1155/2023/9418666 .

[6] Trigka, M., & Δρίτσας. (2023, January 20). Long-Term Coronary Artery Disease Risk Prediction with Machine Learning Models. Sensors; Multidisciplinary Digital Publishing Institute. https://doi.org/10.3390/s23031193.

[7] Khan, A., Qureshi, M., Daniyal, M., & Tawiah, K. (2023, February 20). A Novel Study on Machine Learning Algorithm-Based Cardiovascular Disease Prediction. Health & Social Care in the Community; Wiley-Blackwell. https://doi.org/10.1155/2023/1406060.

[8] Chandrasekhar, N., & Peddakrishna, S. (2023, April 14). Enhancing Heart Disease Prediction Accuracy through Machine Learning Techniques and Optimization. Processes; Multidisciplinary Digital Publishing Institute. https://doi.org/10.3390/pr11041210.

[9] Rani, S., & Masood, S. (2020, January 2). Predicting congenital heart disease using machine learning techniques. *Journal of Discrete Mathematical Sciences and Cryptography*, 23(1), 293–303. https://doi.org/10.1080/09720529.2020.1721862.

[10] Abdulsalam, G., Meshoul, S., & Shaiba, H. (2023). Explainable Heart Disease Prediction Using Ensemble-Quantum Machine Learning Approach. Intell. Autom. Soft Comput, 36, 761-779.

[11] Nandy, S., Adhikari, M., Balasubramanian, V., Menon, V. G., Li, X., & Zakarya, M. (2021, May 27). An intelligent heart disease prediction system based on swarm-artificial neural network. Neural Computing and Applications; Springer Science+Business Media. https://doi.org/10.1007/s00521-021-06124-1.

[12] Salman Shukur, B., & Mohsin Mijwil, M. (2023, April 1). Involving machine learning techniques in heart disease diagnosis: a performance analysis. *International Journal of Electrical and Computer Engineering* (IJECE), 13(2), 2177. https://doi.org/10.11591/ijece.v13i2.pp2177-2185.

[13] Özcan, M., & Peker, S. (2023, November 1). A classification and regression tree algorithm for heart disease modeling and prediction. Healthcare Analytics; Elsevier BV. https://doi.org/10.1016/j.health.2022.100130.

[14] Biswas, N., Ali, M. M., Rahaman, M. A., Islam, M., Mia, M. R., Azam, S., Ahmed, K., Bui, F. M., Al-Zahrani, F. A., & Moni, M. A. (2023, May 2). Machine Learning-Based Model to Predict Heart Disease in Early Stage Employing Different Feature Selection Techniques. BioMed Research International; Hindawi Publishing Corporation. https://doi.org/10.1155/2023/6864343.

[15] Biswas, N., Ali, M. M., Rahaman, M. A., Islam, M., Mia, M. R., Azam, S., Ahmed, K., Bui, F. M., Al-Zahrani, F. A., & Moni, M. A. (2023, May 2). Machine Learning-Based Model to Predict Heart Disease in Early Stage Employing Different Feature Selection Techniques. BioMed Research International; Hindawi Publishing Corporation. https://doi.org/10.1155/2023/6864343.

[16] AlArfaj, A. A., & Mahmoud, H. A. H. (2023, January 1). Deep Learning Prediction Model for Heart Disease for Elderly Patients. Intelligent Automation and Soft Computing; Taylor & Francis. https://doi.org/10.32604/ia.

[17] Kumar, Anil , Gupta, Amit Kumar , Panwar, Deepak , Chaurasia, Sandeep & Goyal, Dinesh. Operating system security with discrete mathematical structure for secure round robin scheduling method with intelligent time quantum, *Journal of Discrete Mathematical Sciences and Cryptography*, 26:5, 1519–1533 (2023).

[18] Kushwaha, Satpal Singh, Joshi, Sandeep & Gupta, Amit Kumar. An efficient approach to secure smart contract of Ethereum blockchain using hybrid security analysis approach, *Journal of Discrete Mathematical Sciences and Cryptography*, 26:5, 1499–1517 (2023), DOI: https://doi.org/10.47974/JDMSC-1815.

[19] Joshi, Ruchi, Mathur, Priya, Gupta, Amit Kumar, Singh, Suyesha, Paliwal, Vismita & Nayar, Sejal. Mathematical modeling of intelligent system for predicting effectiveness of premenstrual syndrome, *Journal of Interdisciplinary Mathematics,* 26:3, 551-562 (2023).