Predictive Analysis for Early Detection of Heart Disease and Diabetes: A Machine Learning Approach

[1]Mr. Rishi Cheravath, [2]Dr. Abhijit Banubakode

[1], [2] MET-Institute of Computer Science, Mumbai, India mca22_1307ics@met.edu,

principal_ics@met.edu

*Abstract –**Diabetes and heart disease pose significant health risks globally, leading to substantial costs and challenges for healthcare systems. This research investigates the utilization of logistic regression and support vector machines (SVM) in early disease detection, focusing on both heart disease and diabetes using clinical data. Despite encountering challenges such as restricted dataset size and quality, our study reveals the potential of both models in achieving accurate disease detection. These findings underscore the versatility of logistic regression and SVM as effective tools for proactive disease identification, providing valuable insights for enhancing patient healthcare.***

Keywords – Diabetes, Heart disease, Machine learning, Support vector machines (SVM), Logistic regression, Predictive analysis, early disease detection.

1. Objectives

A. Explore the concept of predictive modeling for early detection of heart disease and diabetes.

B. Develop and validate predictive models using machine learning algorithms.

C. Assess the performance of predictive models in terms of accuracy and effectiveness.

2. INTRODUCTION

Diabetes and heart disease are two serious health problems worldwide that burden healthcare systems and affect many people. The Early detection of these illnesses is critical to both reducing healthcare costs and improving patient care. Machine learning techniques such as logistic regression and support vector machines (SVM) can be used to evaluate medical data and provide promising means of identifying diseases before they become serious. In this study, we examine the possible uses of logistic regression and SVM in the risk assessment for diabetes and heart disease. By developing healthcare procedures and addressing the benefits and drawbacks of predictive disease detection methods, we want to enhance patient care.

Furthermore, the integration of machine learning algorithms like logistic regression and SVM into healthcare systems enables the study of massive volumes of patient data, and medical history. Additionally, as research and data are gathered, these algorithms' dependability and accuracy are

continually enhanced, increasing their value over time in proactive disease treatment. The further machine learning is used in healthcare, the more crucial it is that data scientists, physicians, and doctors collaborate to properly exploit these technologies for the benefit of people worldwide.

3. METHODOLOGY

3.1 Data Collection

The dataset utilized in this study was obtained from Kaggle, a widely recognized platform known for hosting various datasets and machine learning challenges. Specifically, we employed two datasets for our analysis: one focusing on diabetes and the other on heart disease. The diabetes dataset consists of 768 rows and 9 columns, while the heart disease dataset comprises 303 rows and 14 columns

Diabetes Dataset

Data Element	Description
Pregnancies	Number of pregnancies
Glucose	Glucose Level(mg/dl)
Blood Pressure	Blood Pressure(mm Hg)
Skin Thickness	Skin Thickness(mm)
Insulin	Insulin Levels
BMI	Body Mass Index
DiabetesPedigreeFunction	Diabetes pedigree function
Age	Age(Years)
Outcome	Diabetes outcome(0: No, 1:Yes)

Table 1. Feature Description (Diabetes)

Heart Disease Dataset

Data Element	Description
Age	Age(Years)
Sex	Gender(0:Female,1:Male)
Cp	Chest Pain Type (0: Typical angina, 1: Atypical angina, 2: Non-angina pain, 3: Asymptomatic)
Trestbps	Resting BP (mm Hg)
Chol	Cholesterol (mg/dl)
Fbs	Fasting BS (1: >120 mg/dl, 0: <=120 mg/dl)
Restecg	Resting ECG
Thalach	Max HR (bpm)
Exang	Exercise Angina (1: Yes, 0: No)
Oldpeak	ST Depression
Slope	ST Slope

Ca	Major Vessels (0-3)
Thal	Thalium Stress Test
Target	Heart Disease (1: Yes, 0: No)

Table 2. Feature Description (Heart Disease)

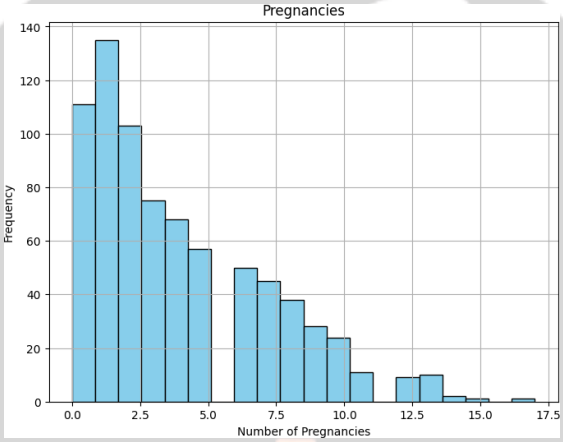
3.2 Data-Preprocessing

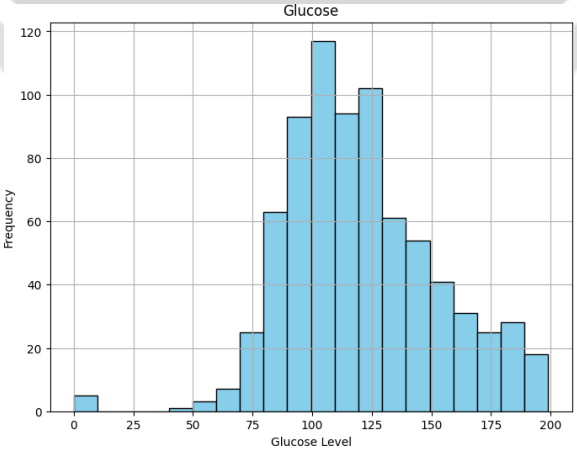
In the data preprocessing step, missing data must be addressed, and normalization or standardization must occur. To improve the quality and use of raw data for precise and insightful analysis.

1) Data Visualization and Data Cleaning

Initially, conducted an assessment for missing values.

Then, in order to examine the distribution of the data, we plotted several histograms using the features of the dataset.





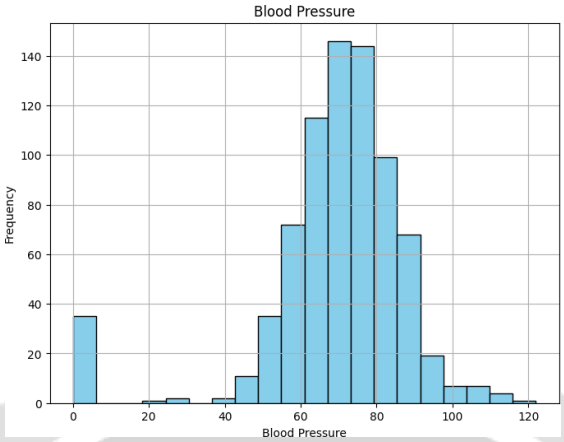
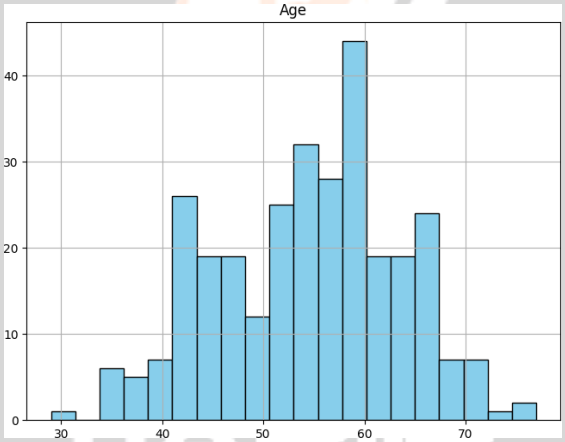
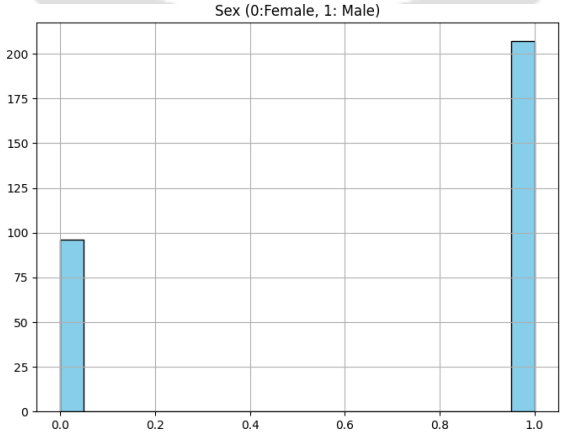


Figure 1. Pregnancies, Glucose and Blood Pressure Levels (Diabetes)





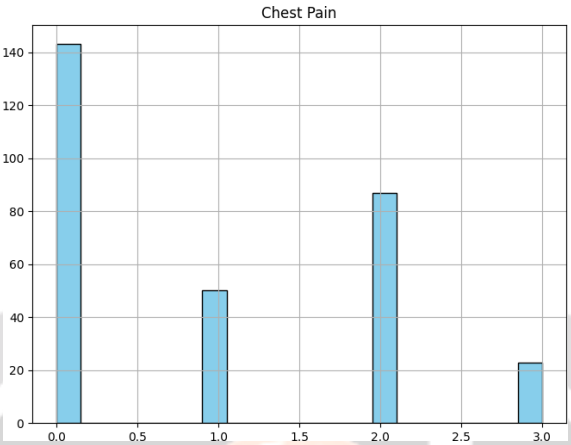
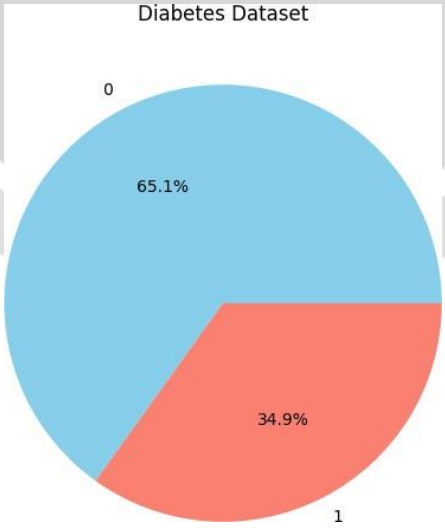


Figure 2. Age, Sex and Chest Pain (Heart Disease)

2) Assessing for Imbalances:

In our databases, we investigate whether an individual has a certain condition, such as diabetes or heart disease. A class imbalance simply suggests that one group (people without the condition) may have significantly more data than the other (people with the condition).



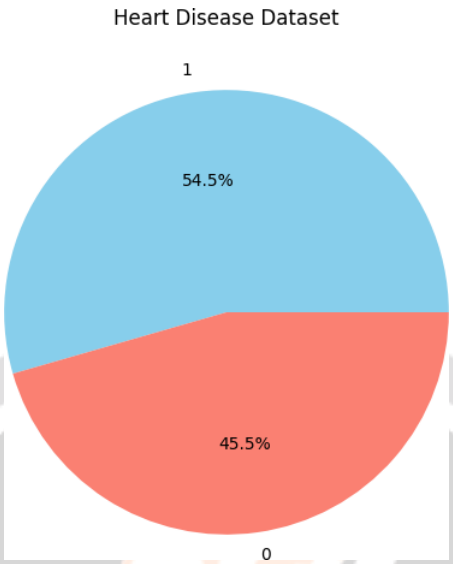


Figure 3. Imbalances

4. APPLIED ALGORITHMS

1) Logistic Regression:

Logistic regression is a widely used statistical method for problems involving binary classification, where the output variable may have two possible outcomes. According to our research, a variety of clinical and demographic factors can be used to predict a person's likelihood of developing heart disease using the logistic regression model. Similarly, for diabetes, logistic regression proves effective in utilizing clinical and demographic factors to predict the likelihood of diabetes.

The logistic regression approach utilizes a linear combination of predictor factors to assess the probability that a person has heart disease, including factors such as blood pressure, cholesterol levels, age, sex, and other relevant features. Similarly, for diabetes prediction, predictor variables such as BMI, blood sugar levels, and other relevant factors are considered in the logistic regression model.

Several metrics, including accuracy, precision, recall, and F1-score, were employed to evaluate the logistic regression model's performance in predicting heart disease and diabetes.

2) Support Vector Machine:

Support Vector Machine (SVM) is a popular supervised learning method that excels in handling categorization issues by determining the optimal hyperplane to separate data

points into different classes, thereby optimizing the margin between them. In our research, SVM has proven to be a valuable model for predicting both the likelihood of diabetes and heart disease based on a range of clinical variables.

The SVM model generates a decision boundary (hyperplane) in the feature space that separates individuals with diabetes and heart disease from those without. Features of the SVM model for diabetes include age, BMI, blood sugar levels, and other relevant factors. Features of the SVM model for heart disease prediction include blood pressure, cholesterol levels, age, sex, and other relevant factors.

The effectiveness of the SVM model was assessed using a variety of metrics, including F1- score, recall, accuracy, and precision in predicting heart disease and diabetes.

5. RESULTS AND ANALYSIS

5.1 Model Performance Metrics:

Evaluation metrics gauge the quality and effectiveness of machine learning models.

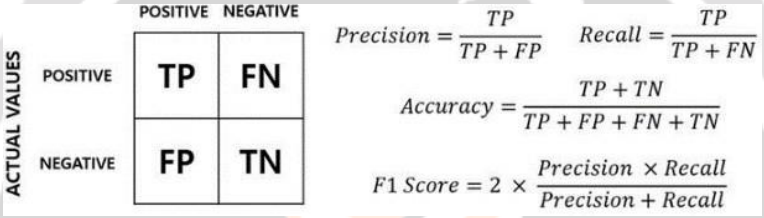


Figure 4. Evaluation Metrics

Diabetes

	Logisitc Regression	SVM
Accuracy	0.74	0.76
Precision	0.63	0.72
Recall	0.67	0.56
F1 Score	0.65	0.63

Table 3. Scores of applied algorithms (Diabetes)

The logistic regression model achieved an accuracy of 74%, with a precision of 63%, recall of 67%, and an F1-score of 65%. In contrast, the SVM model exhibited slightly superior performance, attaining an accuracy of 76%, precision of 72%, recall of 56%, and an F1-score of 63%.

Heart Disease

	Logisitc Regression	SVM
Accuracy	0.88	0.70
Precision	0.87	0.66
Recall	0.90	0.87
F1 Score	0.89	0.75

Table 4. Scores of applied algorithms (Heart Disease)

While the logistic regression model demonstrated strong performance with an accuracy of 88%, precision of 87%, recall of 90%, and an F1-score of 89%, the SVM model exhibited comparatively lower metrics, achieving an accuracy of 70%, precision of 66%, recall of 87%, and an F1-score of 75%.

5.2 Confusion Matrix:

A 'confusion matrix' is a tool that we use in our study to assess the performance of our prediction models. It enables us to determine whether or not our models are producing reliable estimations. We will examine in detail the confusion matrices for the two models that we employed to predict heart disease and diabetes. For both predictions, Support Vector Machine (SVM) and Logistic Regression were used. Each model was trained and assessed using different datasets, one for diabetes and the other for heart disease.

The confusion matrix has a square matrix structure, with rows and columns denoting the expected and actual class labels, respectively. In the matrix, each cell represents the number of cases that are categorized based on their true and expected labels. The diagonal elements reflect accurately identified cases, such as true positives (TP) and true negatives (TN). Off-diagonal elements are indicators of misclassified situations, such as false positives (FP) and false negatives (FN).

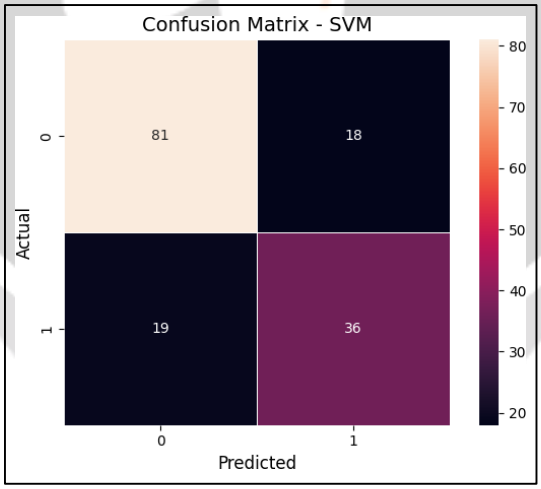


Figure 5. Diabetes Confusion Matrix (SVM)

1. The SVM model correctly identified diabetes in 81 cases, representing a true positive rate of 81/(81+19) = 81%.

2. False positives occurred in 18 cases, resulting in a false positive rate of 18/(18+36) = 33.3%.
3. The false negative rate, indicating the proportion of actual positive cases incorrectly identified as negative, was 19/(19+81) = 19%.
4. The true negative rate, representing the proportion of actual negative cases correctly identified, was 36/(36+18) = 66.7%.

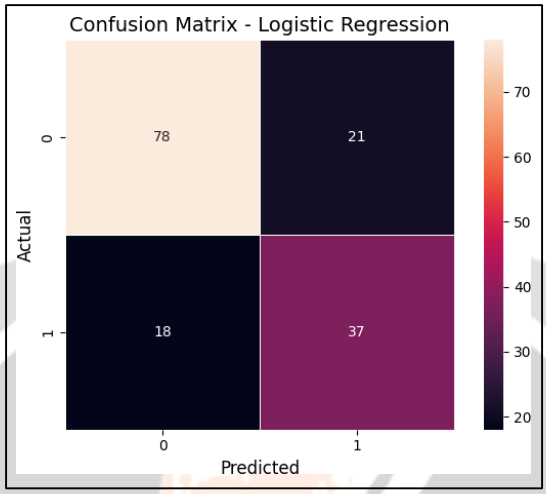


Figure 6. Diabetes Confusion Matrix (Logistic Regression)

1. The Logistic Regression model correctly identified diabetes in 78 cases, representing a true positive rate of 78/(78+18) = 81.25%.
2. False positives occurred in 21 cases, resulting in a false positive rate of 21/(21+37) = 36.21%.
3. The false negative rate, indicating the proportion of actual positive cases incorrectly identified as negative, was 18/(18+78) = 18.75%.
4. The true negative rate, representing the proportion of actual negative cases correctly identified, was 37/(37+21) = 63.79%

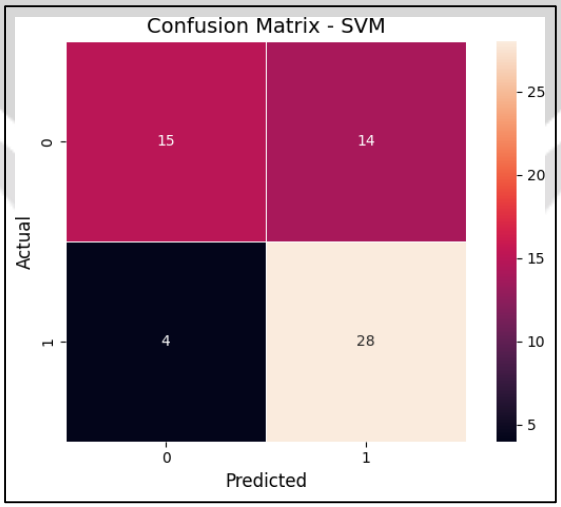


Figure 7. Heart Disease Confusion Matrix (SVM)

1. The SVM model correctly identified cases of heart disease in 15 instances, representing a true positive rate of 15/(15+4) = 78.95%.
2. False positives occurred in 14 cases, resulting in a false positive rate of 14/(14+28) = 33.33%.
3. The false negative rate, indicating the proportion of actual positive cases incorrectly identified as negative, was 4/(4+15) = 21.05%.
4. The true negative rate, representing the proportion of actual negative cases correctly identified, was 28/(28+14) = 66.67%.

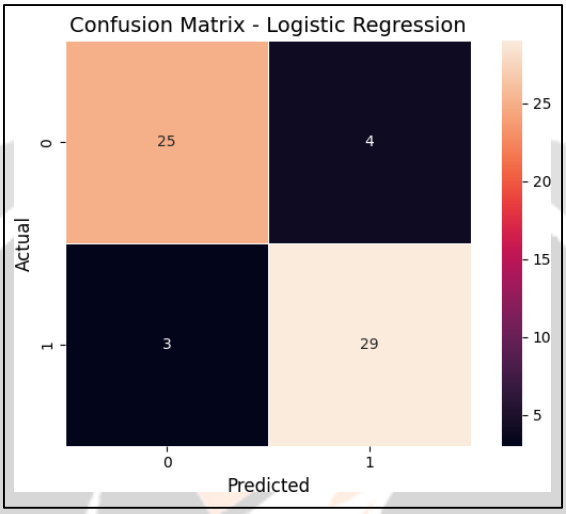


Figure 8. Heart Disease Confusion Matrix (Logistic Regression)

1. The Logistic Regression model correctly identified cases of heart disease in 25 instances, representing a true positive rate of 25/(25+3) = 89.29%.
2. False positives occurred in 4 cases, resulting in a false positive rate of 4/(4+29) = 12.12%.
3. The false negative rate, indicating the proportion of actual positive cases incorrectly identified as negative, was 3/(3+25) = 10.71%.
4. The true negative rate, representing the proportion of actual negative cases correctly identified, was 29/(29+4) = 87.88%.

6. LIMITATIONS

A limitation on our research is the dataset that we obtained from Kaggle. The dataset had shortcomings including missing data and possible biases, but it also offered insightful information about the relationship between predictor variables and the desired outcome. These limitations might have affected the accuracy of our analysis.

Furthermore, there's a chance that the Kaggle dataset's sample size was limited, which might have had an impact on the accuracy of the findings. The dataset may not accurately represent the diversity of the target population because it came from a single source. These factors should be considered while assessing the data because they may limit the accuracy of the results.

The logistic regression and SVM techniques that we utilized in our work for predictive modeling may have limits in regards to model flexibility and complexity when applied to the dataset. The modeling assumptions and hyperparameter selections may cause bias or have an impact on the reliability of the models. These techniques need to be kept in mind while interpreting the results.

It's important to be aware that the dataset we utilized could have unique features or restrictions. This implies that not everyone or every real-life scenario will benefit from the findings of our study. Future research could make use of alternative datasets or sources to see if our findings hold true for a larger population.

7. FUTURE SCOPE

Working with healthcare organizations and professionals to validate the models in clinical settings may help to facilitate the integration of predictive models into healthcare decision support systems. It would be essential to carry out future studies or clinical trials to assess the models' efficiency, usability, and impact on patient outcomes prior to their practical application.

Additional research could be done to optimize and improve the current prediction models in order to increase their accuracy, precision, and overall performance. This could involve looking into new machine learning approaches or adding new features or data sources in order to achieve even higher levels of accuracy than what we now have.

More understandable computer predictions could boost public trust in the application of these technologies. Individuals can understand why the machine is generating certain predictions by applying strategies such as prioritizing particular criteria or utilizing more basic models. Using the computer's advice, this can help individuals make more informed decisions.

8. CONCLUSION

In conclusion, our research has clarified the use of machine learning methods in the prognosis of diabetes and heart disease. Through the development and evaluation of predictive models, we have learned important things about the factors that contribute to these common health issues.

Our results show how well logistic regression and support vector machine (SVM) models predict the risk of heart disease and diabetes. By identifying those who are more prone to develop certain problems, our models enable proactive intervention by allowing preventive measures to be taken to limit the progression of these health conditions.

In summary, our results demonstrate the importance of applying machine learning methods for early disease detection and management, which pave the way for improved patient outcomes and the progress of public health.

REFERENCES

[1] Shuge Ouyang, “Research of Heart Disease Prediction Based on Machine Learning”, Published in: 2022 5th International Conference on Advanced Electronic Materials, Computers and Software Engineering (AEMCSE), 16 November 2022, DOI: 10.1109/AEMCSE55572.2022.00071.

[2] Krishna Battula, Raghupatruni Durgadinesh, Killi Suryapratap and Gedela Vinaykumar, “Use of Machine Learning Techniques in the Prediction of Heart Disease ” , Published in: 2021 International Conference on Electrical, Computer,

Communications and Mechatronics Engineering (ICECCME), 07-08 October 2021, DOI: 10.1109/ICECCME52200.2021.9591026.

[3] Rahul Katarya and Polipireddy Srinivas, "Predicting Heart Disease at Early Stages using Machine Learning: A Survey", Published in: 2020 International Conference on Electronics and Sustainable Communication Systems (ICESC), 02-04 July 2020, DOI: 10.1109/ICESC48915.2020.9155586.

[4] P. Sujatha and K. Mahalakshmi, "Performance Evaluation of Supervised Machine Learning Algorithms in Prediction of Heart Disease", Published in: 2020 IEEE International Conference for Innovation in Technology (INOCON), 06-08 November 2020, DOI: 10.1109/INOCON50539.2020.9298354.

[5] Aman Solanki, Anand Vardhan, Aman Jharwal and Prof. Narender Kumar I, "Heart Diseases Prediction Using Machine Learning", Published in: 2023 14th International Conference on Computing Communication and Networking Technologies (ICCCNT), 06-08 July 2023, DOI: 10.1109/ICCCNT56998.2023.10307839.

[6] Kundan Kumar and Arjit Tomar, "Diabetes Prediction System Using Machine Learning", Published in: 2023 International Conference on Advances in Computation, Communication and Information Technology (ICAICCIT), 23- 24 November 2023, DOI: 10.1109/ICAICCIT60255.2023.10466034.

[7] Bhavesh Rathi and Filipe Madeira, "Early Prediction of Diabetes Using Machine Learning Techniques", Published in: 2023 Global Conference on Wireless and Optical Technologies (GCWOT), 24-27 January 2023, DOI: 10.1109/GCWOT57803.2023.10064682.

[8] Vinod Jain, "Diabetes Prediction using Support Vector Machine, Naive Bayes and Random Forest Machine Learning Models", Published in: 2022 6th International Conference on Electronics, Communication and Aerospace Technology, 01-03 December 2022, DOI: 10.1109/ICECA55336.2022.100092413

[9] Srishti Mahajan, Pradeepta Kumar Sarangi, Ashok Kumar Sahoo and Mukesh Rohra, "Diabetes Mellitus Prediction using Supervised Machine Learning Techniques", Published in: 2023 International Conference on Advancement in Computation & Computer Technologies (InCACCT), 05-06 May 2023, DOI: 10.1109/InCACCT57535.2023.10141734.

[10] Lalit Mohan, Priyanka Suyal, Ravindra Singh Koranga, Bhawna Tewari and Amit Mittal, "Evaluation on Diabetes Care Using Machine Learning", Published in: 2023 14th International Conference on Computing Communication and Networking Technologies (ICCCNT), 06-08 July 2023, DOI: 10.1109/ICCCNT56998.2023.10306462.

[11] Narendra Mohan and Vinod Jain, "Performance Analysis of Support Vector Machine in Diabetes Prediction" , Published in: 2020 4th International Conference on Electronics, Communication and Aerospace Technology (ICECA), 05-07 November 2020, DOI: 10.1109/ICECA49313.2020.9297411