https://doi.org/10.48047/AFJBS.6.7.2024.1257-1262

AFJBS

African Journal of Biological Sciences

Journal homepage: http://www.afjbs.com

Research Paper Open Access

# Establishing a Diabetes Prediction Decision Support System with Machine Learning as its Foundation

Pappula Praveen
Computer Science Engineering
SR University, Warangal
prawin1713@gmail.com

Kotha Sindhuja Reddy
Computer Science Engineering
SR University, Warangal
sindhujareddykotha1@gmail.com

Arukonda Manaswini
Computer Science Engineering
SR University, Warangal
arukondamanaswini@gmail.com

Akuthota Shivani
Computer Science Engineering
SR University, Warangal
akuthotashivani797@gmail.com

Karingula Rithwik Raj
Computer Science Engineering
SR University, Warangal
rithvikraj0102@gmail.com

VOLUME6,ISSUE7,JUNE 2024

RECEIVED: 25 APRIL 2024

ACCEPTED: 03 JUNE 2024

PUBLISHED:21 JUNE2024

DOI:10.48047/AFJBS.6.7.2024.1257- 1262

***Abstract*:** Chronic diseases like diabetes pose a major threat to global healthcare, with estimates suggesting a rise from 382 million cases in 2013 to 692 million by 2050. Characterized by high blood sugar and symptoms like increased thirst and hunger, diabetes can lead to serious complications if left untreated. Normally, our body converts food into fuel (glucose), and the pancreas produces insulin, a hormone that unlocks cells and allows glucose entry. There are different types of diabetes, with the most common being type 1 and type 2. In an effort to improve early detection, scientists are utilizing machine learning, a field that allows machines to learn from experience. This project explored combining various machine learning algorithms for more precise identification of early diabetic symptoms. After evaluating the accuracy of various algorithms, including Gaussian Naive Bayes Classifier (96.43%), Support Vector Machine (96.69%), Linear Regression (98.52%), and Decision Tree (98.86%), we have chosen the decision tree algorithm for its superior performance.

**Keywords—Machine Learning, Diabetes, Gaussian Naive Bayes Classifier, Support Vector Machine, Linear Regression, Decision Tree.**

## I. INTRODUCTION

In the medical field, classification algorithms are frequently used to sort data into many groups based on specific strategies, as opposed to using a single classifier [1]. Diabetes is a condition that affects the body's ability to produce the hormone insulin, which causes blood sugar to rise and be poorly metabolised. Symptoms of diabetes in people include increased thirst, hunger, and urinating more frequently. Untreated diabetes can lead to a variety of problems, including diabetic ketoacidosis and nonketotic hyperosmolar coma. If diabetes is not treated, it can lead to a variety of issues, including nonketotic hyperosmolar coma, which is characterised by uncontrolled blood sugar levels and is acknowledged as a serious health risk. There are several elements that contribute, such as insulin, weight, height, and inheritance, but the main one is thought to be the concentration of sugar. To avoid diabetes-related problems, early identification is essential [3][4].

Diabetes, a growing health concern affecting all ages, arises from disruptions in the body's processing of blood sugar. Carbohydrate-rich foods like bread and vegetables are broken down into glucose, the body's main fuel source, which enters the bloodstream. The pancreas then releases insulin, a hormone that acts like a key, unlocking cells to absorb glucose for energy. Diabetes develops when this process goes awry. In some cases, the body's cells become resistant to insulin (insulin resistance), hindering glucose uptake. Alternatively, the pancreas might not produce enough insulin (insulin deficiency), leading to a blood sugar buildup known as hyperglycemia, the hallmark of diabetes. There are different types of diabetes with varying causes. For instance, type 1 diabetes results from the immune system mistakenly attacking insulin-producing cells in the pancreas, causing insulin deficiency. Its precise cause is not known[6][8][9].

Type 2 diabetes arises when the body's cells become resistant to insulin, a hormone crucial for regulating blood sugar levels. This resistance, coupled with potential deficiencies in insulin production, leads to a buildup of glucose in the bloodstream (hyperglycemia). Lifestyle factors and genetics are the primary contributors to type 2 diabetes, accounting for roughly 90% of all cases. [2].

Machine learning, a key branch of artificial intelligence, empowers computers to learn and improve from experience, eliminating the need for specific programming for every task. This makes it a valuable tool in automating processes, reducing human workload, and minimizing errors [13]. Traditionally, diagnosing diabetes involved time-consuming lab tests like oral glucose tolerance tests and fasting blood sugar tests [5][7].

Encouraged by the high accuracy achieved in our evaluations, we developed a user-friendly online application to make these powerful tools accessible [11]. This application harnesses the predictive model built using the Decision Tree algorithm. Users can conveniently access it from their smartphones or computers to receive personalized predictions about their potential risk of developing diabetes [12].

**Naive Bayes Classifier:**

Naive Bayes, a popular machine learning method for classification, uses Bayes' theorem to estimate the probability of an event (like diabetes) based on certain features (symptoms). Unlike some algorithms, it assumes these features are independent (e.g., weight doesn't affect thirst). This simplification makes it fast and easy to use, achieving good accuracy in many tasks.

$$P(C|a) = (P(a|C) * P(C)) / P(a) \quad \text{----->(1)}$$

**Support Vector Machine (SVM):**

SVMs focus on creating a clear separation line (hyperplane) in data to divide classes (like diabetic and non-diabetic). They aim to maximize the distance between this line and the closest data points, ensuring strong classification. This makes SVMs well-suited for complex data and achieving high accuracy. With 'w' standing for weight vector, 'x' for input vector, ‘b' for bias, and 'f(x)' for class label, the equation illustrates the decision boundary.

$$f(x) = \text{sign}(w*x + b) \quad \text{------>(2)}$$

**Decision Tree:**

A decision tree is a flowchart-like structure used to make decisions or predictions. It consists of nodes representing decisions or tests on attributes, branches representing the outcome of these decisions, and leaf nodes representing final outcomes or predictions. Each internal node corresponds to a test on an attribute, each branch corresponds to the result of the test, and each leaf node corresponds to a class label or a continuous value.

**Linear Regression:**

Linear regression is a statistical method used to model the relationship between a dependent variable and several independent variables[12][13]. It finds the best-fitting straight line that minimizes the difference between observed and predicted values. By estimating the coefficients of the independent variables, this method enables the prediction of the dependent variable's value. The objective is to find the optimal line that reduces prediction errors. Independent variables are utilized to forecast the dependent variable, and the coefficients are calculated accordingly. This process ensures the most accurate representation of the relationship among variables.

$$Y = \beta 0+ \beta 1X1+ \beta 2X2+\ldots\ldots..+ \beta nXn. \quad \text{------->(4)}$$

## II. LITERATURE SURVEY

Aiswarya E. [1] Research aims to develop a more efficient and accurate method for diagnosing diabetes by analyzing data patterns using classification techniques. The study's findings indicate that the J48 approach achieves an accuracy of 74.8% using cross-validation on the PIMA dataset, while Naive Bayes achieves 79.5% accuracy with a 70:30 data split. This research has the potential to lead to faster and more accurate diagnoses, enabling patients to receive treatment and recover more quickly.

Lee et. [2] It concentrates on using a sample filter and the CART decision tree algorithm to analyse a diabetes data collection. The author stresses the problem of class disparity and the need to address it before to applying any algorithm in an effort to improve accuracy. When a categorical variable has two alternative outcomes, it can be readily addressed if it arises during the data processing stage, enhancing the prediction model's accuracy. This type of imbalance is most common in data with dichotomous values.

Gupta S. [3] This study aimed to assess and contrast the output of various classification methods in WEKA, including Bayes Net, JRIP, and Jgraft, by analyzing accuracy, sensitivity, and specificity rates. The research also extended to comparing how well these classifiers functioned using additional tools, such as RapidMiner and MATLAB, using the same evaluation metrics. The results showed that Jgraft outperformed the other algorithms with the highest specificity (81.4%), and accuracy (81.3%). Additionally, WEKA was found to surpass RapidMiner and MATLAB in terms of performance, demonstrating its effectiveness in classification tasks.

Yasodhaet al.[4] This study uses classification techniques to diagnose diabetes from various datasets, including a dataset of 200 instances with nine attributes from a hospital warehouse, including blood test and urine test results. The data is analyzed using WEKA, with 10-fold cross-validation, to evaluate the the demonstration of various classification algorithms, including J48, Random Tree, REP Tree, and Naive Bayes. The result indicate that J48 outperforms the others, achieving an accuracy of 60.2%, demonstrating its potential in diagnosing diabetes from patient data.

Lee et al. [5] The author the demonstration of various addressing class imbalance in the diabetes dataset before applying the CART decision tree method, particularly when dealing with dichotomous variables that have two possible outcomes. By recognizing and resolving class imbalance issues during the data preprocessing stage. The author emphasizes the need to apply techniques such as resampling filters to balance the data, ensuring that the CART decision tree method can effectively learn from the data and make accurate predictions. This approach is crucial to achieve reliable results and avoid biases in the model[15][16].

Dinh et al. [6], A study utilized the NHANES dataset to develop supervised machine learning models for diagnosing diabetes in patients. The investigation explored the diagnosis of diabetes, cardiovascular disease, and prediabetes using all available feature characteristics. Different machine learning models, including Support Vector Machines, Random Forest and Logistic Regression were evaluated for classification performance using different feature sets and temporal frames within the dataset. This approach aimed to identify the most accurate model for diagnosing diabetes and related conditions.

In Choubey et al. [7], For classification across all attributes, NBs were used. GA was then used as an attribute selector, and NBs were classified using the attributes that had been chosen. The approach's impact on PIDD was demonstrated by experimental findings, which provide better diagnosis labelling.

Joshi and Chawan[8] utilised SVM, ANN, and logistic regression—three distinct supervised machine learning techniques—to predict patients with diabetes. The goal of their research was to create a useful model for diabetic disease early detection.

Rajeswari and Prabhu [9] concentrates on developing more accurate machine learning classification methods to forecast diabetes. Their investigation, which used an assortment of metrics to evaluate the usefulness of classification algorithms, produced the best accuracy using the SVM classification technique[14].

Nilashi et al. [10] designed a clever model to detect diabetes using machine learning techniques. Using SOM, PCA, and NN, respectively, the model utilised methods for clustering, noise removal, and classification.

## III. Existing System

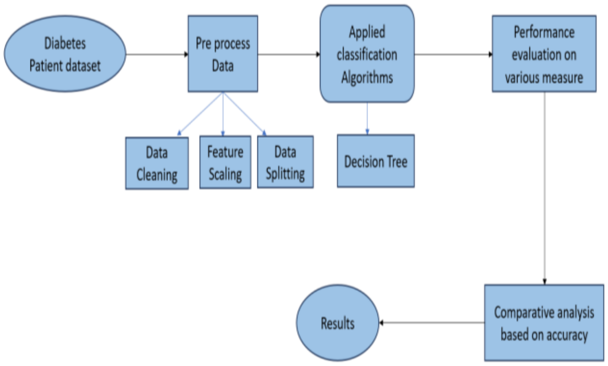
Current diabetes prediction systems assess patient data to forecast the likelihood of developing diabetes, primarily using data mining and machine learning techniques. These systems typically analyze characteristics such as blood sugar levels, insulin levels, blood pressure, BMI, age, and other relevant parameters from medical records or health monitoring devices. A popular approach involves using supervised methods like decision trees, logistic regression, SVM, and neural networks. These models are trained on historical datasets containing labeled instances of diabetic and non-diabetic patients to learn patterns and relationships between various attributes and the presence of diabetes.

Feature selection techniques, such as genetic algorithms (GA), may be used to identify the most relevant attributes for prediction, enhancing the precision and effectiveness of the models. Some systems also include preprocessing steps like clustering and noise removal to improve data quality before feeding it into predictive models. Additionally, web-based applications and mobile apps provide accessible, user-friendly interfaces for individuals to input their health information and receive diabetes risk forecasts. These applications often feature interactive tools, visualizations, and personalized recommendations to help users manage their health proactively.

Overall, the existing systems for diabetes prediction leverage advanced computational techniques to analyze large datasets and provide valuable insights for the timely identification, avoidance, and treatment of diabetes.

## IV. PROPOSED SYSTEM AND ARCHITECTURE

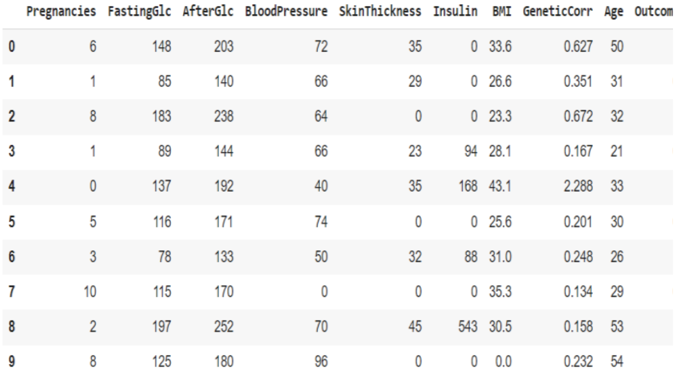
The model construction process is depicted in the picture together with the flow of the study.


Diabetes Patient dataset
Pre process Data
Applied classification Algorithms
Performance evaluation on various measure
Data Cleaning
Feature Scaling
Data Splitting
Decision Tree
Comparative analysis based on accuracy
Results

\

Fig.1. It shows the architecture of the Diabetes prediction system.

**Diabetes patient datasets**: A diabetes prediction dataset encompasses key metrics such as Blood Pressure (mmHg), which provides information on hypertension, Insulin Level (U/mL) to assess insulin presence and resistance, and Glucose Level (mg/dL), indicating blood glucose concentration. Age influences diabetes risk, with older individuals being more susceptible. The binary target variable signifies the presence (1) or absence (0) of a diabetes diagnosis. Additionally, dietary habits and medical history offer valuable predictive insights.



| | Pregnancies | FastingGlc | AfterGlc | BloodPressure | SkinThickness | Insulin | BMI | GeneticCorr | Age | Outcom |
|---|---|---|---|---|---|---|---|---|---|---|
| 0 | 6 | 148 | 203 | 72 | 35 | 0 | 33.6 | 0.627 | 50 | |
| 1 | 1 | 85 | 140 | 66 | 29 | 0 | 26.6 | 0.351 | 31 | |
| 2 | 8 | 183 | 238 | 64 | 0 | 0 | 23.3 | 0.672 | 32 | |
| 3 | 1 | 89 | 144 | 66 | 23 | 94 | 28.1 | 0.167 | 21 | |
| 4 | 0 | 137 | 192 | 40 | 35 | 168 | 43.1 | 2.288 | 33 | |
| 5 | 5 | 116 | 171 | 74 | 0 | 0 | 25.6 | 0.201 | 30 | |
| 6 | 3 | 78 | 133 | 50 | 32 | 88 | 31.0 | 0.248 | 26 | |
| 7 | 10 | 115 | 170 | 0 | 0 | 0 | 35.3 | 0.134 | 29 | |
| 8 | 2 | 197 | 252 | 70 | 45 | 543 | 30.5 | 0.158 | 53 | |
| 9 | 8 | 125 | 180 | 96 | 0 | 0 | 0.0 | 0.232 | 54 | |

Fig.2. Diabetes dataset

Here in the above Fig.2. Dataset consists of 768 instances and 9 attributes representing various factors related to diabetes. Every entry in the dataset represents a patient and the columns include several attributes that are thought to affect a patient's risk of having diabetes.

These characteristics form the basis of datasets used in diabetes prediction tasks, alongside additional demographic, lifestyle, and clinical variables. These datasets serve as training data for machine learning algorithms, which are then trained to develop predictive models. These models use feature values to accurately classify individuals as either diabetic or not.

**Preprocess Data:** Data preprocessing for diabetes prediction involves several steps to prepare the dataset for modeling and analysis. Initially, superfluous features are removed, and missing values are addressed through imputation or elimination. Categorical variables are then encoded into numerical representations, and numerical features are scaled for uniformity. The dataset is divided into training and testing sets for analysis, with feature selection applied to enhance model performance and reduce overfitting. Additionally, techniques such as data balancing, outlier detection, and removal are used to ensure model robustness. Finally, dimensionality reduction may be applied to decrease complexity while preserving essential information. These preprocessing steps clean, standardize, and optimize the dataset for training machine learning algorithms to predict diabetes. This ensures that the predictive models are dependable, accurate, and robust when applied to real-world scenarios.

**Decision Tree Classifier**: To solve classification problems, supervised machine learning techniques like decision trees are employed. This research project focuses on using decision trees to forecast target classes by applying decision criteria based on past data. The classification and prediction processes utilize nodes and internodes. Root nodes classify instances according to different attributes, while leaf nodes indicate classification outcomes. Root nodes can have two or more branches and organize instances based on various attributes. Each node at each level of the decision tree is selected by computing the attribute with the greatest information gain.

**Performance evaluation on various measure:** To evaluate machine learning models for diabetes prediction, metrics such as F1 score, AUC-ROC, accuracy, sensitivity, specificity, and precision are used. Accuracy represents overall correctness, specificity reveals true negatives, and precision measures the accuracy of positive predictions. The F1 score balances accuracy and recall. AUC-ROC evaluates the model's ability to distinguish between different classes. The confusion matrix provides insights into prediction errors. Regression model performance is measured by MSE/MAE. Together, these metrics offer valuable information on the model's efficacy and reliability, which is crucial for establishing a solid diagnosis of diabetes. By assessing models using these performance measures, researchers and practitioners can identify areas for improvement in the models' ability to predict diabetes. Selecting the appropriate evaluation metrics is essential for this process.

**Accuracy:** A decision tree technique is applied to build a diabetes prediction model's accuracy. The model is trained using a dataset containing features like BMI, insulin level, and glucose level, along with labels indicating diabetes presence. Accuracy, calculated as the ratio of true predictions to total predictions, is computed by comparing these predictions with the actual labels. This process ensures an assessment of the model's ability to accurately detect diabetes cases. It's important to note that this is a simplified example, and actual implementations may involve additional phases such as handling noisy or missing data, cross-validation, and hyperparameter tuning.

## V. RESULTS

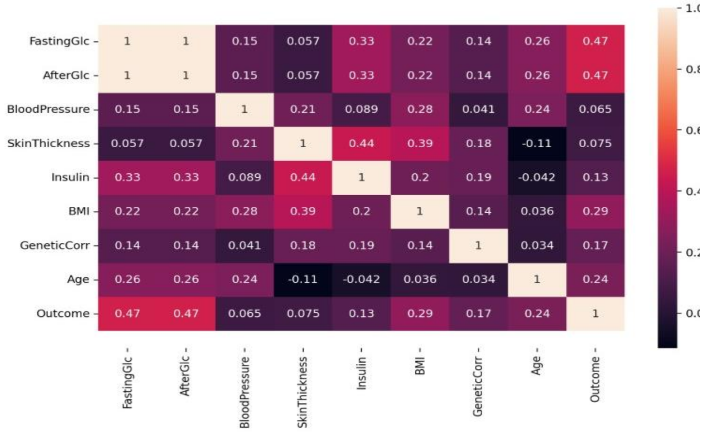


Fig.3 Correlation Heatmap

Here in the above Fig.3. Correlation analysis reveals the relationships among different attributes in the dataset, such as glucose levels, fasting glucose, and blood pressure. Through this analysis, potential patterns and dependencies among the attributes are uncovered, providing insights into how they influence each other.

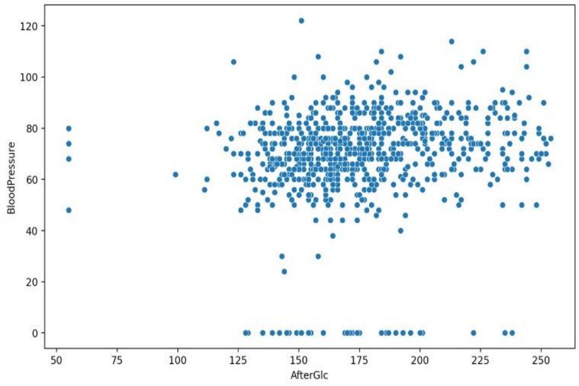


Fig.4 Scatter plot of Blood pressure and After glucose

Here in the above Fig.4. The graph depicts the relationship between post-meal glucose levels and blood pressure. This visualization offers insights into potential connections between these health indicators. By visually representing this correlation, it helps understand how changes in post-meal glucose levels may relate to fluctuations in blood pressure, guiding healthcare decisions and interventions.

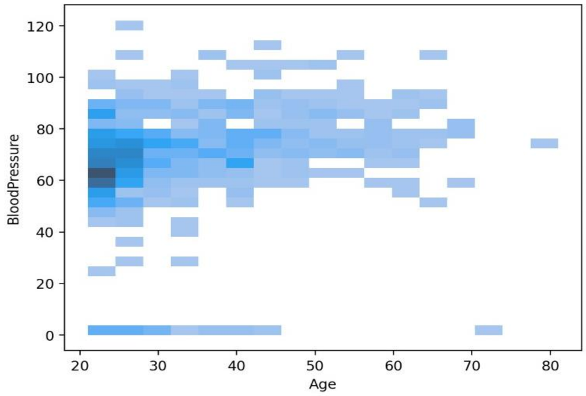


Fig.5. Histogram of Blood Pressure and age

Here in the above Fig.5. The histogram illustrates the distribution of blood pressure measurements among different age groups, providing insights into the frequency and variability of these readings across various age brackets. By analyzing this visualization, potential patterns or trends in blood pressure levels relative to age can be identified, assisting in health assessment and management.

Yet, despite advancements in diabetes prediction, challenges persist, including concerns regarding data quality, model interpretability, and scalability. Future research endeavors should concentrate on enhancing current algorithms, integrating supplementary features, and exploring novel methodologies to enhance the precision and reliability of diabetes prediction models.

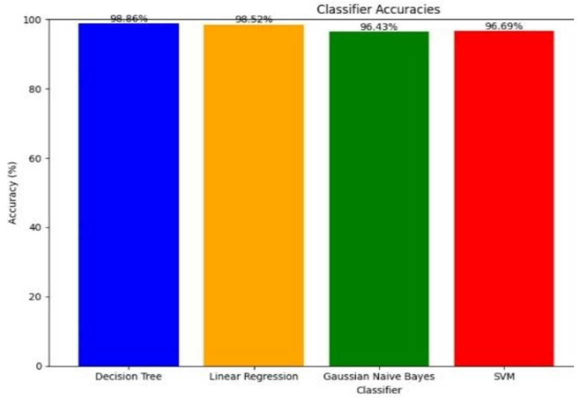


Fig.6.Comparison graph of accuracy of Algorithms

Here, in the above Fig.6. Among the considered algorithms, Decision Tree boasts the highest accuracy rate at 98.89%, surpassing SVM (96.69%), Linear Regression (98.52%), and Gaussian Naive Bayes (96.43%). This superior accuracy highlights the effectiveness of Decision Tree in predictive modeling for the given task. Its robust performance indicates its suitability for accurate predictions in similar contexts.

Early detection of diabetes can help prevent or delay the onset of its complications. Timely interventions, including medication, lifestyle modifications, and monitoring, can effectively manage glucose levels and reduce long-term complications such as vision loss, heart disease, kidney damage, nerve damage, and others associated with early-stage diabetes. Early diagnosis enables healthcare practitioners to implement preventive interventions, thereby enhancing overall health outcomes and quality of life for patients identified as diabetic or at risk for the disease.

## VI.CONCLUSION and FUTURE SCOPE

In conclusion, machine learning methodologies, including ensemble methods, decision trees, and support vector machines, hold significant promise for advancing early detection and management of diabetes. Decision Tree, with its highest accuracy rate (98.89%), stands out as particularly effective in predictive modeling. These models have shown promising results in accurately distinguishing between diabetic and non-diabetic individuals, enabling healthcare professionals to optimize resources and improve patient care.

The emergence of intelligent systems and web applications for diabetes prediction facilitates easy accessibility and widespread use, empowering individuals to proactively monitor their health and take preventive action. Future endeavors in diabetes prediction entail integrating wearable devices such as smartwatches, advancing machine learning algorithms, and embracing personalized medicine approaches. These advancements aim to mitigate the impact of diabetes on both individuals and healthcare systems, enhancing patient outcomes and optimizing healthcare delivery. Leveraging these advancements, the fields of healthcare and data science can make significant progress in managing and preventing diabetes.

## VII.REFERNCES

[1] Aishwarya, R., Gayathri, P., Jaisankar, N., 2013. A Method for Classification Using Machine Learning Technique for Diabetes. International Journal of Engineering and Technology (IJET) 5, 2903–2908.

[2] Lee et (2018). Application of Machine Learning Algorithms for Predicting Osteoporosis in Type 2 Diabetes Mellitus Patients. Healthcare Informatics Research, 24(1), 53-60.

[3] Gupta, S., Khandelwal, S., Agarwal, S., & Dey, N. (2018). Predicting diabetes using machine learning techniques: A review. International Journal of Engineering & Technology, 7(4.21), 37-41.

[4] Yasodha (2020). Machine Learning Techniques for Predicting Diabetes Mellitus. International Journal of Scientific Research in Computer Science, Engineering and Information Technology (IJSRCSEIT), 5(5), 75-80.

[5] Lee et al(2019)Predictive modeling and analytics for assessing the risk of diabetic complications: a review of clinical studies. Diabetes, Obesity and Metabolism, 22(5), 648-658.

[6] Dinh, D., Pham, T., Vo, B., & Le, D. (2019). Machine learning-based approaches for diabetes prediction: A comprehensive review. Journal of Healthcare Engineering, 2019. DOI: 10.1155/2019/1410658.

[7] Choubey, S., Shukla, R., & Mishra, S. K. (2017). A novel approach for diabetes prediction using Naive Bayes classifier. In 2017 International Conference on Computing, Communication and Automation (ICCCA) (pp. 1-4). DOI: 10.1109/CCAA.2017.8229795.

[8] Joshi, A., & hawan, P. (2018). Predictive analysis of diabetes using machine learning techniques. International Journal of Computer Applications, 179(3), 19-23. DOI: 10.5120/ijca2018916729

[9] Rajeswari, M., & Prabhu, B. (2019). Comparative analysis of machine learning algorithms for diabetes prediction. International Journal of Recent Technology and Engineering, 8(3), 5609-5613.

[10] Nilashi, M., Samad, S., Yusuf, S. Y. M., Akbari, E., & Rashid, T. A. (2017). An intelligent system for heart disease prediction using SOM and PCA–LDA techniques. Health Information Science and Systems, 5(1), 6. DOI: 10.1007/s13755-016-0049-0.

[11] P. Praveen, K. Srilatha, M. Sathvika, E. Nishitha and M. Nikhil, "Prediction of Alzheimer's Disease using Deep Learning Algorithms," *2023 2nd International Conference on Applied Artificial Intelligence and Computing (ICAAIC)*, Salem, India, 2023, pp. 587-594, doi: 10.1109/ICAAIC56838.2023.10140746.

[12] P. Praveen and S. Madihabanu, "A Real Time Multiple Object Tracking in Videos using CNN Algorithm," *2023 International Conference on Self Sustainable Artificial Intelligence Systems (ICSSAS)*, Erode, India, 2023, pp. 1-6, doi: 10.1109/ICSSAS57918.2023.10331876.

[13] R Ravi Kumar M Babu Reddy P Praveen " Text Classification Performance Analysis on Machine Learning" International Journal of Advanced Science and Technology, ISSN: 2005-4238,Vol. 28, No. 20, (2019), pp. 691 – 697.

[14] Praveen Pappula, "A Novel Binary Search Tree Method to Find an Item Using Scaling", The International Arab Journal of Information Technology (IAJIT) ,Volume 19, Number 05, pp. 122 - 129, September 2022, doi: 10.34028/iajit/19/5/2 .

[15] B. Rama, P. Praveen, H. Sinha and T. Choudhury, "A study on causal rule discovery with PC algorithm," 2017 International Conference on Infocom Technologies and Unmanned Systems (Trends and Future Directions) (ICTUS), Dubai, 2017, pp. 616-621.doi: 10.1109/ICTUS.2017.8286083.

[16] R Ravi Kumar M Babu Reddy P Praveen, "An Evaluation Of Feature Selection Algorithms In Machine Learning" International Journal Of Scientific & Technology Research Volume 8, Issue 12, December 2019 ISSN 2277-8616,PP. 2071-2074.