Diabetes Risk Prediction Using Feature Selection Algorithms and Advanced Machine Learning Models

Deepanshu Goyal[1], Jasjit Singh[2], and Apurva Vashist[2]

[1]Department of Computer Science Ambedkar, DSEU Shakarpur
[2]Department of Computer Science Ambedkar DSEU Shakarpur

February 27, 2025

Abstract

Diabetes is a persistent metabolic disorder that impacts millions globally, presenting a significant health challenge worldwide with increasing prevalence and significant healthcare implications. Early and accurate prediction of diabetes can aid in timely intervention and disease management. This study investigates the efficacy of machine learning algorithms in predicting diabetes using the Sylhet Diabetes Hospital dataset, which consists of clinical records from 520 patients. Various feature selection methodologies, including Pearson correlation analysis, Genetic Algorithm, Chi-Square test, and Recursive Feature Elimination (RFE), were employed to identify the most biologically significant predictors associated with diabetes onset. Five machine learning models-Support Vector Machine (SVM), K-Nearest Neighbours (KNN), Decision Tree (DT), Random Forest (RF), and Logistic Regression (LR)-were trained and validated through cross-validation techniques. Among these techniques, the Random Forest algorithm exhibited the highest predictive performance, achieving an accuracy of 94.70% and an Area Under the Receiver Operating Characteristic Curve (AUC-ROC) score of 98.22%, indicating its superior ability to differentiate between diabetic and non-diabetic cases. Making it the most Dependable for diabetes prediction. These outcomes highlight the potential of machine learning in enhancing diabetes diagnosis and risk assessment. Future work includes integrating deep learning techniques and expanding datasets to improve generalizability and predictive performance.

Diabetes Risk Prediction Using Feature Selection Algorithms and Advanced Machine Learning Models

Deepanshu Goyal [1] | **Jasjit Singh [2]** | **Apurva Vashist[3]**

[1]Department of Computer Science

Ambedkar DSEU Shakarpur Campus – 1

New Delhi – 110092 , India

goyaldeepanshu482@gmail.com

[2]Department of Computer Science

Ambedkar DSEU Shakarpur Campus – 1

New Delhi – 110092 , India

Singhjasjit15@gmail.com

[3]Department of Computer Science

Ambedkar DSEU Shakarpur Campus – 1

New Delhi – 110092 , India

apurva.vashist@gmail.com

Abstract

Diabetes is a persistent metabolic disorder that impacts millions globally, presenting a significant health challenge worldwide with increasing prevalence and significant healthcare implications. Early and accurate prediction of diabetes can aid in timely intervention and disease management. This study investigates the efficacy of machine learning algorithms in predicting diabetes using the Sylhet Diabetes Hospital dataset, which consists of clinical records from 520 patients. Various feature selection methodologies, including Pearson correlation analysis, Genetic Algorithm, Chi-Square test, and Recursive Feature Elimination (RFE), were employed to identify the most biologically significant predictors associated with diabetes onset. Five machine learning models—Support Vector Machine (SVM), K-Nearest Neighbours (KNN), Decision Tree (DT), Random Forest (RF), and Logistic Regression (LR)—were trained and validated through cross-validation techniques. Among these techniques, the Random Forest algorithm exhibited the highest predictive performance, achieving an accuracy of 94.70% and an Area Under the Receiver Operating Characteristic Curve (AUC-ROC) score of 98.22%, indicating its superior ability to differentiate between diabetic and non-diabetic cases.

Making it the most Dependable for diabetes prediction. These outcomes highlight the potential of machine learning in enhancing diabetes diagnosis and risk assessment. Future work includes integrating deep learning techniques and expanding datasets to improve generalizability and predictive performance.

Keywords

Diabetes Prediction, Machine Learning, Feature Selection, Random Forest, Sylhet Diabetes Hospital Dataset, Genetic Algorithm, Chi-Square Method, Recursive Feature Elimination (RFE), Pearson Correlation, Cross-Validation, Classification Models.

1.Introduction:

Diabetes is a long-term metabolic condition that impairs the body's ability to maintain stable blood sugar levels. It encompasses various forms of diabetes, each necessitating distinct management strategies. If inadequately controlled, persistent hyperglycemia can precipitate severe complications, such as cardiovascular disease, cerebrovascular events, nephropathy, neuropathy, and progressive vision impairment. The worldwide predominance of sort 2 diabetes has been consistently rising, posturing a critical open wellbeing challenge. This condition could be a major supporter to different complications, counting maladies influencing the eyes, kidneys, heart, and lower appendages. It not as it were lessens quality of life but moreover increments the hazard of untimely passing.

Right now, diabetes influences around 537 million people matured 20–79, ...with projections indicating an increase to 0.578 billion cases by 2030 and 0.7 billion by 2045.. The far-reaching predominance of diabetes forces a significant monetary burden on worldwide healthcare frameworks, essentially expanding per capita therapeutic costs.

In 2021 alone, diabetes was dependable for 6.7 million passings, making it the ninth driving cause of mortality around the world. Gauges recommend that within the same year, the infection accounted for around 747,000 passings and caused an amazing $10 billion in treatment costs.[1]

Despite the findings of the World Health Organization (WHO), nearly 71% of annual deaths are attributed to non-communicable diseases. Among these, diabetes greatly elevates the risk of severe complications, including strokes and cardiovascular diseases. Given its widespread impact, diabetes has emerged as a major global healthcare crisis.

In 2019, approximately 463 million people worldwide were affected by diabetes, representing 8.8% of the global adult population. The disease is not only a growing concern among adults but also poses a severe threat to children and adolescents, who face heightened risks of fatal complications. Research indicates that the prevalence of diabetes is similar among men and women, with projections suggesting a continued rise in cases. The healthcare sector encounters major challenges, such as managing electronic health records, enhancing computer-aided diagnosis, and achieving accurate disease prediction. To mitigate these issues and reduce healthcare costs, there is an increasing shift toward personalized medicine. Machine learning has evolved into a powerful tool, leveraging advanced algorithms and predictive models to enhance diabetes diagnosis and management. Recent studies highlight the high accuracy of machine learning algorithms in diabetes prediction, establishing them as a crucial asset for early detection and preventive care.

Common Side Effects of diabetes:

Common symptoms of diabetes include increased thirst, frequent urination, intense hunger, and unintended weight loss. Other nonspecific signs may also appear, such as fatigue, blurred vision, sweet-smelling urine or semen, and genital irritation due to Candida infections. Notably, approximately half of affected individuals may remain asymptomatic. Type 1 diabetes typically presents suddenly following a preclinical phase, whereas type 2 diabetes has a more insidious onset, often remaining asymptomatic for several years.

1.1 Types of diabetes:

A. Type 1 Diabetes

Also referred to as juvenile diabetes, type 1 diabetes arises when the body ceases insulin production, the hormone essential for blood sugar regulation. While it is commonly diagnosed in childhood or adolescence, it can also develop in adulthood. Managing type 1 diabetes requires lifelong insulin therapy, delivered through injections or an insulin pump.[1]

B. Type 2 Diabetes

Type 2 diabetes is the most common form of the disease, strongly linked to obesity and lifestyle factors. It occurs when the body either does not produce enough insulin or becomes resistant to its effects. Unlike type 1 diabetes, type 2 diabetes can sometimes be managed without insulin through medication, dietary changes, and regular exercise. However, in instances where blood glucose levels remain elevated despite these measures, insulin therapy may be required. While traditionally considered a condition of adulthood, type 2 diabetes can manifest at any stage of life, including in children and adolescents, largely due to the increasing prevalence of obesity. [1]

C. Gestational Diabetes

This sort of diabetes happens amid pregnancy and as a rule goes absent after childbirth.

This form of diabetes manifests during pregnancy and typically resolves following childbirth. However, women who develop gestational diabetes are at an increased risk of progressing to type 2 diabetes later in life. Effective management through regular monitoring, adherence to a well-balanced diet, and consistent physical activity can help regulate blood glucose levels throughout pregnancy.[1]

Diabetes is not only a concern for adults but also affects children and teenagers, increasing their risk of serious health complications. While some types of diabetes are influenced by diet and physical inactivity, others, like type 1 diabetes, develop due to genetic and autoimmune factors.

Managing diabetes requires a combination of lifestyle changes, medications, and continuous blood sugar monitoring. Advances in medical research, including machine learning and artificial intelligence, are helping improve diabetes prediction and treatment. Early diagnosis and effective management can substantially mitigate the risk of

complications, promoting better health outputs and an improved quality of life for individuals affected by diabetes.

Diagnosis:

Diabetes mellitus [23] is diagnosed through the assessment of blood glucose levels, with confirmation based on any of the following criteria:

1. Fasting Plasma Glucose (FPG) Level ≥ 7.0 mmol/L (126 mg/dL)

- This test requires a blood sample collected after a fasting period, typically in the morning before breakfast, ensuring the patient has fasted overnight or for a minimum of eight hours prior to the test.

2. Plasma Glucose Level ≥ 11.1 mmol/L (200 mg/dL) Two Hours Post-Oral Glucose Tolerance Test (OGTT)

- This measurement is taken after administering a 75-gram oral glucose dose to assess the body's capacity to manage blood sugar levels.

3. Symptoms of hyperglycemia, along with a plasma glucose level of ≥ 11.1 mmol/L (200 mg/dL), regardless of fasting status.

- Diagnosis may also be established in individuals exhibiting clinical symptoms of elevated blood sugar levels, whether fasting or non-fasting, with plasma glucose concentrations meeting or exceeding the threshold of 11.1 mmol/L (200 mg/dL).

Previous studies have investigated various factors associated with the onset of diabetes, especially in predictive research utilizing machine learning (ML). As a specialized field within artificial intelligence (AI), ML enables the creation of predictive models that continuously improve their accuracy through data-driven learning, eliminating the need for explicit programming of predefined tasks. One of ML's greatest strengths is its ability to identify hidden risk factors that may contribute to diabetes by analyzing large datasets with enhanced learning capabilities. Unlike traditional statistical models, which struggle with non-linear data, ML algorithms can efficiently handle such complexities, leading to more precise and reliable predictions.

In this study, we utilized multiple ML models, including Support Vector Machine (SVM), Logistic Regression (LR), Random Forest, K-Nearest Neighbors (KNN), and Decision Tree, incorporating cross-validation techniques to enhance the reliability of our predictions. The models were assessed using key performance metrics such as accuracy, precision, F1-score, and the ROC-AUC curve. For training and validation, we leveraged real-world patient data obtained from the Sylhet Diabetes Hospital [3].

This dataset contains 520 records and 18 columns that provide information about different symptoms and characteristics of individuals such as: Age, Gender, Polyuria, Polydipsia, Blurred Vision, Obesity, Rapid Weight Loss, Weakness etc. The most important column is Result, which indicates whether the person has diabetes (Positive) or not (Negative).

To understand the above-mentioned dataset, we created the following graphical representation:

Figure 1: Pie Chart – Percentage of Diabetes Cases by Gender

This pie chart shows the percentage of Male and Female people with diabetes. The blue area represents males and the red area represents females.

Observations:

The graph shows that a larger percentage of men in the dataset have been diagnosed with diabetes compared to women.

This suggests that the prevalence of diabetes may be slightly higher in men. However, further analysis would be needed to confirm this trend.

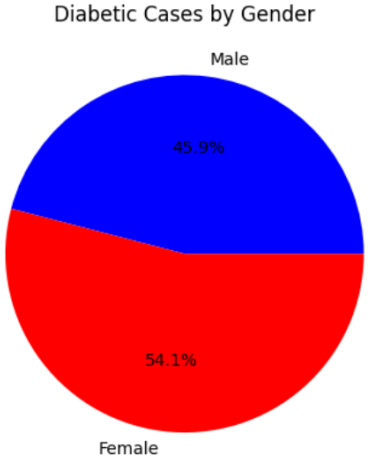


Figure 2: Bar Chart – Distribution of Diabetes Cases by Gender

This bar chart compares the amount of Diabetic (positive) cases and non-Diabetic (negative) cases in both men and women.

Observations:

The red bars represent non-diabetic individuals and the green bars represent diabetic individuals. There are more diabetic individuals than non-diabetic individuals in both men and women, so the dataset may consist primarily of individuals at high risk for diabetes.

There appear to be more diabetic cases in men compared to women in this dataset.

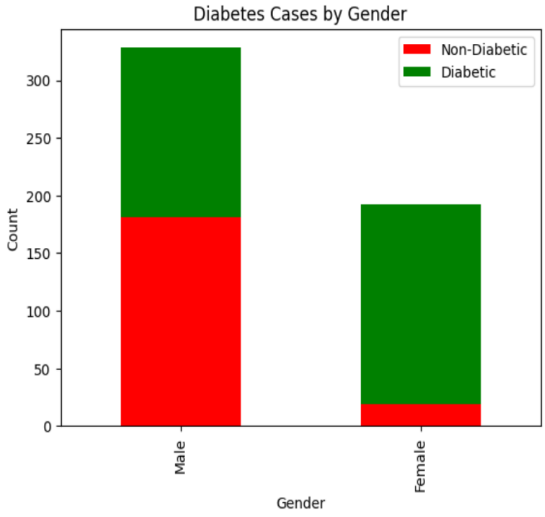


Figure 3: Histogram - Age distribution of diabetics and non-diabetics

This histogram shows the age distribution of diabetics (red) and non-diabetics (green).

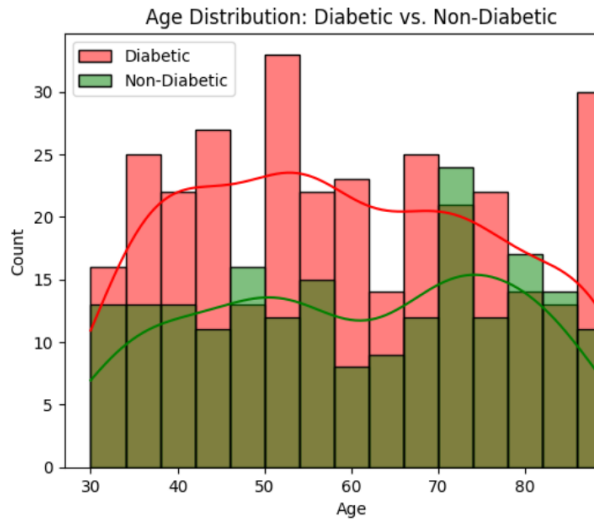
Observations:

The red peak shows the age group where diabetes is most prevalent.

The green peak represents people without diabetes.

The age distribution is slightly different:

Most cases of diabetes seem to occur between 40 and 60 years of age, suggesting that middle-aged people are more commonly affected.



2.Literature Review:

N. Kushal Kumar Raju and Keshav Krishnamurthy [4] utilized machine learning techniques and an ensemble approach to improve early diabetes prediction. They implemented four distinct algorithms—Support Vector Machine (SVM), K-Nearest Neighbors (KNN), Logistic Regression, and Random Forest Classifier—conducting a comparative analysis to determine the most accurate model. Their findings revealed that the Random Forest Classifier achieved the highest accuracy of 98.08%. As a result, they incorporated this optimal model into a web application aimed at assessing an individual's risk of developing diabetes.

Malathy S. and Santhiya M. [5] aimed to develop a highly accurate diabetes prediction system using various machine learning models. Their research focused on constructing a predictive model for diabetes occurrence by employing classification techniques such as Naïve Bayes, Support Vector Machine (SVM), and Artificial Neural Networks (ANN). Their results demonstrated that the neural network algorithm outperformed the other models. To enhance the robustness and reliability of their system, they implemented their project using the R programming language.

Israt Jahan Kakoly and Rakibul Hoque [1] carried out a study to predict diabetes risk factors using machine learning techniques. They enhanced predictive accuracy by employing feature selection methods such as Principal Component Analysis (PCA) and Information Gain. The study evaluated the performance of five machine learning models—Decision Tree, Random Forest, Support Vector Machine (SVM), Logistic Regression, and K-Nearest Neighbors (KNN). Their results showed a classification accuracy of over 82.2% and an Area Under the ROC Curve (AUC) of 87.2%. Ultimately, the study concluded that clinical risk factors are the most reliable predictors of diabetes, with dietary determinants following in significance.

Nidhi Kumari and Madhu Gautam [6] conducted extensive research highlighting the potential of machine learning algorithms to enhance disease detection while minimizing medical errors, ultimately contributing to improved patient outcomes. Their study addressed challenges such as outliers and missing values caused by class imbalance within the dataset. The primary objective was to develop a precise and resilient approach for diabetes prediction and classification. Using the Pima Indian Diabetes (PID) dataset from Kaggle,

they implemented five machine learning algorithms—Support Vector Machine (SVM), Naïve Bayes (NB), Random Forest (RF), K-Nearest Neighbors (KNN), and Decision Tree (DT). Their findings indicated that the Random Forest algorithm achieved the highest predictive accuracy, reaching 92%.

Arwatki Chen Lyngdoh and Nurul Amin Choudhary [7] applied machine learning techniques such as K-Nearest Neighbors (KNN), Naïve Bayes (NB), Decision Tree (DT), Random Forest (RF), and Support Vector Machine (SVM) for diabetes prediction. Their study revealed that the KNN classifier achieved the highest and most stable accuracy of 76%, while the other models consistently maintained an accuracy of approximately 70%. The research aimed to identify the optimal model that balances accuracy and computational efficiency for effective diabetes prediction.

Isfafuzzaman Tasin, Sanjida Islam and Riasat Khan [8], there study focuses on early diabetes detection using machine learning and AI, addressing the growing global prevalence of diabetes. It combines the Pima Indian dataset with local Bangladeshi data to enhance prediction accuracy. Feature selection via mutual information, class imbalance handling with SMOTE/ADASYN, and models like XGBoost, SVM, and random forests. Machine learning significantly improves diabetes prediction, aiding early diagnosis and risk management.

Orlando Iparraguirre and Karina Espinola [9] conducted a study investigating machine learning-based methods for the early detection and classification of type 2 diabetes. They evaluated five machine learning models—K-Nearest Neighbors (KNN), Decision Tree (DT), Random Forest (RF), Support Vector Machine (SVM), and Logistic Regression (LR)—to identify the most effective predictive approach. Their findings indicated that the Random Forest model achieved the highest accuracy, underscoring the potential of machine learning in improving diabetes diagnosis and facilitating early intervention.

The study by KM Joyti Rani [10] "Diabetes Prediction Using Machine Learning" aims to serve as a powerful tool for early diabetes detection, enabling more accurate and efficient diagnosis. The study demonstrates that among the evaluated models—K-Nearest Neighbours (KNN), Logistic Regression (LR), Random Forest (RF), Support Vector Machine (SVM), and Decision Tree (DT)—the most effective algorithm is identified based on its predictive accuracy. These findings underscore the potential of machine learning in enhancing early diagnosis and facilitating timely medical intervention for diabetes management. techniques can significantly enhance the accuracy of early diabetes diagnosis, potentially improving patient outcomes.

The study "A Survey on Diabetes Risk Prediction Using Machine Learning Approaches" by Shimoo Firdous, Gowher A Wagai, and Kalpana Sharma explores the application of machine learning models for early diabetes mellitus detection. The authors conduct a comprehensive review of existing research, focusing on the use of machine learning for diabetes risk prediction. Their analysis covers several prominent techniques, including Support Vector Machine (SVM), K-Nearest Neighbour (KNN), and Random Forest (RF). This research contributes to the growing body of knowledge on how advanced engineering and computational methods can be leveraged to improve healthcare outcomes through proactive and accurate disease prediction.

The survey concludes that machine learning approaches, particularly classification algorithms like SVM, KNN, and RF, show high accuracy in predicting diabetes at an early stage, emphasizing the potential of these techniques in healthcare.

Authors and year	Objective	Techniques/Algorithms Used	Dataset Used	Performance (Accuracy, AUC, etc.)	Key Findings
N. Kushal & Keshav Krishnamurthy (2024)[4]	Early prediction of diabetes using ML	SVM, KNN, Logistic Regression, Random Forest	Sylhet Diabetes Hospital	Random Forest achieved 98.08% accuracy	Integrated the best model into a web application
Malathy S. & Santhiya M. (2021)[5]	Develop a diabetes prediction system	Naïve Bayes, SVM, ANN	Primary health care centres	ANN performed the best	Implemented using R programming language
Israt Jahan Kakoly & ME Rakibul Hoque (2023)[1]	Identify diabetes risk factors using ML	PCA, Information Gain + Decision Tree, RF, SVM, Logistic Regression, KNN	Bangladeshi surveys	Accuracy: 82.2%, AUC: 87.2%	Clinical risk factors were effectively identified, including dietary factors
Nidhi Kumarii & Madhu Gautam (2023)[6]	Enhance disease detection and minimize errors	SVM, NB, RF, KNN, DT	Pima Indian Diabetes (Kaggle)	Random Forest achieved 92% accuracy	Addressed missing values and imbalanced data
Arwakti Chen Lyngdoh & Nurul Amin Choudhary (2020)[7]	Optimize accuracy and computational efficiency	KNN, NB, DT, RF, SVM	Pima Indians Diabetes	KNN achieved the highest accuracy of 76%	Other classifiers maintained around 70% accuracy
Isfafuzzaman Tasin & Riasat Khan (2022)[8]	Automated diabetes prediction using ML	XGBoost, SVM, Random Forest	Pima Indian + Bangladeshi dataset	Improved prediction accuracy	Feature selection with mutual information, handled class imbalance using SMOTE/ADASYN
Orlando Iparraguirre, Karina Espinola, (2023)[9]	Timely identification and classification of diabetes.	KNN, Decision Tree, RF, SVM, LR	Pima Indian dataset	Random Forest performed best	ML models enhance diabetes risk assessment
KM Jyoti Rani (2020)[10]	Diabetes prediction using ML	KNN, LR, RF, SVM, Decision Tree	Pima Indian dataset	Improved accuracy in early diagnosis	ML models significantly improve prediction accuracy
Shimoo Firdous & Kalpana Sharma (2022)[11]	Review of ML approaches for diabetes prediction	SVM, KNN, Random Forest	Pubmed	High accuracy with classification models	Survey concluded SVM and RF are most effective
Pradeepa Sampath&Gurupriya Elangovan (2024)[12]	Diabetic prediction using ensemble ML	AdaBoost and XGBoost	Pima Indians Diabetes	AUC of 0.968+/-0.015	Significantly improve diabetes prediction
Shahid Mohm Ganie & Saurav Malik	Boosted ensemble learning for	ML metrics and k-fold techniques	Tima Diabetes Dataset	Accuracy rate of 92.85%	Enhanced detection and

(2023)[13]	diabetes prediction				prediction of diabetes
Aishwarya Mujumdar, &V Vaidehi. (2019)[2]	Diabetes Prediction using Machine Learning Techniques	LR, Pipelines	Not given	Logistic Regression – 96% Pipeline-98.85	Precision of diabetes prediction

Table:1 Literature Review

3.Methodology:

This study presents a diabetes prediction model leveraging machine learning techniques, built upon the Sylhet Diabetes Hospital dataset comprising 520 patient records. The methodology involves a systematic approach starting from data pre-processing, feature selection, model training and evaluation to determine the best machine learning techniques for diabetes prediction. The complete workflow is shown in the flowchart:

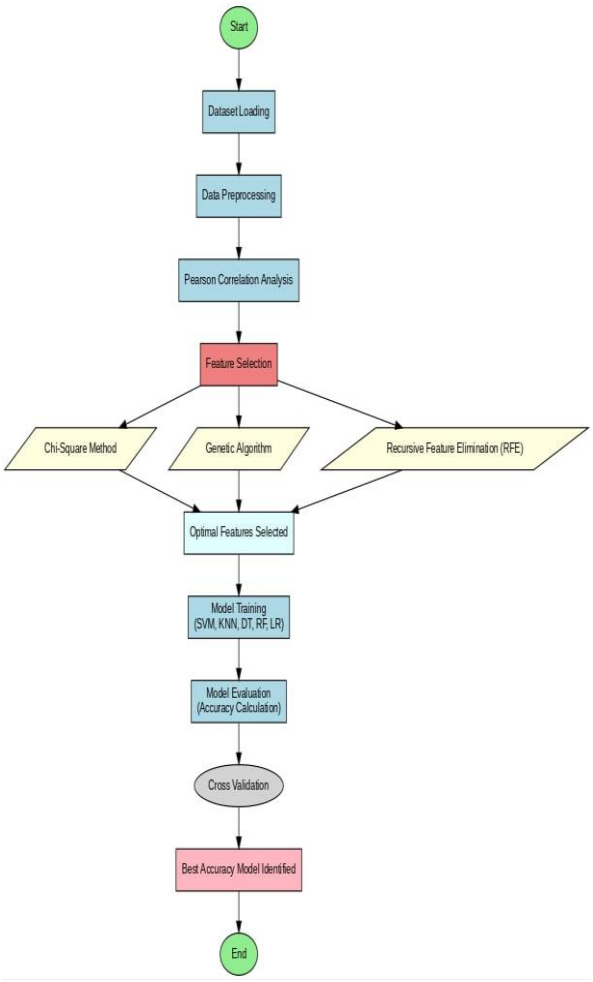


Figure:4 Flowchart of the Proposed Method

1. Start

- The process begins at the Start node which marks the beginning of the entire workflow.

2. Loading the dataset

Datasets used: Sylhet Diabetes Hospital dataset (520 records)

- The dataset consists of various medical and lifestyle-based attributes that help in predicting diabetes.

2.1 **Possible features include:**

- Age (years)

- Gender -Either it could be a male or a female

- Polyuria - Polyuria refers to excessive urine production, typically exceeding 3 liters per day in adults.

- Polydipsia- Polydipsia is excessive thirst, often associated with diabetes.

- Sudden weight loss- Sudden weight loss in diabetes is an unintended, rapid reduction in body weight caused by insulin deficiency, leading to fat and muscle breakdown for energy.

- Weaknesses

- Obesity (BMI related data)

Why is the choice of dataset important?

-The Sylhet Diabetes dataset is highly relevant for clinical applications as it provides real-world data from hospitalized patients.

- The 520 dataset ensures a balance between computational power and model accuracy.

3. Data Preprocessing

Data preprocessing is a vital initial step in building accurate predictive models, improving data integrity, and preparing the dataset for analysis. Failure to preprocess data can lead to significant issues such as inconsistencies, errors, noise, and missing values, increasing the risk of overfitting. To

assess the true influence of preprocessing on a classification algorithm for diabetes prediction, we systematically compared outcomes with and without preprocessing. During this phase, we explored three distinct feature selection techniques and identified the most effective one to enhance the model's accuracy and reliability.

4. Pearson Correlation Analysis

The Pearson Correlation Coefficient (PCC) measures the strength of the relationship between independent variables and diabetes occurrence. Features that are highly correlated with diabetes are kept, while features with weak correlation may be removed to avoid overfitting.

Pearson Corelation Formula:

$$r = \frac{n(\sum xy) - (\sum x)(\sum y)}{\sqrt{[n\sum x^2 - (\sum x)^2][n\sum y^2 - (\sum y)^2]}}$$

Figure 5: Pearson Correlation

a. "r" represents Pearson's correlation coefficient, indicating the strength and direction of the relationship between two variables.
b. "n" denotes the total number of stock pairs under consideration.
c. "∑xy" represents the summation of the products obtained by multiplying each corresponding stock pair.
d. "∑x" refers to the total sum of all values of the first variable.
e. "∑y" denotes the total sum of all values of the second variable.
f. "∑x²" signifies the summation of the squared values of the first variable.
g. "∑y²" represents the summation of the squared values of the second variable.

Example:

Pearsons correlation BETWEEN polyuria and polydipsia: 0.599

5. Feature Selection (Use 3)

Feature selection improves the accuracy of the model by removing irrelevant or redundant data.[16] Three different methods are used:

5.1 Chi-square (statistical approach):

-Tests independence between traits and the target variable (incidence of diabetes).

-Selects the features with highest statistical significance for prediction.

5.2 Genetic Algorithm (Optimization-Based Approach):

-Mimics herbal choice to discover the great subset of capabilities.

-Iteratively evolves the characteristic choice method to maximize prediction accuracy.[17]

Formula: We start by calculating a fitness value (let's call it "fl") for each person in a group (population). We do this for each person in the group, one by one (j = 1, 2, ... up to n).

$$p_i = f_i \backslash \sum_{j=1}^{n} \quad \text{..... (2)}$$

Next, we add up all these fitness values to get a total sum. Think of it like adding up scores.

$$F = \sum_{j=1}^{n} f_i \quad \text{..... (3)}$$

To decide which person gets selected, we use a formula to calculate the probability of each person being chosen. We'll call this calculated value "k." We repeat this for each person in the group (k = 1, 2. ... up to n).

..... (4)

$$p_k = f_k / \sum_{j=1}^{n} f_i$$

Figure 6: Genetic Algorithm

5.3 Recursive Feature Elimination (RFE) (Machine Learning-Based Approach)

-Trains a version iteratively and gets rid of the least vital capabilities step with the aid of using step.

-Ensures that best the maximum applicable capabilities are retained for very last training.

6. Optimal Feature Selection

After making use of Chi-Square, Genetic Algorithm, and RFE, the great-decided on capabilities are finalized.

These capabilities are then exceeded to the device mastering fashions for training.

7. Model Training (Five ML Models Used)

The decided-on capabilities are used to educate 5 extraordinary device mastering fashions:

Model Description:

a) Support Vector Machine (SVM): Works nicely with small datasets, unearths the great hyperplane to categories diabetic vs. non-diabetic cases.[14]
b) K-Nearest Neighbors (KNN): Predicts diabetes primarily based totally on comparable affected person cases (neighbors).[19]
c) Decision Tree (DT): Splits statistics into selection nodes primarily based totally on characteristic importance.
d) Random Forest (RF): Uses a couple of selection bushes to enhance prediction accuracy.[18]
e) Logistic Regression (LR): A statistical version normally used for binary class issues like diabetes detection.[15]

8. Model Evaluation (Accuracy Calculation)

- The performance of each model is tested as follows:
- Accuracy – measures the overall accuracy[21]
- Precision and Recall – important for medical prediction
- F1 score – The F1 score ensures an optimal balance between precision and recall, providing a comprehensive measure of a model's performance.
- ROC curve and AUC score – measures the reliability of the model

9. Cross-Validation

Cross-validation is a key machine learning technique for evaluating a model's generalization ability on unseen data. It involves splitting the dataset into multiple folds, where each fold serves as a validation set once while the model is trained on the remaining data. This iterative process ensures comprehensive assessment, and the final performance metrics are averaged to provide a more reliable estimate of the model's predictive accuracy. By mitigating overfitting and enhancing generalizability, cross-validation plays a crucial role in selecting an optimal model for deployment in real-world scenarios.

10. The most accurate model is determined

The most accurate model is selected for the final diabetes prediction.

11. The End

2.Model Evaluation

This process ends with a fully trained, validated and optimized machine learning model that is ready for real-time diabetes prediction.

12. Result:

In the given Table , the application of the Pearson correlation method indicates that among the seven input factors in the Sylhet Diabetes hospital dataset namely, Polyphagia, Obesity, Sudden Weight loss, Vision Blurring, Polyuria, Polydipsia and Itching the analysis yields the highest level of correlation among these variables

Table No. 2 Pearsons Correlation

S. No.	Corelation Between	Outcome
1	Polyphagia & Obesity	0.030
2	Sudden weight loss & Obesity	0.169
3	Vision Blurring & Obesity	0.109
4	Polyuria & Polydipsia	0.599
5	Polyuria & Weakness	0.263
6	Weakness & Polydipsia	0.332
7	Itching & Weakness	0.309

Upon performing Pearson Correlation analysis, a significant correlation was observed between Polyuria and Polydipsia. Subsequently, feature selection models were applied to refine the dataset. Thereafter, a heatmap was generated to visually illustrate the interrelationship between Polyuria and Polydipsia.

1.Feature Selection

By applying various feature selection technique, the following features are selected:

Genetic Algorithm: Female, Polyuria, Polydipsia, Visual Blurring, Delayed Healing, Sudden weight loss, Polyphagia, Partial Paresis.

Recursive Feature Elimination: Male, Female, Polyuria, Polydipsia, Irritability.

Chi-Square Method: Female, Polyuria, Partial Paresis, Polydipsia, Sudden weight loss.

Best features (Optimal Features) are generated by Genetic Algorithm (GA)

This study evaluated five different machine learning models for diabetes prediction: Support Vector Machine (SVM), K Nearest Neighbor (KNN), Decision Tree (DT), Random Forest (RF), and Logistic Regression (LR). Each techinique was evaluated based on several performance metrics to determine its effectiveness in predicting diabetes.

Comparison Of Model Performance

Each model was trained and tested using the Sylhet Diabetes Hospital dataset, which consists of 520 records, following an 80:20 train-test split. The performance evaluation outcomes are summarized in the following table

Model Evaluation	Accuracy	Precision	Recall	F1-Score	ROC_AUC
Decision Tree	93.74%	96.78%	93.34%	95.21%	96.19%
Random Forest	94.70%	96.47%	94.90%	95.66%	98.22%
SVM	89.19%	88.46%	94.90%	91.52%	95.39%
Logistic Regression	90.03%	90.31%	94.51%	92.33%	95.18%
KNN	91.58%	95.23%	91%	93.02 %	96.01%

Table 3: Performance Evaluation

Findings And Best Model Selection

The results indicate that Random Forest outperformed all other models, achieving the highest accuracy (94.70%), F1 score (95.66%) and AUC-ROC (98.22%), making it the most reliable model for diabetes prediction.

Decision Tree also performed well at 96.19%, making it another viable choice.

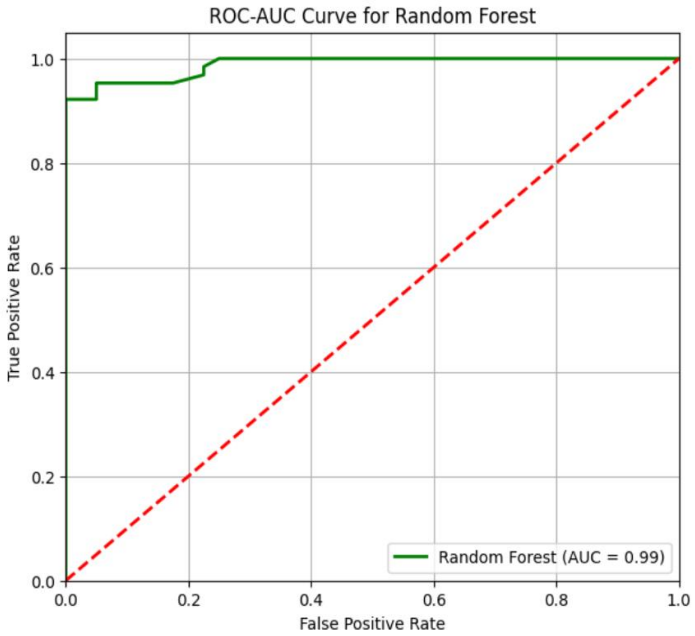


Figure 7: AUC-ROC Curve

4.Heat Map Of Polyuria And Polydipsia

3.Auc-Roc Curve For Random Forest Classifier

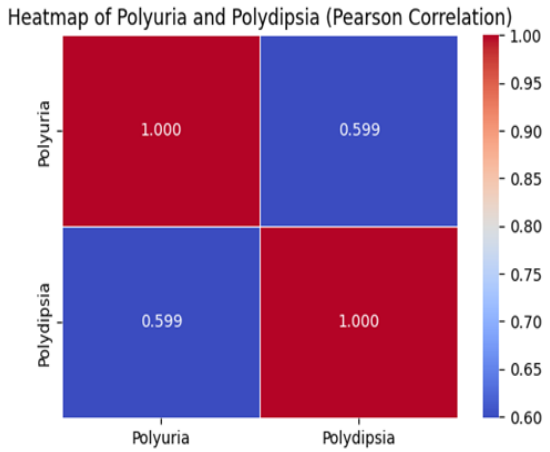


Figure 8: Heat Map

13.Conclusion:

This study investigates the effectiveness of machine learning algorithms in diabetes prediction using the Sylhet Diabetes Hospital dataset. To improve predictive accuracy, various feature selection techniques—Pearson correlation, Genetic Algorithm, Chi-Square test, and Recursive Feature Elimination (RFE)—were applied to identify the most relevant attributes. The optimized feature set was then used to train and evaluate five machine learning models: Support Vector Machine (SVM), K-Nearest Neighbors (KNN), Decision Tree (DT), Random Forest (RF), and Logistic Regression (LR).

Among these models, Random Forest showed the highest accuracy of 98.22%, making it the most reliable model in diabetes prediction. Decision Tree model also performed well and showed competitive results. These findings support the effectiveness of machine learning in early detection of diabetes and may greatly aid healthcare professionals in risk assessment and prevention.

This study highlights the potential of machine learning models to improve diagnostic accuracy and reduce classification errors through advanced feature selection and classification techniques. Integrating these predictive models into clinical practice could enable early intervention, improving patient outcomes and lowering healthcare costs.

14.Future Work:

While the study achieved promising results, several areas warrant further exploration to enhance the robustness and applicability of the proposed model:

14.1 Expansion of Dataset

Incorporating a larger and more diverse dataset from different geographical regions and demographics would improve the model's generalizability.

Including real-time clinical data from electronic health records (EHRs) could further enhance prediction accuracy.

14.2 Feature Engineering & Optimization

Additional feature engineering methods could be explored to refine the dataset and eliminate potential noise.

Incorporating domain-specific features, such as genetic predisposition and lifestyle habits, could improve prediction accuracy.

14.3 Ensemble & Hybrid Models

Exploring ensemble techniques (e.g., stacking, boosting) that combine multiple models for improved accuracy and robustness.

Investigating reinforcement learning for adaptive diabetes prediction.

15.Reference:

[1]. Israt Jahan Kakoly, Md. Rakibul Hoque and Najmul Hasan," Data-Driven Diabetes Risk Factor Prediction Using Machine Learning Algorithms with Feature Selection Technique", Kakoly, I.J.; Hoque, M.R.; Hasan, N. Data-Driven Diabetes Risk Factor Prediction Using Machine Learning Algorithms with Feature Selection Technique. Sustainability 2023, 15, 4930.

[2]. Aishwarya Mujumdara , Dr. Vaidehi V,"Diabetes Prediction using Machine Learning Algorithms", International Conference on Recent Trends in Advanced Computing 2019.

[3]. Kaggle Dataset by Sylhet Diabetes Hospital – Diabetes

[4] N. Kushal Kumar Raju and Keshav Krishnamurthy," Diabetes Prediction Using Machine Learning and Flask", Biomedical & Pharmacology Journal, June 2024. Vol. 17(2), p. 1307-1316

[5]. Malathy S. and Santhiya M.." Diabetes Disease Prediction Using Artificial Neural Network with Machine Learning Approaches", 2021 5th International Conference on Electronics, Communication and Aerospace Technology (ICECA) | 978-1-6654-3524-6/21/$31.00 ©2021 IEEE | DOI: 10.1109/ICECA52323.2021.9676094

[6]. Nidhi Kumari and Madhu Gautam.” Analysis of Diabetes Disease Prediction Using Machine Learning Algorithms”, 2023 International Conference on Circuit Power and Computing Technologies (ICCPCT) | 979-8-3503-3324-4/23/$31.00 ©2023 IEEE |

[7] Arwatki Chen Lyngdoh and Nurul Amin Choudhary .” Diabetes Disease Prediction Using Machine Learning Algorithms”. 2020 IEEE-EMBS Conference on Biomedical Engineering and Sciences (IECBES) | 978-1-7281-4245-6/21/$31.00 ©2021 IEEE |

[8] Isfafuzzaman Tasin, Tansin Ullah Nabil, Sanjida Islam and Riasat Khan,” Diabetes prediction using machine learning and explainable AI techniques”, Healthc Technol Lett. 2022 Dec 14;10(1-2):1–10. doi: 10.1049/htl2.12039

[9]. Orlando Iparraguirre-Villanueva , Karina Espinola-Linares,” Application of Machine Learning Models for Early Detection and Accurate Classification of Type 2 Diabetes”, Diagnostics (Basel). 2023 Jul 15;13(14):2383. doi: 10.3390/diagnostics13142383

[10]. The study by KM Joyti Rani "Diabetes Prediction Using Machine Learning" International Journal of Scientific Research in Computer Science Engineering and Information Technology

[11]. Shimoo Firdous, Gowher A Wagaiand , Kalpana Sharma, “ A survey on diabetes risk prediction using machine learning approaches”, J Family Med Prim Care. 2022 Nov;11(11):6929-6934. doi: 10.4103/jfmpc.jfmpc_502_22.

[12]. Pradeepa Sampath&Gurupriya Elangovan, “ Diabetic prediction using ensemble ML”, Scientific Reports volume 14, Article number: 28984 (2024)

[13]. Shahid Mohm Ganie & Saurav Malik, “An ensemble learning approach for diabetes prediction using boosting techniques”, Sec. Computational Genomics Volume 14 – 2023

[14] A. Mir and S. N. Dhage, “Diabetes disease prediction using machine learning on big data of healthcare,” in 2018 Fourth International Conference on Computing Communication Control and Automation (ICCUBEA), 2018, pp. 1–6.

[15] J. Smith, J. Everhart, W. Dickson, W. Knowler, and R. Johannes, “Using the adap learning algorithm to forcast the onset of diabetes mellitus,” Proceedings- Annual Symposium on Computer Applica tions in Medical Care, vol. 10, 11 1988.

[16] P. S. Kohli and S. Arora, “Application of machine learning in disease prediction,” in 2018 4th International Conference on Computing Communication and Automation (ICCCA), 2018, pp. 1–4.

[17] Jasjit Singh , Deepanshu Goyal and Apurva Vashisht , “Feature Selection Using Correlation Analysis for Accurate Breast Cancer Diagnosis”, IGI Global- Applications of Synthetic high dimensional data . CH – 7

[18] Talha Mahboob Alam [a], Muhammad Atif Iqbal, “ A model for early prediction of diabetes”, Informatics in Medicine Unlocked

Volume 16, 2019, 100204

[19] Shetty, Deeraj, et al. "Diabetes disease prediction using data mining." 2017 international conference on innovations in information,

embedded and communication systems (ICIIECS). IEEE, 2017.

[20] Gopi Battineni 1,ORCID,Getu Gamo Sagaro, “Comparative Machine-Learning Approach: A Follow-Up Study on Type 2 Diabetes Predictions by Cross-Validation Methods” , Machines 2019, 7(4), 74

[21] Shadman Sakib and Nowrin Yasmin , “ Performance Analysis of Machine Learning Approaches in Diabetes Prediction”, 2021 IEEE 9th Region 10 Humanitarian Technology Conference (R10-HTC)

[22] Kanchan, B. Dhomse, and M. Mahale Kishor. "Study of machine

learning algorithms for special disease prediction using principal of

component analysis." 2016 international conference on global trends in

signal processing, information computing and communication

(ICGTSPICC). IEEE, 2016.

[23] Vijayan, V. Veena, and C. Anjali. "Prediction and diagnosis of diabetes mellitus—A machine learning approach." 2015 IEEE Recent Advances in Intelligent Computational Systems (RAICS). IEEE, 2015