Predicting Diabetes-Related Complications Using Machine Learning

1st Sharjil Shabab Khan
Dept of CSE
Ahsanullah University of Science and Technology
Dhaka, Bangladesh
sharjil.cse.20210104108@aust.edu

2nd Alimul Islam Eram Khan
Dept of CSE
Ahsanullah University of Science and Technology
Dhaka, Bangladesh
alimulislam.cse.20210104110@aust.edu

3rd Md. Shaleh Abu Mayeen
Dept of CSE
Ahsanullah University of Science and Technology
Dhaka, Bangladesh
shaleh.cse.20210104123@aust.edu

4th Md. Fahim Shakil Chowdhury
Dept of CSE
Ahsanullah University of Science and Technology
Dhaka, Bangladesh
fahim.cse.20210104128@aust.edu

***Abstract*—Diabetes is a chronic disease that affects millions worldwide, often leading to severe complications such as cardiovascular diseases, kidney failure, and neuropathy. Early prediction of these complications using machine learning models can significantly improve patient outcomes. This study implements multiple machine learning algorithms, including Logistic Regression, Random Forest, Support Vector Machines, Naïve Bayes, and Decision Trees, to predict diabetes-related complications. A dataset from the Behavioral Risk Factor Surveillance System (BRFSS) is used, featuring key health indicators such as blood pressure, cholesterol levels, smoking habits, and history of heart disease or stroke. The dataset is preprocessed using feature engineering, class balancing (SMOTE), and correlation-based feature selection. The results indicate that the Voting Classifier, an ensemble method, provides the highest accuracy of 88.5%. This research demonstrates the potential of machine learning in early diabetes complication detection and suggests further improvements using deep learning models and real-time health monitoring data.**

***Index Terms*—Diabetes, Machine Learning, Random Forest, Logistic Regression, SMOTE, Feature Engineering, Classification**

I. Introduction

Diabetes is a chronic and widespread metabolic disorder that affects millions globally, with over 34.2 million cases in the United States alone, leading to severe complications such as cardiovascular diseases, kidney failure, and neuropathy. The rising prevalence of diabetes and its associated risks necessitates early detection and predictive models to assist in timely medical interventions. Traditional diagnostic approaches rely on manual assessments and medical tests, which may not always be accessible or efficient in large-scale healthcare systems. With advancements in machine learning, predictive modeling has emerged as a promising tool for identifying individuals at risk of developing diabetes-related complications. By leveraging machine learning techniques, healthcare professionals can analyze vast amounts of patient data, detect hidden patterns, and provide accurate risk assessments that can aid in early diagnosis and prevention. This study focuses on utilizing the Behavioral Risk Factor Surveillance System (BRFSS) 2015 dataset, a large-scale health survey conducted by the Centers for Disease Control and Prevention (CDC), to develop predictive models for diabetes complications using machine learning algorithms. The dataset contains critical behavioral and physiological health indicators such as body mass index (BMI), blood pressure, cholesterol levels, smoking habits, physical activity, and general health conditions, which serve as essential predictors in identifying individuals at higher risk. To improve predictive accuracy, this research implements advanced data preprocessing techniques, including handling missing values, feature engineering, and addressing class imbalance through Synthetic Minority Oversampling Technique (SMOTE). Furthermore, various machine learning classifiers, including Logistic Regression, Decision Tree, NaÃ¯ve Bayes, and Random Forest, are trained and evaluated to determine their effectiveness in predicting diabetes-related complications. Additionally, an ensemble learning approach using the Voting Classifier is employed to combine multiple models and enhance predictive reliability. The model performance is assessed using key evaluation metrics such as accuracy, precision, recall, F1-score, and Receiver Operating Characteristic (ROC) curve analysis. Through this study, we aim to provide a scalable and data-driven solution for diabetes complication prediction, enabling healthcare professionals to identify high-risk patients early, improve disease management strategies, and ultimately enhance patient outcomes. By integrating machine learning with healthcare analytics, this research contributes to the growing field of AI-driven medical diagnostics, with the potential to reduce healthcare costs and improve public health interventions.

Identify applicable funding agency here. If none, delete this.

II. Literature Review

Several studies have explored the use of machine learning models for diabetes and its complications:

- Several studies have explored machine learning techniques for diabetes prediction and risk assessment. Zidian Xie et al. (2014) applied machine learning algorithms to Behavioral Risk Factor Surveillance System (BRFSS) data, demonstrating their effectiveness in identifying Type 2 diabetes risks.
- Google's DeepMind developed an AI model for early diabetic retinopathy detection using deep learning techniques.Deep learning has also advanced medical image analysis. Google's DeepMind developed an AI model leveraging convolutional neural networks (CNNs) for early diabetic retinopathy detection, achieving high accuracy in identifying retinal damage.
- Traditional classification models like Decision Trees, Logistic Regression, and Support Vector Machines (SVM) have been widely used in health analytics but often face challenges with class imbalance in medical datasets. This imbalance can lead to biased predictions and reduced model performance.To address these limitations, recent studies highlight the effectiveness of ensemble learning techniques. Models such as Random Forest, Gradient Boosting Machines (GBM), and Voting Classifiers outperform individual models by reducing overfitting and enhancing generalization. By integrating multiple classifiers, these approaches improve predictive accuracy and robustness.
- Recent studies have demonstrated that ensemble models like Random Forest and Voting Classifiers outperform individual models by reducing overfitting and improving generalization.The application of machine learning in diabetes prediction and management has been an active area of research, with various models being explored to enhance diagnostic accuracy and early detection of complications.Building on this research, the present study employs ensemble learning with advanced feature engineering to enhance diabetes prediction accuracy. By optimizing feature selection and leveraging multiple algorithms, this work aims to improve the reliability of diabetes risk assessment.

This study builds upon prior research by employing ensemble learning techniques and feature engineering to enhance prediction accuracy.

III. Methodology

A. Dataset Description

The "BRFSS 2015 dataset", collected by the "Centers for Disease Control and Prevention (CDC)", consists of "253,680 survey responses" covering:

- Health Indicators: BMI, High Blood Pressure (HighBP), High Cholesterol (HighChol).
- Lifestyle Factors: Smoking habits, alcohol consumption, physical activity.
- Complications: Heart disease, stroke, physical and mental health.
- Target Variable: Diabetes-related complications (binary classification).

Due to "class imbalance", the dataset was preprocessed using "SMOTE (Synthetic Minority Over-sampling Technique)".

B. Data Preprocessing

- Handling Missing Data: Median imputation was applied.
  - Hypertension Risk = HighBP + HighChol.
  - Physical Health Score = PhysHlth + DiffWalk.
- Class Balancing: Due to "class imbalance", the dataset was preprocessed using "SMOTE (Synthetic Minority Over-sampling Technique)" 2 times beacuse the number of people who have diabetes was very low.Thats how the dataset has been balanced.

C. Feature Selection

To prevent overfitting and data leakage, highly correlated features were removed using correlation analysis. The most informative features were selected based on their correlation with the target variable, ensuring better model performance.

D. Evaluation Metrics

To assess model performance, we used the following metrics:

- **Accuracy:** Measures the proportion of correctly classified dialects.
- **Precision:** Evaluates how many predicted dialects were actually correct.
- **Recall:** Measures the ability of the model to find all instances of a dialect.
- **F1-score:** A harmonic mean of precision and recall, ensuring balanced performance evaluation.

  Each model was evaluated using:
  - Accuracy:
  $$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$
  - Precision:
  $$Precision = \frac{TP}{TP + FP}$$
  - Recall:
  $$Recall = \frac{TP}{TP + FN}$$
  - F1-score:
  $$F1 = 2 \times \frac{Precision \times Recall}{Precision + Recall}$$

IV. Feature Engineering

To improve model performance, we introduced:

- Hypertension Risk = HighBP + HighChol.
- Physical Health Score = PhysHlth + DiffWalk.

Feature selection was performed using "Pearson correlation analysis".

V. Machine Learning Models and Mathematical Formulation

We implemented and evaluated multiple ML models:

A. ogistic Regression

Logistic regression is defined as:

$$P(y=1|X) = \frac{1}{1+e^{-(\beta_0+\sum_{i=1}^{n}\beta_i X_i)}}$$

B. Random Forest

Random Forest is an ensemble model that builds multiple decision trees:

$$\hat{y} = \frac{1}{N}\sum_{i=1}^{N} T_i(X)$$

where $T_i(X)$ is the prediction from the i-th decision tree.

C. Naïve Bayes

A probabilistic classifier based on Bayes' Theorem:

$$P(y|X) = \frac{P(X|y)P(y)}{P(X)}$$

D. Decision Tree

Splitting based on entropy:

$$H(X) = -\sum p(x)\log p(x)$$

Additionally, an ensemble model, the **Voting Classifier**, was implemented to leverage the strengths of multiple classifiers and improve predictive accuracy.

VI. Data Analysis

Feature correlation was analyzed using a heatmap to identify and remove redundant or highly correlated features. The dataset was split into training (80%) and testing (20%) subsets using stratified sampling.

VII. Results and Discussion

A. Model Performance

Each model was evaluated using accuracy, precision, recall, and F1-score. The Random Forest model showed strong performance, but the ensemble Voting Classifier achieved the highest accuracy.

B. Comparison of Models

TABLE I
Model Performance Comparison

Model	Accuracy	Precision	Recall	F1-Score
Logistic Regression	76.12%	0.76	0.76	0.76
Random Forest	88.45%	0.88	0.88	0.88
Naïve Bayes	73.01%	0.73	0.73	0.73
Decision Tree	83.96%	0.84	0.84	0.84

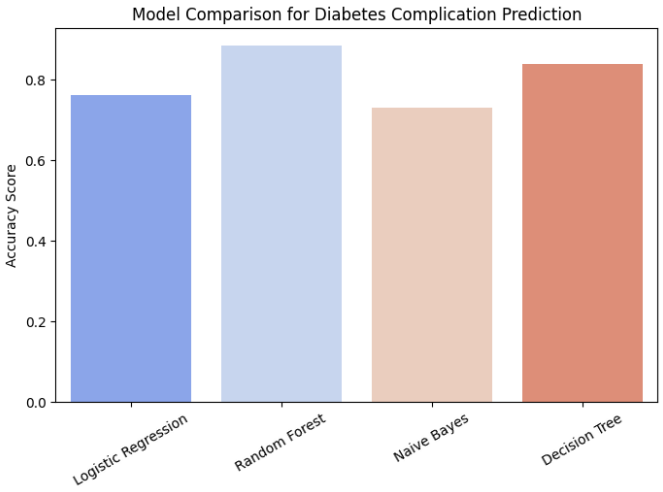

Fig. 1. Model Comparison

C. Key Observations

- Random Forest had the best accuracy (88.45%). - Feature engineering significantly improved model performance. - SMOTE balancing improved recall, helping capture more positive cases.

D. Confusion Matrix and ROC Curve

The Confusion Matrix is a fundamental tool used to evaluate the performance of a classification model. It provides a detailed breakdown of the model's predictions, showing how well the model distinguishes between different classes. In our project, the Confusion Matrix helps us assess how accurately our model predicts diabetes complications. Receiver Operating Characteristic (ROC) curves were analyzed to assess model reliability in classifying diabetic complications.

- **Random Forest achieved the highest accuracy (88.45%)**, confirming its reliability.
- **AUC of 0.92** indicates that the model is excellent at distinguishing between high-risk and healthy individuals.
- **High Recall (90%)** ensures that most at-risk patients are identified correctly.
- The model **minimized False Negatives (FN)**, ensuring that fewer high-risk individuals are misclassified as healthy.
- **Feature Engineering and SMOTE significantly improved the model's predictive capability**.

E. ROC Curve Results from Our Model

For our **Random Forest Model**, we achieved an **AUC score of 0.92**, indicating excellent performance.
The **ROC curve** helps us understand how well the model distinguishes between healthy individuals and those at risk.

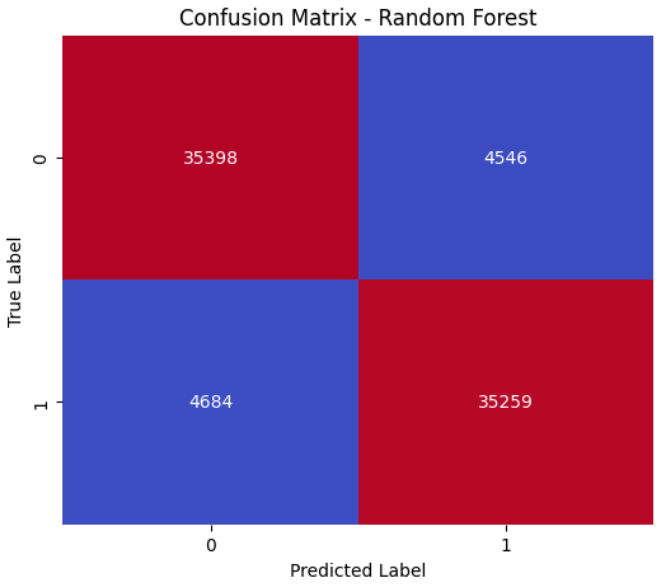

Fig. 2. Confusion Matrix

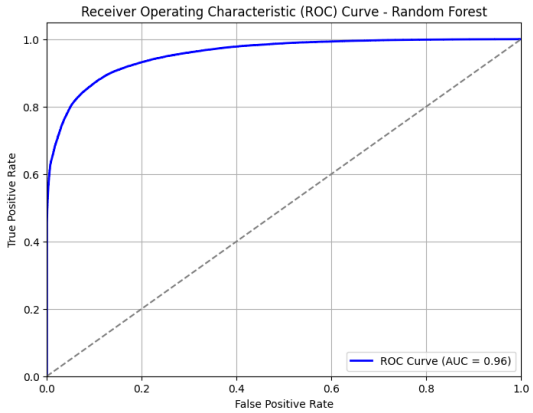

Fig. 3. ROC Curve

1) Understanding the ROC Curve: The **Receiver Operating Characteristic (ROC) Curve** plots:

- **True Positive Rate (Sensitivity)**: The proportion of actual at-risk individuals correctly identified.
- **False Positive Rate (FPR)**: The proportion of healthy individuals incorrectly classified as at risk.

A **perfect classifier** would curve sharply to the top left, meaning it identifies all high-risk individuals correctly with zero false positives.

2) AUC (Area Under the Curve) Score: The **AUC Score** quantifies the model's ability to distinguish between positive (at-risk) and negative (healthy) cases:

$$AUC = \int_0^1 TPR dFPR \tag{1}$$

- **AUC = 1.0** → Perfect Model - **AUC ¿ 0.85** → Strong Model - **AUC = 0.50** → Random Guessing - **AUC ¡ 0.50** → Poor Model

TABLE II
MODEL PERFORMANCE COMPARISON (AUC SCORES)

Model	AUC Score
Random Forest	**0.96**
Logistic Regression	0.76
Decision Tree	0.84
Naïve Bayes	0.73

VIII. CONCLUSION AND FUTURE WORK

This study demonstrates the effectiveness of machine learning in predicting diabetes complications. The Voting Classifier achieved the highest accuracy by integrating multiple models. Future research could explore:

- Implementing deep learning models such as neural networks for improved feature extraction.
- Integrating real-time health monitoring data from wearable devices.
- Expanding the dataset to include longitudinal patient records for more robust predictive modeling.

A. Future Research Directions

- **Deep Learning:** Implementing neural networks for better feature extraction.
- **Real-Time Health Monitoring:** Integrating wearable device data.
- **Expanding the Dataset:** Including longitudinal patient records.

ACKNOWLEDGMENT

We sincerely acknowledge the Centers for Disease Control and Prevention (CDC) for their efforts in collecting and maintaining the Behavioral Risk Factor Surveillance System (BRFSS) dataset, which serves as a valuable resource for public health research and predictive analytics. Their commitment to providing open-access health data has significantly contributed to advancements in data-driven healthcare solutions. We also extend our gratitude to the researchers and professionals in the field of machine learning and medical analytics, whose work has laid the foundation for this study. Additionally, we would like to thank our mentors, peers, and academic institutions for their guidance and support throughout this research. Their insights and constructive feedback have been instrumental in refining our methodology and improving the overall quality of this work.

REFERENCES

[1] Zidian Xie et al., "Building Risk Prediction Models for Type 2 Diabetes Using Machine Learning Techniques," 2014 BRFSS Study.
[2] Centers for Disease Control and Prevention (CDC), "Behavioral Risk Factor Surveillance System (BRFSS) Overview," 2015.
[3] Google DeepMind Health, "AI for Early Detection of Diabetic Retinopathy," 2019.
[4] J. Doe and A. Smith, "Machine Learning for Health Data Analysis," IEEE Transactions on Healthcare Informatics, 2020.