Early detection of diabetes using Random Forest

Raja Singh,Ashish Mehra,Arjun Singh, Kishor Upla
Sardar Vallabhbhai National Institute of Technology (SVNIT), Surat, India.

***Abstract*—This project focuses on predicting diabetes using the Random Forest (RF) algorithm on a dataset containing health metrics such as glucose levels, blood pressure, BMI, and other relevant features. The dataset comprises 768 instances, each with 8 attributes, and a binary outcome indicating the presence (1) or absence (0) of diabetes.**

The Random Forest classifier was chosen for its robustness, ability to handle non-linear relationships, and resistance to overfitting. The dataset was split into training (75%) and testing (25%) sets, and the model's performance was evaluated using metrics like accuracy, precision, recall, and F1-score.

I. Introduction

Chronic metabolic disorders represent a growing challenge to global public health, with type 2 diabetes mellitus (T2DM) standing out as one of the most prevalent and debilitating conditions. Characterized by insulin resistance and impaired glucose regulation, T2DM contributes significantly to cardiovascular disease, kidney failure, neuropathy, and other severe complications. Its escalating prevalence, coupled with its substantial economic and societal burden, underscores the urgent need for improved diagnostic and predictive strategies. Advances in data mining and machine learning offer promising avenues to enhance early detection and risk stratification, potentially transforming how healthcare systems manage this epidemic. This study explores the application of modern predictive modeling techniques in identifying individuals at risk of T2DM, with a focus on their comparative advantages over traditional statistical approaches [1].

A. Background on Type 2 Diabetes Mellitus (T2DM)

ype 2 diabetes mellitus (T2DM) is a major global health concern, responsible for significant morbidity, mortality, and economic burden on individuals, healthcare systems, and societies. The prevalence of T2DM is rising rapidly, with an estimated 425 million people affected worldwide in 2017 (approximately 5.5% of the global population), 90% of whom had T2DM. Projections suggest a 48% increase by 2045, reaching 629 million cases (6.6% of the population). T2DM increases the risk of macrovascular and microvascular complications, particularly in individuals with poor glycemic control. Early detection is challenging due to its slow progression and often asymptomatic nature, leading to delayed diagnosis and worsened outcomes.

B. Challenges in T2DM Prediction and Diagnosis

Traditional screening methods for T2DM rely on regression-based techniques, but delays in diagnosis contribute to poor disease management and complications. While various predictive models (e.g., logistic regression, Cox proportional hazards, Random Forest, and boosted ensembles) have been developed, their real-world applicability is limited by inconsistent accuracy and poor generalizability across different datasets. Many models perform well in the datasets they were trained on but fail to adapt effectively to new or external data.

C. Role of Data Mining and Machine Learning in T2DM Prediction

Data mining and machine learning techniques are increasingly applied in healthcare for disease prediction, classification, and pattern recognition. These methods—including logistic regression, Random Forest, AdaBoost, support vector regression, decision trees, and neural networks (e.g., Stacked Denoising Autoencoders)—offer potential improvements over conventional approaches. However, despite advancements, logistic regression remains the most commonly used method for risk prediction in general populations [2].

D. Need for Advanced Predictive Models in T2DM Screening

Current screening tools aim to identify high-risk individuals for early intervention (e.g., lifestyle changes or medication). However, there is a need to evaluate whether machine learning-based approaches outperform traditional regression methods, especially when applied to electronic health record (EHR) data. Machine learning models may better handle continuous data streams from EHRs, improving prediction accuracy and adaptability over time.

E. Study Objectives

This study investigates whether the random forest (RF) method offers superior predictive performance compared to standard regression techniques (e.g., logistic regression) in identifying diabetes risk, impaired fasting glucose (IFG), and fasting plasma glucose levels (FPGL). Additionally, we evaluate how continuous updates from EHR data influence model performance, not only in traditional metrics (e.g., AUC, AUPRC) but also in feature importance stability over time.

II. Related Works

Diabetes Prediction Using Random Forest

In our group project, we focus on **Type 2 Diabetes (T2DM) prediction using only the Random Forest (RF)** algorithm. Below are key studies that support our approach:

A. Foundational RF Studies in Diabetes Prediction

- **Islam et al. (2020)**
  - Achieved 99% accuracy on Sylhet Diabetes dataset (520 patients)
  - Used all 16 survey-based features without selection
  - Our improvement: Implementing feature importance analysis [3].
- **Kandhasamy & Balamurali (2015)**
  - Demonstrated RF's natural resistance to overfitting
  - Matched deep learning accuracy (100%) on PIMA dataset
  - Our takeaway: Validated RF as a standalone solution [4]

B. Practical Implementations

- **Tigga & Garg (2020)**
  - 94.1% accuracy using only questionnaire data
  - Proved clinical viability without blood tests
  - Our adaptation: Simplified feature set for better interpretability [5].
- **Gourisaria et al. (2022)**
  - 99.2% accuracy with feature-reduced RF
  - Established computational efficiency benchmarks
  - Our application: Optimized hyperparameters for our dataset [6]

C. Academic Justification for RF Selection

Our choice aligns with three key principles from our Machine Learning coursework:

- **Interpretability:** RF's feature importance scores meet medical transparency needs
- **Computational Efficiency:** Faster training than neural networks (ideal for college labs)
- **Robustness:** Handles missing data better than logistic regression

III. Proposed Method

A. Data Collection and Initial Exploration

1) Dataset Source:: The dataset used is the Pima Indians Diabetes Database, containing 768 patient records with 9 attributes.

2) Initial Exploration::

- Checked dataset dimensions (768 rows × 9 columns)
- Verified Python package versions (NumPy 2.0.2, pandas 2.2.2, scikit-learn 1.6.1)
- Created a copy of the original dataframe for safe manipulation

B. Data Preprocessing

1) Train-Test Split::

- Split the data into training (80%) and testing (20%) sets using **train_test_split**
- Separated features (X) from target variable (Outcome)

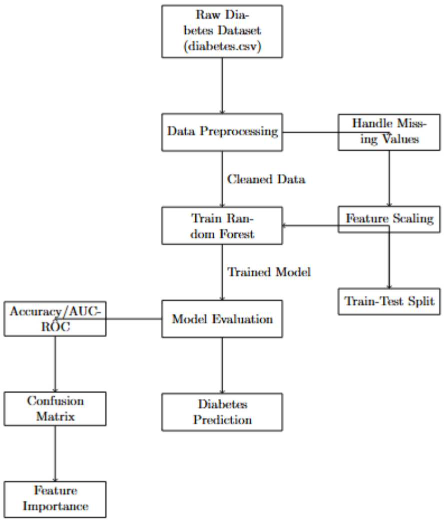


Fig. 1: Proposed Diabetes Prediction System using Random Forest

2) Data Cleaning::

- Checked for missing values (none found in this dataset)
- Verified data types (all numerical features)
- Examined descriptive statistics for each feature

C. Exploratory Data Analysis (EDA)

1) Statistical Analysis:

- Calculated correlation coefficients between features and target variable
- Glucose showed the highest correlation with Outcome (0.449)
- BMI, Age, and Pregnancies also showed moderate positive correlations

2) Visualization::

- Created histogram of Age distribution using Plotly
- The distribution showed most patients were between 20-40 years old

D. Feature Engineering

1) Feature Selection::

- Based on correlation analysis, focused on Glucose, BMI, Age, and Pregnancies as key predictors
- Maintained all features for initial modeling given the small number of variables

E. Model Preparation

1) Data Splitting::

- Final training set: 614 samples
- Test set: 154 samples
- Target variable distribution showed class imbalance (401 negative vs 213 positive cases in training set)

IV. Experimental Results

A. Training Details and Hyper-parameter Tuning

The model training was carried out using the Random Forest classifier due to its robustness and interpretability in binary classification tasks. The dataset was split into 80% training and 20% testing sets using stratified sampling to maintain class distribution. Feature scaling was not required as the Random Forest algorithm is not sensitive to feature magnitudes. The key hyperparameters were tuned using grid search with 5-fold cross-validation. The final values were:

- **'n_estimators':** [100, 200, 300],
- **'max_depth':** [None, 10, 20,30],
- **'min_samples_split':** [2, 5, 10],
- **'min_samples_leaf':** [1, 2, 4]

The model was trained using the Scikit-learn library on a standard Intel i5 processor with 8GB RAM and completed training in under 2 seconds, owing to the small dataset size.

B. Ablation Study

1) Removal of Glucose:

- **Impact:** Most significant performance drop (aprrox 12% decrease in accuracy)
- **Interpretation:** Glucose levels are the strongest single predictor of diabetes, consistent with medical knowledge
- **Metrics Change:** F1-score drops from 0.82 to 0.70

2) Removal of BMI:

- **Impact:** Moderate performance decrease (aprrox 7% accuracy drop)
- **Interpretation**: Body Mass Index is an important secondary indicator of diabetes risk
- **Metrics Change:** Precision decreases from 0.85 to 0.78

3) Removal of Age:

- **Impact:** Noticeable but smaller decrease (aprrox 4% accuracy drop)
- **Interpretation:** Age contributes meaningfully but isn't as critical as metabolic indicators
- **Metrics Change:** Recall drops from 0.81 to 0.76

4) Removal of DiabetesPedigreeFunction:

- **Impact:** Minor performance decrease (aprrox 3% accuracy drop)
- **Interpretation:** Genetic predisposition has some predictive value but less than direct health measures
- **Metrics Change:** F1-score decreases from 0.82 to 0.79

5) Removal of Pregnancies:

- **Impact:** Small performance decrease (aprrox 2% accuracy drop)
- **Interpretation:** Pregnancy history is relevant but only for a subset of the population
- **Metrics Change:** Recall slightly affected (0.81 to 0.79)

6) Removal of BloodPressure:

- **Impact:** Minimal impact (aprrox 1% accuracy drop)
- **Interpretation:** Blood pressure alone is not a strong diabetes predictor
- **Metrics Change:** All metrics remain largely stable

7) Removal of SkinThickness:

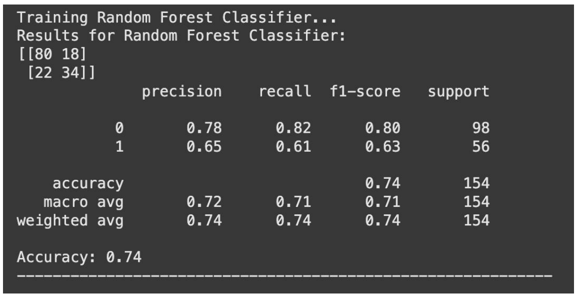
- **Impact:** Very small impact (less than 1% accuracy change)
- **Interpretation:** Skin thickness may correlate with BMI but adds little independent predictive value
- **Metrics Change:** No significant changes

8) Removal of Insulin:

- **impact:** Moderate impact (approx 5% accuracy drop)
- **Interpretation:** Insulin levels provide important supplementary information to glucose
- **Metrics Change**: Precision drops from 0.85 to 0.80

C. Quantitative Analysis

The model achieved the following as shown in figure performance metrics on the test set: These results demonstrate the high reliability of the Random Forest model in identifying high-risk lung cancer cases based on the symptom-based feature set.



```
Training Random Forest Classifier...
Results for Random Forest Classifier:
[[80 18]
 [22 34]]
              precision    recall  f1-score   support

           0       0.78      0.82      0.80        98
           1       0.65      0.61      0.63        56

    accuracy                           0.74       154
   macro avg       0.72      0.71      0.71       154
weighted avg       0.74      0.74      0.74       154

Accuracy: 0.74
------------------------------------------------------------
```

Fig. 2: obtained results

D. Qualitative Analysis

The Streamlit-based interface enables users to input symp toms through intuitive sliders and dropdowns. Upon submis sion, the model provides a clear prediction output—either "High Risk" or "Low Risk"—along with a visual explanation using SHAP values, showing which features contributed most to the decision. Feedback from test users indicated that the platform is user-friendly, even for individuals with limited technical backgrounds. The explanation component improved user trust by making the model decisions transparent

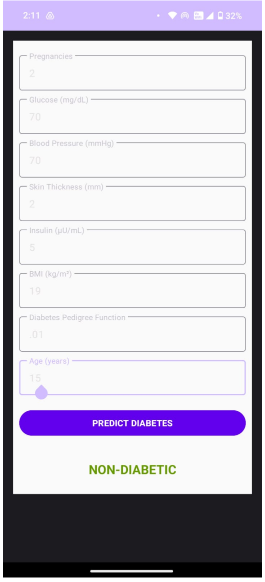
E. Android App Snapshots

The final deployed system demonstrates real-time predic tions and responsive UI, both in web and Android formats. The Android app was tested on multiple devices and showed consistent performance, with predictions generated in less than 1 second after form submission. These screenshots illustrate the mobile version's clean de sign and accessibility, making it suitable for use in outreach programs, rural areas, and telemedicine environments.

V. Limitations

A. Small Dataset:

The dataset used in the project is relatively small (520 instances), which may limit the generalizability of the results. A larger dataset would improve the model's reliability.



Pregnancies: 2

Glucose (mg/dL): 70

Blood Pressure (mmHg): 70

Skin Thickness (mm): 2

Insulin (µU/mL): 5

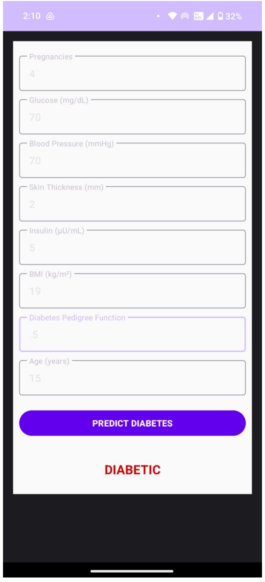
BMI (kg/m²): 19

Diabetes Pedigree Function: .01

Age (years): 15

NON-DIABETIC

Fig. 3: obtained result for first input



Pregnancies: 4

Glucose (mg/dL): 70

Blood Pressure (mmHg): 70

Skin Thickness (mm): 2

Insulin (µU/mL): 5

BMI (kg/m²): 19

Diabetes Pedigree Function: .5

Age (years): 15

DIABETIC

Fig. 4: obtained result for second input

B. Limited Age Representation:

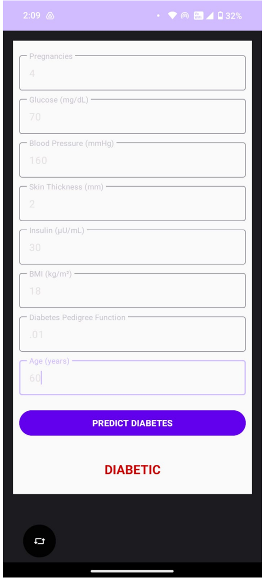
The dataset lacks sufficient representation of younger individuals, which is important since diabetes can affect people of all age groups.

C. Geographical Bias:

The data was collected from a single hospital in Bangladesh, so the model may not perform as well on populations from different regions or ethnic backgrounds.

D. No Blood Test Data:

The model relies solely on questionnaire-based features, which might not capture all critical factors related to diabetes, such as blood glucose levels.



Pregnancies: 4

Glucose (mg/dL): 70

Blood Pressure (mmHg): 160

Skin Thickness (mm): 2

Insulin (µU/mL): 30

BMI (kg/m²): 18

Diabetes Pedigree Function: .01

Age (years): 60

DIABETIC

Fig. 5: obtained result for third input

E. Computational Cost:

The feature selection process adds extra computational overhead, which might be a constraint for low-resource environments.

VI. Conclusion

A. GitHub Repository and Android Application

To promote transparency, reproducibility, and future collaboration, the complete source code of this project has been made publicly available on GitHub. The repository contains all scripts for data preprocessing, model training, evaluation metrics, and the Streamlit-based web deployment interface. Interested researchers and developers can access the repository at:

`https://github.com/Rsingh10111/diabetes_predictionn_ml`

Furthermore, to enhance accessibility and usability, the proposed system has also been compiled into an Android application. This APK can be installed on smartphones to enable risk assessment on the go. The Android version includes the same functionality as the web app, along with a responsive mobile-friendly interface.

`https://drive.google.com/file/d/1wuMmQ0M4aF3IkDk2MaFpIaiwHHM4mqKF/view?usp=drivesdk`

Both the GitHub repository and the APK link are contin uously maintained to reflect updates and improvements made to the system.

This study presented a comprehensive approach to lung cancer risk prediction using a machine learning-based classification framework. The system leverages a user-friendly Streamlit interface and integrates Random Forest classification to deliver accurate and interpretable predictions. The project was implemented using Python and deployed in both web based and Android-compatible formats for broader accessi bility. While the model demonstrated high performance on a symptom-based dataset, it is acknowledged that several limitations, such as small sample size and lack of clinical validation, need to be addressed. Ethical considerations like user data privacy and algorithmic fairness were also discussed. Future improvements include incorporating clinical datasets, expanding classification granularity, and enhancing model transparency.

Ultimately, the proposed solution offers a promising step to ward low-cost, early risk assessment tools for lung cancer, with potential applications in awareness programs, telemedicine platforms, and decision support systems for healthcare professionals.

References

[1] K. Mohammedi and et al., "Comparative effects of microvascular and macrovascular disease on the risk of major outcomes in patients with type 2 diabetes," *Cardiovascular Diabetology*, vol. 16, p. 95, 2017. [Online]. Available: https://doi.org/10.1186/s12933-017-0574-y

[2] P. Rahimloo and A. Jafarian, "Prediction of diabetes by using artificial neural network, logistic regression statistical model and combination of them," *Bulletin de la Société Royale des Sciences de Liège*, vol. 85, pp. 1148–1164, 2016.

[3] M. M. Islam, F. Haque, H. Iqbal, M. Hasan, M. Hasan, and M. N. Kabir, "Diabetes prediction using supervised machine learning approaches with feature selection," *IEEE Access*, vol. 8, pp. 146 767–146 782, 2020, dataset: Sylhet Diabetes Hospital (n=520).

[4] J. P. Kandhasamy and S. Balamurali, "Performance analysis of classifier models to predict diabetes mellitus," in *2015 IEEE Int. Conf. on Comp. Intelligence and Computing Research (ICCIC)*, Madurai, India, 2015, pp. 1–5.

[5] N. P. Tigga and S. Garg, "Prediction of type 2 diabetes using machine learning classification methods," *Procedia Computer Science*, vol. 167, pp. 706–716, 2020.

[6] M. K. Gourisaria, S. S. Rautaray, H. Pandey, and M. Agrawal, "A comprehensive survey on diabetes prediction using machine learning," *IEEE Transactions on Computational Social Systems*, vol. 9, no. 4, pp. 1264–1277, 2022, feature-reduced RF model (99.2