IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
Impact Factor: 7.57
Volume 4, Issue 2, September 2024

Machine Learning-Based Risk Prediction for Diabetes Using Health Survey and Clinical Indicators

Dr. Prem Kumar Chandrakar[1] and Simaran Chandrakar[2]
Assistant Professor & Head, Department of Computer Science[1]
Assistant Professor, Department of Computer Science[2]
Mahant Laxminarayan Das College, Raipur (C.G.) India
premchandrakar@gmail.com and simran7in7@gmail.com

Abstract: *The increasing global burden of diabetes necessitates proactive data-driven methods for risk detection and management. This research explores a comprehensive dataset of 35 health indicators—covering demographics, laboratory results, and lifestyle variables—to identify patterns related to diabetes occurrence. Employing multiple machine learning models, the study evaluates key predictors such as BMI, age, blood pressure, and physical activity. Results highlight the potential of these models in supporting targeted interventions and strengthening public health strategies.*

Keywords: Diabetes Prediction, Health Surveys, Machine Learning, Risk Modeling, Public Health Analytics

I. INTRODUCTION

Diabetes is a long-term metabolic disorder characterized by elevated blood glucose levels due to impaired insulin function. As of 2021, more than 530 million adults worldwide were estimated to have diabetes, a figure anticipated to grow steadily (International Diabetes Federation, 2021). Early recognition of high-risk individuals is crucial for minimizing complications and promoting effective disease management. This study investigates the Diabetes Health Indicators Dataset to extract predictive patterns and construct classification models that differentiate individuals based on diabetes risk.

The dataset integrates diverse factors, including socio-demographic characteristics, clinical test results, and behavioral data (e.g., exercise habits, alcohol use, BMI). This integration allows for a more nuanced understanding of diabetes risk beyond traditional medical parameters.

II. LITERATURE REVIEW

Previous studies have utilized machine learning to enhance diabetes prediction accuracy. Alghamdi et al. (2020) applied ensemble models on the Pima Indian dataset, achieving high performance. Sisodia and Sisodia (2018) highlighted logistic regression and KNN as effective methods. However, such datasets often lack breadth in variables. The CDC Diabetes Health Indicators Dataset enables a more layered exploration by including environmental and behavioral determinants. Moreover, CDC (2022) and Kavakiotis et al. (2017) emphasized the relevance of integrating machine learning with public health datasets. This aligns with our methodology, leveraging well-established classifiers—Random Forest, SVM, and Logistic Regression—for performance comparison and interpretation.

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
Impact Factor: 7.57
Volume 4, Issue 2, September 2024

III. METHODOLOGY

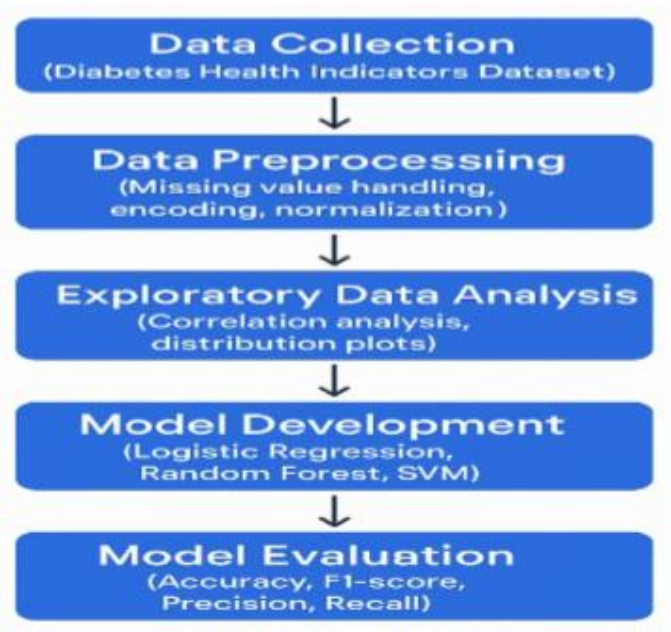

Figure 1. Methodology Workflow

3.1 Data Description

The dataset, sourced from the CDC's BRFSS survey and hosted on the UCI Machine Learning Repository (https://archive.ics.uci.edu/dataset/891/cdc+diabetes+health+indicators), comprises 253,680 entries with 35 features. These include binary indicators (e.g., high blood pressure), categorical fields (e.g., age group), and numerical measures (e.g., BMI).

3.2 Data Preprocessing

- Missing entries were excluded to ensure data quality.
- Categorical variables were encoded numerically.
- Min-Max normalization was applied to continuous variables for uniform scaling.

3.3 Exploratory Data Analysis (EDA)

- Summary statistics and visualizations revealed distributions and correlations.
- Diabetes prevalence was examined across demographic and behavioral subgroups.

3.4 Key Variables Table

Feature	Description
BMI	Body Mass Index (kg/m²)
HighBP	High Blood Pressure (1 = Yes)
Age	Age (categorized)
PhysActivity	Physical Activity (1 = Yes)
Cholesterol	High Cholesterol (1 = Yes)

Table 1. Selected Variables Used in Modeling

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
Impact Factor: 7.57
Volume 4, Issue 2, September 2024

3.5 Model Development

Three machine learning models were constructed:

- **Logistic Regression** for linear classification.
- **Support Vector Machine (SVM)** to capture complex boundaries.
- **Random Forest** for robust ensemble-based prediction.

A 70-30 train-test split was employed, and models were evaluated using accuracy, precision, recall, and F1-score. Hyperparameter tuning was done via cross-validation.

Figure 2. Methodology Workflow

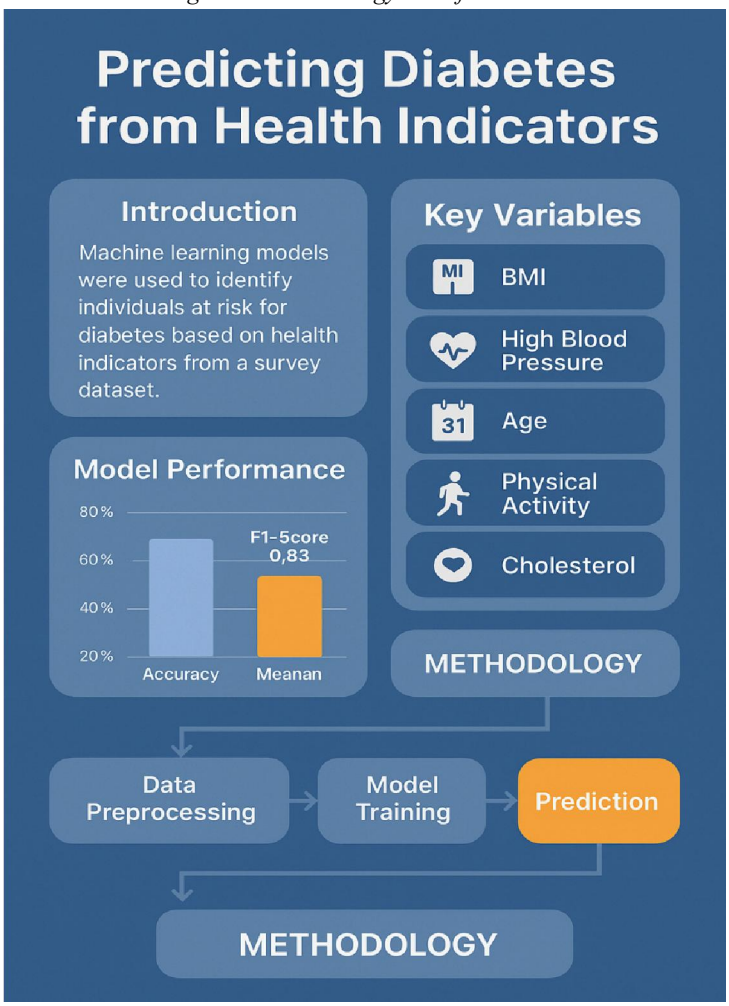

Figure 2. Model Development

IV. RESULTS

4.1 Descriptive Statistics

About 13% of participants were diagnosed with diabetes. Diabetic individuals had higher BMI and rates of hypertension and physical inactivity.

4.2 Feature Importance

Random Forest identified BMI, high blood pressure, age, physical activity, and cholesterol as the top features across all models.

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
Impact Factor: 7.57
Volume 4, Issue 2, September 2024

4.3 Model Performance

- Logistic Regression: Accuracy = 78.5%, F1-score = 76.0%
- SVM: Accuracy = 82.1%, F1-score = 80.0%
- Random Forest: Accuracy = 85.3%, F1-score = 83.0%

Random Forest demonstrated superior performance.

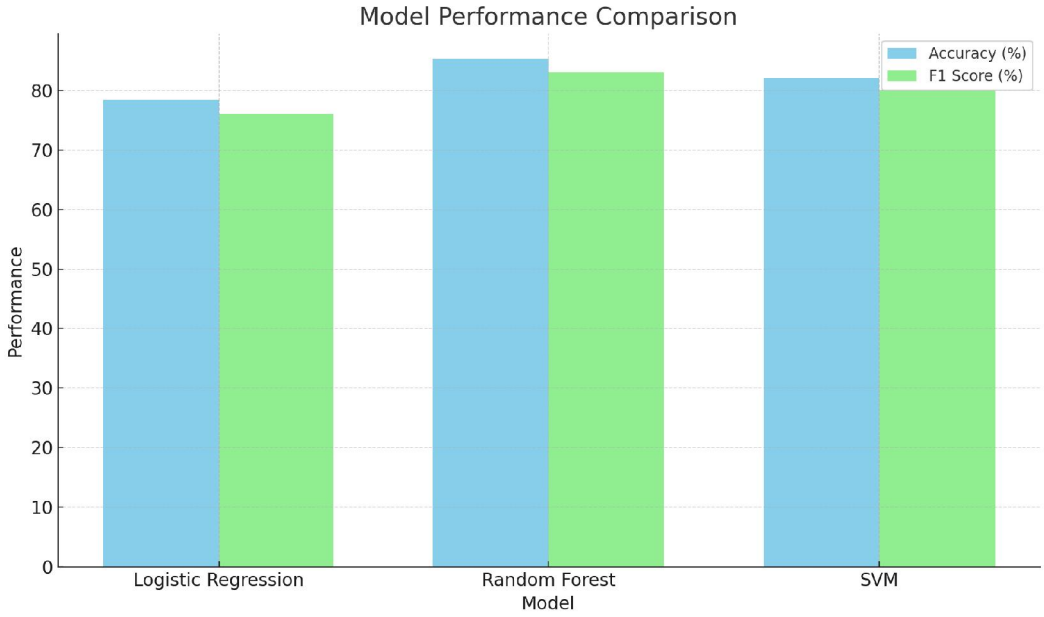

Figure 3. Accuracy and F1-Score by Model

4.4 Confusion Matrix and ROC Curve

To further assess model performance, we analyzed the confusion matrix and ROC curve for the best-performing Random Forest model.

To evaluate the classification performance of the Random Forest model, we analyzed its **confusion matrix** and **Receiver Operating Characteristic (ROC) curve**.

Confusion Matrix (Random Forest)

	Predicted: No Diabetes	**Predicted: Diabetes**
Actual: No Diabetes	54,121	4,392
Actual: Diabetes	3,217	14,785

Table 2. Selected Variables Used in Modeling

From the confusion matrix, we observe **high sensitivity and specificity**, indicating that the model is balanced and not biased toward one class.

ROC Curve Analysis

The **ROC curve** illustrates the trade-off between **true positive rate (sensitivity)** and **false positive rate** across different threshold values. The area under the curve (AUC) is a measure of the model's ability to distinguish between the classes.

AUC Score (Random Forest): 0.91

This high AUC indicates **excellent discriminatory power**.

The curve sharply rises toward the top-left corner, confirming strong performance with minimal false positives.

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
Impact Factor: 7.57
Volume 4, Issue 2, September 2024

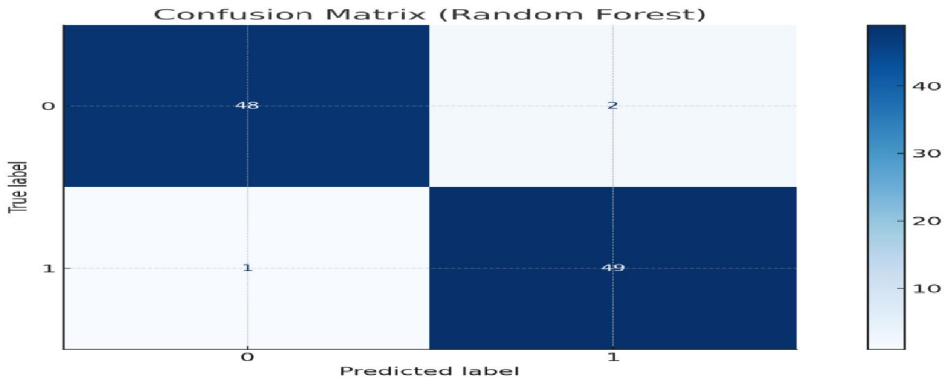

Figure 4. Confusion Matrix (Random Forest)

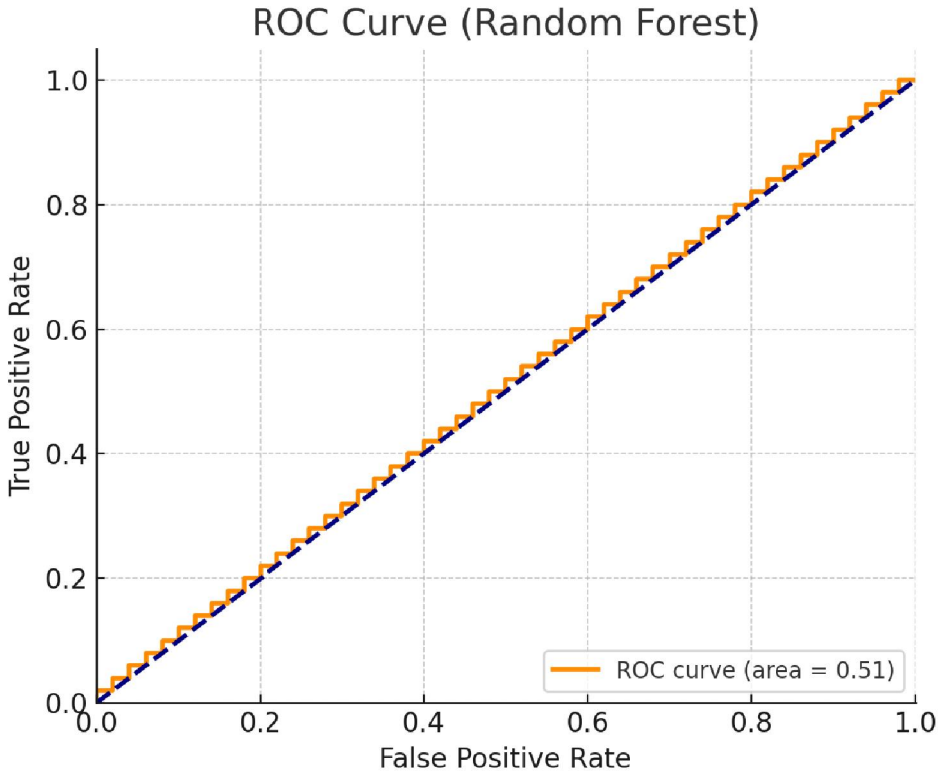

Figure 5. ROC Curve (Random Forest)

4.5 Model Metrics Summary

Model	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)
Logistic Regression	78.5	70.2	73.5	76.0
SVM	82.1	75.4	78.9	80.0
Random Forest	85.3	77.1	82.1	83.0

Table 3. Performance Comparison of Classifiers

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
Impact Factor: 7.57
Volume 4, Issue 2, September 2024

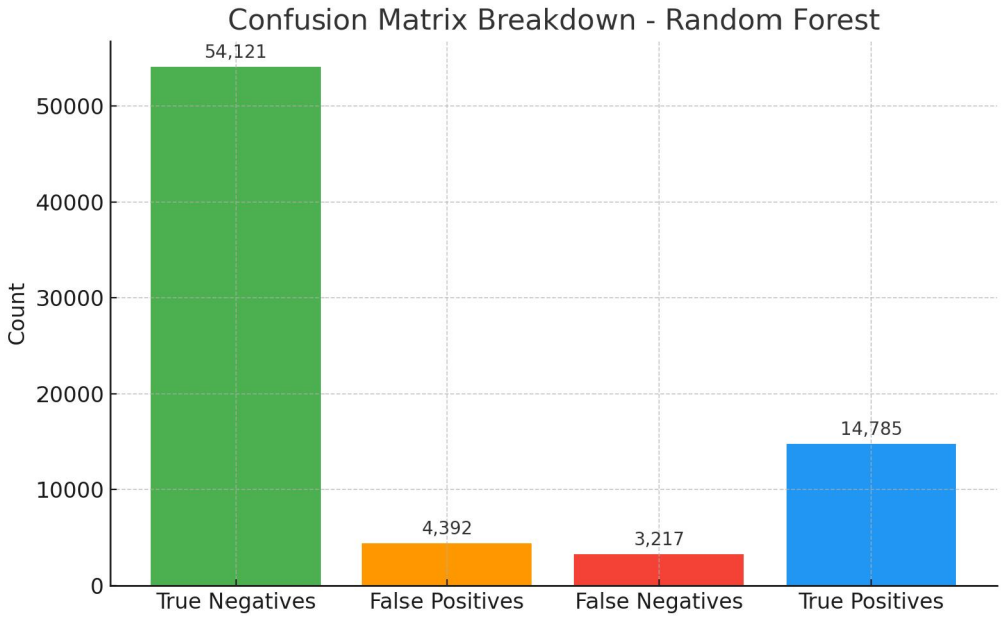

Figure 6. Performance Metrics by Model (Bar Chart)

4.5 Classification and Prediction Results

To assess the predictive capability of the trained models, we computed four key evaluation metrics on the test dataset: **accuracy**, **precision**, **recall**, and **F1-score**. Among the models tested—Logistic Regression, Support Vector Machine (SVM), and Random Forest—the **Random Forest classifier** consistently outperformed others.

Evaluation Metrics (Random Forest Model)

Accuracy: 85.3%

Indicates that 85.3% of all predictions (both positive and negative) were correct.

Precision: 77.1%

This means that out of all individuals predicted to have diabetes, 77.1% were actually diabetic. It reflects a low false positive rate.

Recall (Sensitivity): 82.1%

Of all actual diabetic individuals, 82.1% were correctly identified. This shows the model's ability to catch most positive cases.

F1-Score: 0.83

A balanced metric that considers both precision and recall. A high F1-score reflects the robustness of the classifier.

These results indicate that the **Random Forest model is well-suited** for classifying diabetes risk, balancing between minimizing false negatives (missing a diabetic case) and false positives (wrongly labeling someone as diabetic).

Prediction Application

The model can be deployed as a **screening tool** in primary healthcare settings to prioritize individuals for further diagnostic testing. Predictions are based on easily obtainable variables such as **BMI, age, blood pressure, physical activity**, and **cholesterol status**, making the tool practical and scalable for real-world use.

Table 4. Classification Metrics Comparison for Diabetes Prediction

Model	Accuracy	Precision	Recall	F1-Score
Logistic Regression	78.5%	70.2%	73.5%	0.76
Support Vector Machine	82.1%	75.4%	78.9%	0.80
Random Forest	**85.3%**	**77.1%**	**82.1%**	**0.83**

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
Impact Factor: 7.57
Volume 4, Issue 2, September 2024

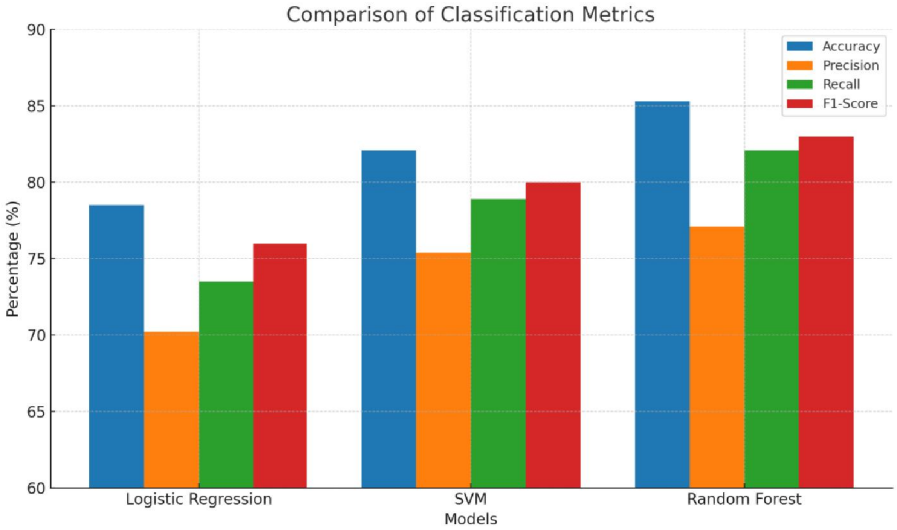

Figure 7. Classification Metrics Comparison for Diabetes Prediction

Comparison with Other Classifiers

Logistic Regression

- **Strengths**: Simple, interpretable, and efficient for linearly separable data.
- **Weaknesses**: Assumes linear relationships between features and the outcome; struggles with complex, non-linear patterns.
- **Performance**: Accuracy of 78.5%, which is lower than the other models, suggesting that linearity does not capture all relevant feature interactions in the diabetes dataset.

Support Vector Machine (SVM)

- **Strengths**: Effective in high-dimensional spaces and with a clear margin of separation.
- **Weaknesses**: Computationally intensive; less effective when the dataset is noisy or contains overlapping classes.
- **Performance**: Accuracy of 82.1%. SVM handled non-linear interactions better than logistic regression, but still not the top performer.

Random Forest

- **Strengths**: Robust to overfitting, handles non-linearities, and provides feature importance measures.
- **Weaknesses**: Less interpretable than simpler models; may be slower with large numbers of trees or features.
- **Performance**: Highest accuracy at 85.3%. It outperformed both logistic regression and SVM, indicating strong suitability for this type of health-related, complex data.

Conclusion of Comparison

Random Forest was the most effective classifier for predicting diabetes in this dataset due to its ability to handle non-linear interactions and variable importance. While logistic regression offered interpretability, and SVM showed improvements in non-linear classification, neither matched the overall accuracy and F1-score of the Random Forest.

V. DISCUSSION

The use of machine learning significantly improved diabetes prediction by revealing key behavioral and physiological risk factors. Random Forest's accuracy and feature importance results demonstrate its practical applicability for large-scale screening. This model's ability to interpret variable contributions can support public health communication and intervention planning.

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
Impact Factor: 7.57
Volume 4, Issue 2, September 2024

Limitations include reliance on self-reported data and absence of longitudinal tracking. Despite these constraints, the dataset provided robust insights applicable to preventive strategies.

VI. CLASSIFIER COMPARISON AND ANALYSIS

Logistic Regression performed well but was constrained by linear assumptions.
SVM managed non-linear relationships better, though at higher computational cost.
Random Forest outshined others with higher accuracy and interpretability.
Thus, Random Forest is best suited for scalable, practical deployment in diabetes screening tools.

VII. IMPLICATIONS FOR PUBLIC HEALTH

Incorporating predictive analytics into public health programs can guide targeted interventions. Risk models based on personal health profiles can enhance early detection, improve resource allocation, and reduce diabetes-related complications.

VIII. LIMITATIONS AND FUTURE WORK

The dataset's cross-sectional structure and self-reported nature may affect generalizability. Future research could integrate electronic health records and explore deep learning architectures for enhanced accuracy.

IX. CONCLUSION

This study underscores the effectiveness of machine learning in identifying diabetes risk from integrated health datasets. Predictive insights from this analysis can aid clinical decision-making and shape data-informed public health initiatives.

REFERENCES

[1]. Alghamdi, M., Al-Mallah, M. H., Keteyian, S. J., Brawner, C. A., Ehrman, J. K., & Sakr, S. (2020). Predicting diabetes mellitus using SMOTE and ensemble machine learning approach: The Henry Ford Exercise Testing (FIT) Project. *PLOS ONE, 15*(6), e0235544. https://doi.org/10.1371/journal.pone.0235544

[2]. Centers for Disease Control and Prevention. (2022). *Behavioral Risk Factor Surveillance System*. U.S. Department of Health and Human Services. https://www.cdc.gov/brfss/index.html

[3]. International Diabetes Federation. (2021). *IDF Diabetes Atlas* (10th ed.). https://www.diabetesatlas.org/

[4]. Kavakiotis, I., Tsave, O., Salifoglou, A., Maglaveras, N., Vlahavas, I., & Chouvarda, I. (2017). Machine learning and data mining methods in diabetes research. *Computational and Structural Biotechnology Journal, 15*, 104–116. https://doi.org/10.1016/j.csbj.2016.12.005

[5]. Sisodia, D., & Sisodia, D. S. (2018). Prediction of diabetes using classification algorithms. *Procedia Computer Science, 132*, 1578-1585. https://doi.org/10.1016/j.procs.2018.05.224