Diabetes Detection: Predicting Type II Diabetes Using Machine Learning

Yujun Ma
Team project (co-authors omitted in public version)

***Abstract*—This project aims to predict type II diabetes using machine learning models and identify the one with the best performance, followed by recommendations for prevention strategies. To enhance data quality and address abnormal values, three distinct imputation methods were explored, and the importance of feature standardization for different models was evaluated. Four models were constructed for comparison: logistic regression with Ridge regularization, support vector machine, Random Forest, and a simple neural network with one hidden layer. Key efforts included feature engineering and hyperparameter tuning through cross-validation to optimize model performance. Additionally, feature importance analysis offered valuable insights into the relationships between health metrics and diabetes risk.**

***Index Terms*—Diabetes Management, Imputation Techniques, Logistic Regression, Ridge Regularization, Support Vector Machine, Random Forest, Neural Network**

I. Introduction

With the increasing emphasis on personal health management, diabetes has become one of the most highly concerning potential health issues. Diabetes is a chronic metabolic disorder characterized by insufficient insulin production or impaired insulin utilization, leading to persistent hyperglycemia. Over time, this condition can cause significant damage to various bodily systems, particularly nerves and blood vessels, often resulting in complications such as cardiovascular disease, stroke, kidney failure, and blindness. According to the World Health Organization (WHO), the prevalence of diabetes has increased substantially, from 108 million people in 1980 to 422 million in 2014, highlighting the growing global health burden of this disease. Individuals with diabetes are at least twice as likely to die prematurely as those without diabetes. In 2012 alone, diabetes caused approximately 1.5 million deaths and is projected to become the seventh leading cause of death by 2030 [1].

Diabetes is generally classified into two main types: type I and type II. Type I diabetes is an autoimmune disorder characterized by the destruction of pancreatic cells β, requiring lifelong insulin injections. Unfortunately, current medical research has not yet identified clear methods for preventing type I diabetes. In contrast, type II diabetes accounts for the majority of diabetes cases and is primarily associated with inadequate insulin secretion or insulin resistance, often caused by unhealthy lifestyle factors such as obesity, lack of physical activity, and poor dietary habits. Type II diabetes is largely preventable and manageable [2].

The data set used for this project was originally collected by the National Institute of Diabetes and Digestive and Kidney Diseases (NIDDK) and subsequently made available on Kaggle by Akshay Dattatray Khare [3]. The data set comprises vital health statistics and Type II diabetes status of individuals 21 years and older who are women of Pima Indian heritage. The Pima community has been the subject of extensive research by NIDDK since 1965 due to their notably high prevalence of Type II diabetes, making it an important population for studying risk factors and diabetes prevention strategies [4].

II. Exploratory Data Analysis

A. Data Structure

The data set contains 768 instances, including 268 diagnosed cases of type II diabetes. It includes 8 numerical features representing various health measurements, along with a factor response variable, "Outcome". Table I provides a summary of all the variables used in the analysis. In particular, the "Diabetes Pedigree Function" is a feature that estimates an individual's likelihood of developing diabetes based on family history and genetic relationships with relatives who have diabetes.

TABLE I
Feature Descriptions and Ranges

Feature	Description	Range
Pregnancies	Times of pregnancies	[0, 17]
Glucose	Glucose concentration (mg/dL)	[0, 199]
BloodPressure	Diastolic blood pressure (mmHg)	[0, 122]
SkinThickness	Triceps skin fold thickness (mm)	[0, 99]
Insulin	2-Hour serum insulin (mu U/mL)	[0, 846]
BMI	Body mass index	[0, 67.1]
DiabetesPedigreeFunction	Scores likelihood of diabetes based on family history	[0.078, 2.42]
Age	Age in years	[21, 81]
Outcome	Diabetes: 1 (Positive), 0 (Negative)	[0, 1]

B. Data Pre-processing

The data set is highly complete and does not contain any missing values. To improve data quality, we verified the range of values for all predictors by consulting medical knowledge to ensure that they fall within reasonable limits. We found that zeros in all predictors, except for "pregnancies," are considered abnormal based on medical standards. For example, a skin thickness value of zero is physiologically impossible. Such

zero values may be the result of measurement errors or could represent extreme small values. Removing all rows where features other than ”pregnancies” have a value of zero would result in significant data loss, reducing the sample size from 768 to 392, which is not feasible. Therefore, we plan to address these zero values using various imputation methods, including mean, median and medically meaningful minimum imputation.

Mean imputation is appropriate for data that are continuously and normally distributed, but is sensitive to outliers and skewed distributions, as it can undermine the overall variability of the data [5]. As shown in Fig.1, the variables ”Insulin” and ”SkinThickness” contain outliers and exhibit right skewness.

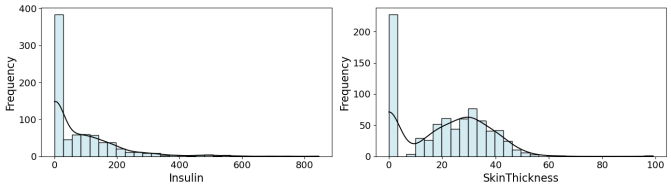

Fig. 1. Features contain outlier and skew distributed

We also devised a method called medically meaningful minimum imputation, where zeros are replaced with a reasonable minimum value based on medical standards. However, this approach has no established evidence to support its effectiveness in the literature. Moreover, we observed that for some features, the medically meaningful minimum values were higher than the nonzero minimum values in the dataset. For example, according to the American Diabetes Association, the recommended fasting glucose target range for diabetes patients is 80-130 mg / dL [6], while the non-zero minimum glucose value in this dataset is 44 mg/dL. We aim to avoid altering the distribution of each feature to maintain the integrity of the data. Therefore, we opted for median imputation, which is suitable for data with skewed distributions or the presence of outliers, as it is robust against outliers. Refer to “(1)”, Median imputation replaces abnormal values with the median of observed values for each feature [5].

$$\hat{x}_{\text{mis}} = \begin{cases} x_{(s)}, & \text{if } n_o \text{ is odd} \\ \frac{x_{(s)}+x_{(s+1)}}{2}, & \text{if } n_o \text{ is even} \end{cases} \quad (1)$$

Where $x(s)$ represents the median observed value and n_o is the size of the observed values. The updated ranges for features previously containing abnormal values are presented in Table II.

TABLE II
FEATURE RANGE COMPARISON

Feature	Range (Before)	Range (After)
Glucose	[0, 199]	[44, 199]
BloodPressure	[0, 122]	[24, 122]
SkinThickness	[0, 99]	[7, 99]
Insulin	[0, 846]	[14, 846]
BMI	[0, 67.1]	[18.2, 67.1]

C. Visualization

1) Relationship between response and single predictor: Initially, we employed paired plots to identify feature distributions that may contribute to classification. The complete paired plot could be found in *Appendix A*. Our focus was primarily on the diagonal plots, which display the Kernel Density Estimate of each variable, grouped by diabetes status: positive cases (orange) and negative cases (blue), as shown in Fig.2.

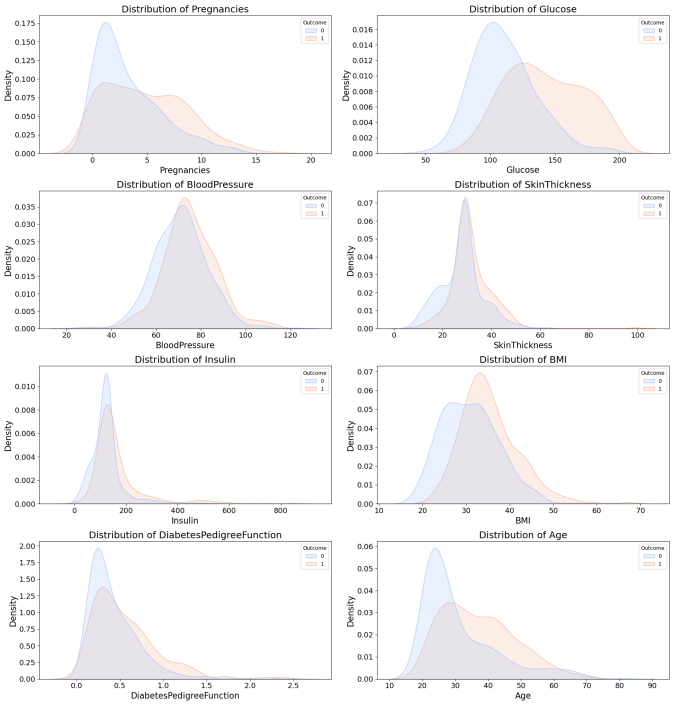

Fig. 2. KDE plot of features by Diabetes Status

Some preliminary conclusions can be drawn:

- All features demonstrate a positive association with the response variable.
- Women with a higher number of pregnancies are more likely to be diagnosed with diabetes
- Diabetic instances tend to have higher glucose levels (>120 mg/dL), whereas non-diabetic individuals exhibit a lower distribution of glucose values.
- Diabetic instances generally have higher BMI values.
- The proportion of Diabetic instances is also higher in the older age group (>50 years), whereas non-diabetic instances are more prevalent among younger age groups.
- Diabetic instances have a higher density in the region with higher Diabetes Pedigree Function scores (>0.6).
- In contrast, the distributions of Blood Pressure, Skin Thickness, and Insulin exhibit considerable overlap between the positive and negative cases, suggesting they may contribute less to the classification.

2) Correlation Heatmap: Secondly, we examined the correlations between features using a correlation coefficient heatmap to assess potential multi-collinearity. In Fig.3, positive correlations are colored red, while negative correlations are colored blue.

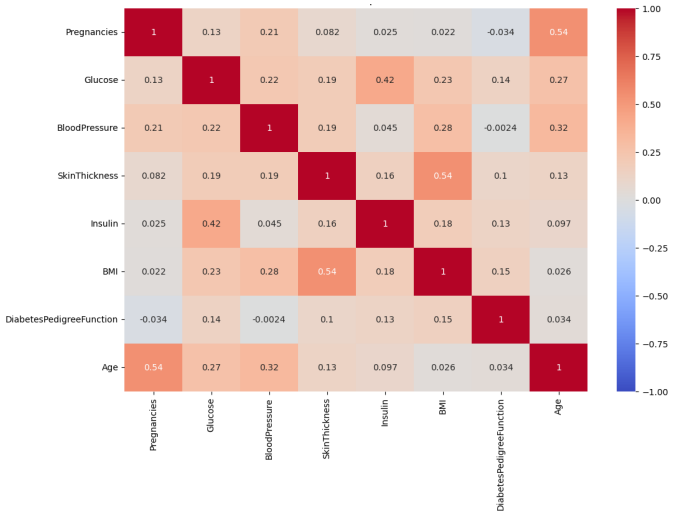

Fig. 3. Correlation Heatmap

From the heatmap, it can be observed that most features exhibit weak correlations with each other, suggesting minimal multicollinearity issues. However, the relatively strong correlations between 'SkinThickness' and 'BMI', as well as 'Age' and 'Pregnancies', are noteworthy. For a linear model such as logistic regression, it is essential to address multicollinearity. We opted to use Ridge regularization, which shrinks the coefficients and helps stabilize the model. Lasso regularization was not chosen because, in our previous KDE analysis, 'BMI' and 'Age' demonstrated significant importance for classification, and we do not want to completely exclude these variables from the model.

III. Analysis

A. Data Splitting and Standardization

We divide the data set into training data and testing data 80% and 20%, resulting in 614 training samples and 154 testing samples. The response variable in our target data set exhibits an imbalanced distribution, with roughly 65% negative cases and 35% positive cases. To ensure that the original class distribution is maintained in both the training and the testing datasets, we employed a stratified data splitting strategy rather than a purely random split. This approach ensures that both the training and the test sets remain representative of the original data set, thereby enhancing the generalization of the model to unseen data and improving the reliability of the evaluation metrics. Specifically, we implemented this stratification by setting the parameter `stratify=y` within the `train_test_split` function.

Models such as logistic regression with Ridge penalty, Support Vector Machines, and neural networks generally perform better when data are scaled to a standard range, as these models are particularly sensitive to feature scaling. Hsu, Chang, and Lin (2003) emphasize the importance of scaling in their practical guide to SVM classification, noting that it helps prevent attributes with larger numerical ranges from dominating those with smaller numeric ranges, while also mitigating numerical difficulties during computation [7]. For the aforementioned models, all features were standardized to ensure equal contribution to the model.

In contrast, Random Forests are generally less sensitive to feature scaling due to their reliance on hierarchical splits and their ensemble nature, which mitigates the effects of varying feature scales. A study by Ozsahin et al. found that, in the context of diabetes prediction, the F1 score of a Random Forest model trained on unstandardized data was 75%, while standardization of characteristics resulted in a reduced F1 score of 71% [8]. Based on these findings, we chose to utilize unstandardized data for the Random Forest model to maintain optimal predictive performance.

B. Logistic Regression

The initial model in our workflow is a standard logistic regression model incorporating Ridge regularization. To determine the optimal regularization strength ($C = 1/\lambda$), we use `GridSearchCV`, which iterates over a range of possible values for C, selecting the value that produces the highest validation score. Subsequently, we fit the model using the optimal regularization strength. To ensure robust model evaluation, we applied *Stratified 5-Fold Cross-Validation* in the grid search stages, with the F1 score serving as the evaluation metric. This stratified approach maintains a consistent class distribution across all folds, thus mitigating the risk of bias from class imbalance and ensuring the evaluation metrics are representative of the entire dataset. The F1 score was chosen specifically for its ability to balance precision and recall, which makes it well suited to evaluate performance in the positive class.

The results indicated that the optimal regularization strength was 1, yielding a mean F1 score of 0.6521 through cross-validation, which fell short of our expectations. Given the limited performance, we hypothesized that the standard logistic regression model might be insufficient to capture potential nonlinear relationships between features. To address this limitation, we extended the model by incorporating second-degree polynomial features, thereby enabling it to capture potential nonlinear patterns that the initial linear model could not. We subsequently tuned the regularization strength and evaluated the extended model using the same cross-validation procedure and scoring metric as used for the standard logistic regression. The results demonstrated that the optimal regularization strength for this enhanced model was 0.01, indicating a stronger penalty for parameter values, and the mean F1 score improved to **0.6647**.

C. Support Vector Machines

Support Vector Machine (SVM) is a widely adopted and effective algorithm to solve binary classification tasks. The primary objective of SVM is to construct an optimal hyperplane that maximally separates two classes of data, ensuring that the hyperplane is positioned as far as possible from the closest members of each class. When the data set is linearly separable, the SVM algorithm identifies a line that divides the two groups of data points. As illustrated by Tristan Fletcher in Fig.4, the hyperplanes H_1 and H_2 represent the boundaries

that contain the data points of two different classes. The data points lying on the decision boundary on either side are known as "support vectors". The margin, which is defined as the equidistant space d_1 and d_2 between the hyperplane and each boundary, represents the gap that SVM aims to maximize for optimal class separation [9].

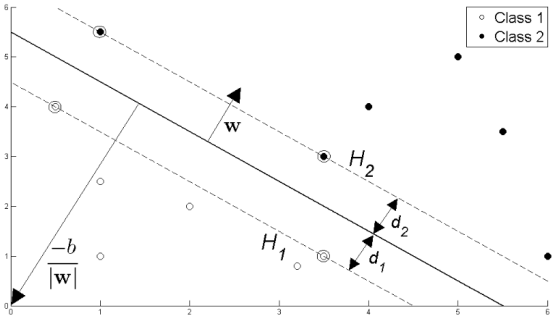

Fig. 4. Linear SVM Hyperplane

In situations where the data are not linearly separable, SVM uses kernel functions to map the input features to a higher-dimensional space. The Radial Basis Function (RBF) kernel, which is commonly employed in such cases, transforms the data such that it becomes linearly separable in the new feature space, allowing for an effective classification using a linear hyperplane, As shown in Fig.5.

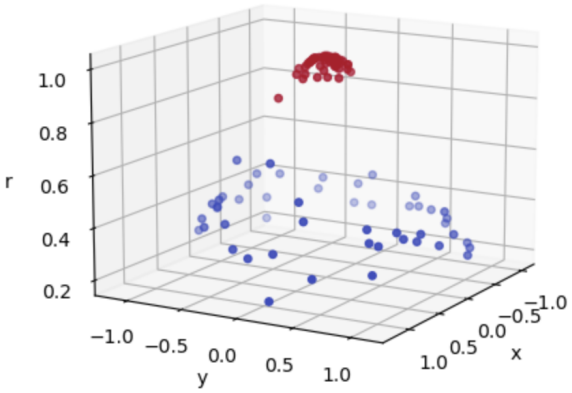

Fig. 5. RBF Kernel Transformation

During model construction, our primary focus was on tuning the hyper-parameters including `kernel`, `C` and `gamma`. We evaluated both the RBF and linear kernels to identify the most suitable model to capture data patterns. The regularization parameter (C) controls the trade-off between low training error and low testing error; a higher value allows the model to fit the training data more closely, potentially leading to overfitting. Gamma (the kernel coefficient for the RBF) determines the influence of individual training instances, with higher values increasing the sensitivity of the model, potentially leading to overfitting. The final model used the RBF kernel with `C=100` and `gamma=0.001`, allowing for sufficient model flexibility while mitigating over-fitting risks. The resulting model achieved an F1 score of **0.6603**. Although efforts were made to enhance the model performance by expanding the hyperparameter search space, the results remained consistent.

D. Random Forest

The approach for training the Random Forest model closely resembles that used for SVM. We tuned four hyperparameters using grid search combined with Stratified 5-Fold Cross-Validation, employing the F1 score as the evaluation metric. As noted in previous section, the model was trained using unstandardized features.

- `n_estimators` determines the number of decision trees used in the forest, where a higher number improves stability by reducing variance.
- `max_depth` controls the maximum depth of each decision tree, with smaller depths preventing overfitting by reducing complexity.
- `min_samples_split` and `min_samples_leaf` specify the minimum number of samples required to split an internal node or define a leaf , with higher values limiting the complexity of the tree.

The search grid and the optimal values obtained are presented in Table III. This configuration contributed to enhancing the model's robustness and its ability to generalize to unseen data. The best F1 score achieved during cross-validation was **0.6536**.

TABLE III
OPTIMAL HYPERPARAMETERS FOR THE RANDOM FOREST MODEL

Parameter	Grid	Optimal Value
`n_estimators`	[100, 200, 300, 400]	200
`max_depth`	[10, 15, 20, None]	15
`min_samples_split`	[5, 10, 15]	6
`min_samples_leaf`	[2, 4, 6]	5

Fig.6 shows the feature importance plot from the Random Forest model, where each bar represents how much each feature contributed to the model's decisions. The result aligns with the conclusion we draw previously. 'Glucose' is the most influential predictor, followed by 'BMI', 'Diabetes pedigree function' and 'Age'. The other four features are considered less significant.

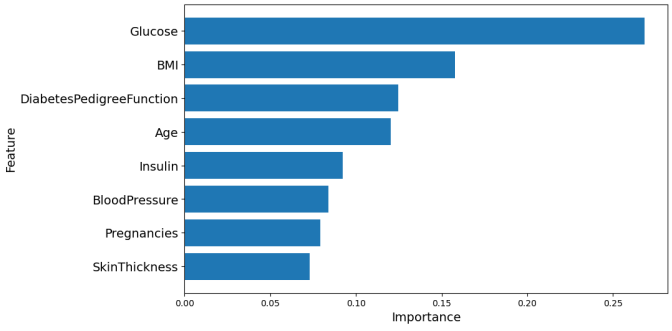

Fig. 6. Feature Importance Plot

E. Neural Network

The neural network implemented in this study is a fully connected feedforward network with a hidden layer. As shown in Fig.7, it consists of an input layer with 8 neurons representing the predictor variables, a single hidden layer with 4 neurons utilizing the ReLU activation function to introduce non-linearity, and an output layer with 2 neurons that produce the non-normalized logits for the binary classification task (diabetic vs. non-diabetic).

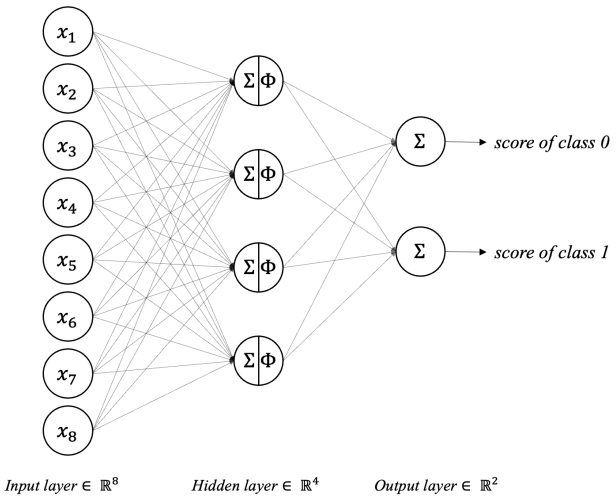

Fig. 7. Architecture of the neural network

To improve convergence and ensure equal contribution of features, we standardized the data. Instead of cross-validation, we created a set of validations (10% of the training data) using stratified sampling to reduce computational cost while effectively evaluating the generalizability of the model.

The `CrossEntropyLoss` function was selected as it is well suited for binary or multiclass classification tasks, measuring the dissimilarity between predicted probabilities and true class labels. The model was trained using the `Adam` optimizer, chosen for its ability to handle non-convex optimization problems and its reduced sensitivity to hyperparameter tuning. A learning rate of 0.001 was used to balance convergence speed and stability, while a batch size of 64 ensured stable gradient updates without excessive computational cost. The model was trained for 4000 epochs, with the validation F1 score monitored during training. To avoid overfitting, the model with the highest validation F1 score was saved.

The final model achieved a best validation F1 score of **0.8**. Training loss steadily decreased, indicating effective learning, while the validation F1 score improved during the early epochs and stabilized after approximately 500 epochs. The training process and results are visualized in Fig.8.

Attempts to improve the model by adding an additional hidden layer and increasing the number of neurons did not yield better results. The modified architecture achieved a lower validation F1 score of 0.7568, likely due to the limited size of the dataset, which may have caused the more complex model to overfit or struggle to generalize effectively.

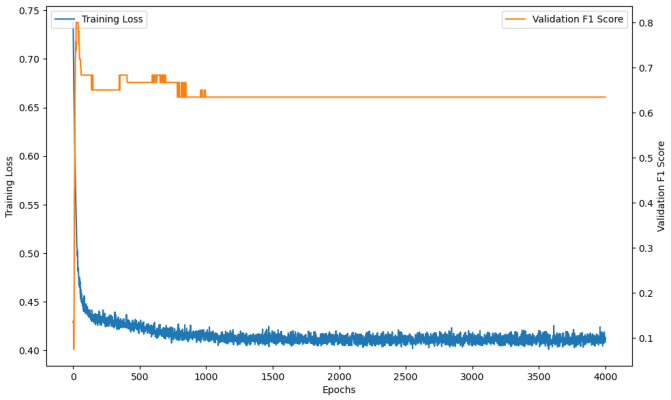

Fig. 8. Training Loss and Validation F1 Score over Epochs

IV. Result

The performance of four models on the test dataset was evaluated using accuracy, F1 score, precision, and recall as metrics. As shown in Table IV, Random Forest outperforms the other models across most metrics, except for recall, where it is slightly outperformed by Logistic Regression. This suggests that Random Forest is the most robust model for identifying diabetic instances in general.

TABLE IV
Comparison of Machine Learning Models

Model	Accuracy	F1 Score	Precision	Recall
Logistic Regression	0.7143	0.5926	0.5926	0.5926
SVM	0.6948	0.5253	0.5778	0.4815
Random Forest	0.7532	0.6122	0.6818	0.5556
Neural Network	0.7078	0.5263	0.6098	0.4630

Additionally, Receiver Operating Characteristic (ROC) and Precision-Recall (PR) curves were generated to evaluate the models' performance. As shown in Fig.9, both Logistic Regression and Random Forest achieved the highest Area Under the ROC Curve (AUC-ROC) of 0.81, demonstrating superior performance in differentiating between diabetic and non-diabetic instances.

Given the imbalanced nature of the data set, the Area Under the Precision-Recall Curve (AUC-PRC) provides more meaningful insights. As illustrated in Fig.10, Logistic Regression and Random Forest again outperformed other models, achieving the highest AUC-PRC values of 0.67. This reflects their ability to maintain a better balance between precision and recall, a critical factor in medical diagnostics to reduce false positives and false negatives.

V. Conclusion

The feature importance analysis provides critical information on the variables that influence the prediction of type II diabetes, as identified by the Random Forest model. Key predictors include glucose, the most significant characteristic, which is strongly correlated with diabetes risk and reflects its role as a primary indicator of the condition. BMI is the second most influential factor, highlighting the connection between increased body fat and increased risk of insulin resistance and

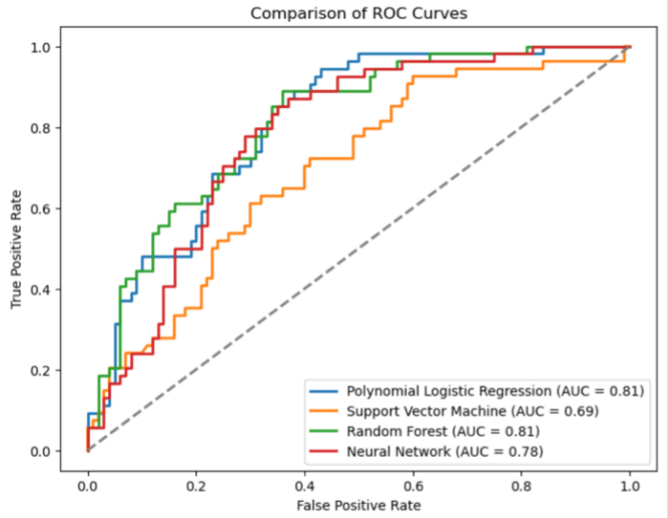

Fig. 9. Comparison of ROC Curves for All Models

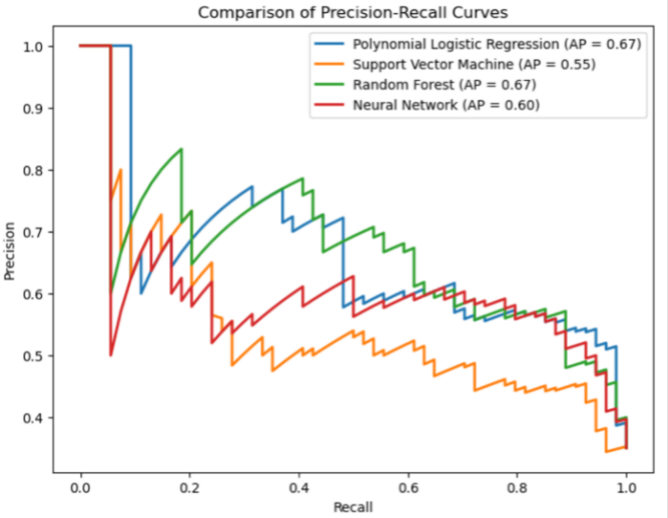

Fig. 10. Comparison of PR Curves for All Models

diabetes. The ”diabetes pedigree function” captures genetic predisposition, with higher values indicating a stronger family history. ’Age’ also plays a significant role, as older individuals are at greater risk due to decreased insulin sensitivity and pancreatic function.

Based on these findings, several prevention strategies are proposed. Regular monitoring of blood glucose levels can facilitate early detection and management, particularly in high-risk individuals. Maintaining a healthy weight through a balanced diet and regular physical activity can lower the risk of insulin resistance. For those with a family history of diabetes, understanding genetic risks and adopting proactive lifestyle changes, including regular medical check-ups, are essential. Age-specific interventions for older adults should focus on improving insulin sensitivity through dietary modifications and exercise. Lastly, public health initiatives should encourage a healthy lifestyle, promote balanced nutrition, physical activity, and regular health screenings to address modifiable risk factors. Collectively, these measures aim to mitigate the prevalence of Type II diabetes and improve public health outcomes.

ACKNOWLEDGMENT

We extend our thanks to the many scholars and researchers whose contributions to diabetes research have significantly advanced understanding and awareness of this critical condition, laying the groundwork for our study. Our appreciation goes to Western University for fostering an academic environment that encourages rigorous inquiry and collaboration, and to our peers for their construc-tive feedback and support throughout this journey. Finally, we acknowledge the individuals who shared their experiences with diabetes, inspiring us to pursue research that not only advances scientific knowledge but also addresses real-world challenges.

REFERENCES

[1] W. H. Organization *et al.*, “Global report on diabetes,” 2016.
[2] R. Kumar, P. Saha, Y. Kumar, S. Sahana, A. Dubey, and O. Prakash, “A review on diabetes mellitus: type1 & type2,” *World Journal of Pharmacy and Pharmaceutical Sciences*, vol. 9, no. 10, pp. 838–850, 2020.
[3] Kaggle, “Diabetes dataset,” oct 2022, [Online]. [Online]. Available: https://www.kaggle.com/datasets/akshaydattatraykhare/diabetes-dataset
[4] J. W. Smith, J. E. Everhart, W. Dickson, W. C. Knowler, and R. S. Johannes, “Using the adap learning algorithm to forecast the onset of diabetes mellitus,” in *Proceedings of the annual symposium on computer application in medical care*. American Medical Informatics Association, 1988, p. 261.
[5] M. Mohammed, H. Zulkafli, M. Adam, N. Ali, and I. Baba, “Comparison of five imputation methods in handling missing data in a continuous frequency table,” in *AIP Conference Proceedings*, vol. 2355, no. 1. AIP Publishing, 2021.
[6] N. A. ElSayed, G. Aleppo, V. R. Aroda, R. R. Bannuru, F. M. Brown, D. Bruemmer, B. S. Collins, M. E. Hilliard, D. Isaacs, E. L. Johnson *et al.*, “6. glycemic targets: standards of care in diabetes—2023,” *Diabetes care*, vol. 46, no. Supplement_1, pp. S97–S110, 2023.
[7] C.-W. Hsu, “A practical guide to support vector classification,” *Department of Computer Science, National Taiwan University*, 2003.
[8] D. U. Ozsahin, M. T. Mustapha, A. S. Mubarak, Z. S. Ameen, and B. Uzun, “Impact of feature scaling on machine learning models for the diagnosis of diabetes,” in *2022 International Conference on Artificial Intelligence in Everything (AIE)*. IEEE, 2022, pp. 87–94.
[9] T. Fletcher, “Support vector machines explained,” *Tutorial paper*, vol. 1118, pp. 1–19, 2009.

APPENDIX
PAIRED PLOT