Evaluation of Predictive Models for Diabetes Using Machine Learning

Tabasum Hamdard

Michigan Technological University
1400 Townsend Drive, Houghton, Michigan 49931-1295 USA

Abstract

This study focuses on evaluating several predictive models to help in detecting individuals at risk of diabetes. The dataset has been selected to predict diabetes using several demographic information, patient history, and laboratory features. The data has 100,000 observations with 9 features. Our methodology involves comparing several models with various predictive metrics using a range of classification algorithms that include Logistic Regression, K-Nearest Neighbor (KNN), Naive Bayes, Decision Trees, Random Forest, and Gradient Boosting. This study shows that compared to other models, Naive Bayes was one of the highest performing model to predict diabetes based on our chosen metrics (recall and specificity) after utilizing the balancing techniques. The findings highlight the major role that advanced classification techniques have in facilitating early identification of diseases to enable timely medical interventions and improve healthcare resource allocation. This paper sets the foundation for future research into improving chronic diseases prediction models and demonstrating the effectiveness of data mining techniques in healthcare applications.

Introduction

Diabetes, or diabetes mellitus, is a metabolic disease caused by high blood sugar, typically develops after prolonged exposure to factors that compromise insulin production or function (World Health Organization 2023). This is one of the most common diseases that affects millions of individuals around the world. If left untreated, it leads to many other health issues, such as different cardiovascular diseases, loss of vision, renal failure and neuropathy (Eid et al. 2019). Although we have numerous technological advancements in medical sciences, the early diagnosis of diabetes still remains a challenge, which emphasizes the importance of innovative strategies to improve disease detection rates (Guan et al. 2023).

In this paper, we have explored the importance of data mining and machine learning methods in predicting diabetes. Our study is motivated by using such innovative technologies that have the potential to transform healthcare and provide early personalized treatments for patients with a particular disease.

The approach to provide the models with the high quality data began with exploring any discrepancies in the dataset, followed by cleaning and preprocessing the data. The study's main goal was to use several classification algorithms, such as Logistic Regression, K-Nearest Neighbors (KNN), Naive Bayes, Decision Trees, Random Forest, and Gradient Boosting, compare them and choose the most effective model for this task. Additionally, each model's ability was evaluated to handle imbalanced data with stratified sampling and compared their performances on both balanced and imbalanced datasets.

After the final evaluation, Na¨ıve Bayes seemed to perform well in diagnosing medical conditions in imbalanced data, showing promising results. On the contrary, ensemble models like Random Forest and Gradient Boosting managed to maintain consistent and balanced predictions across different metrics, both before and after balancing the data. This suggests they perform well in handling balanced datasets.

Related Work

Diabetes Prediction has been extensively studied using many different machine learning algorithms, as it's early detection is crucial for preliminary treatment. These studies have employed several approaches to enhance the accuracy and efficiency of predictive models. This section reviews key studies that have contributed to this area of research, highlighting their findings as well as fitting our work within the ongoing discussion.

One notable study (Ataya 2023) emphasized the use of several machine learning algorithms like Logistic Regression, k-Nearest Neighbors, Support Vector Machine, Random Forests and advanced boosting methods, such as XGBoost and LightGBM for diabetes prediction. The results of their study demonstrated that LightGBM has the highest accuracy (88.5%) among other models to detect this chronic disease. This study is important as it provides a detailed comparative analysis of multiple algorithms, while focusing on performance optimization in clinical decision support systems.

Similarly, in another study a group of researchers explored the effectiveness of boosting techniques in predicting diabetes using datasets from different geographical regions (Shampa, Islam, and Nesa 2023). This study highlighted the use of AdaBoost, CatBoost, and Gradient Boosting to improve the prediction accuracies. The performance of the models were analysed based on their robustness in handling varied dataset characteristics, especially those from Bangladesh, which outperformed other datasets.

Additionally, (D'Souza, Shah, and Singh 2022) studied the application of multiple machine learning models that include K-nearest Neighbor, Naive Bayes and Decision Trees, to classify type 2 diabetes. The research was effective in detecting a model with the highest accuracy, while including both supervised and unsupervised learning techniques. This reflects a distinct approach to understanding and modeling the complex nature of diabetes prediction.

Our approach uses a similar range of classification techniques, however, we focus on comparing these models through specific metrics. Unlike aforementioned studies that primarily focused on accuracy as a metric, our study prioritizes recall and specificity to ensure the high sensitivity required for medical diagnostics.

In contrast to other studies, our research engages with these works by applying and comparing a variety of classification algorithms, while also showing the complexities of data preprocessing and prioritizing different performance metrics. Our study ensures that the model not only predicts accurately but also barely misses cases of diabetes in patients.

Data

This project uses a dataset from Kaggle called 'Diabetes Prediction Dataset' (Mustafa n.d.), which includes a comprehensive set of data encompassing both medical histories and demographic details of patients. This structured dataset contains 100,000 observations with 9 numerical and categorical features. The variables in this dataset are: gender, age, hypertension, heart disease, smoking history, BMI, HbA1c level, blood glucose level, and our target variable, diabetes. Table 1 describes the information related to each feature of the dataset.

After a thorough examination, we found that there were no null (NaN or NA), incomplete, inconsistent, or duplicate attribute values in the dataset. However, we observed approximately 5% imbalance ratio in our dataset, which was handled accordingly before running the models.

Table 1: Diabetes Dataset Information

Feature	**Variable**	**Variable type**	**Value type**
Gender	gender	Nominal	Male or Female
Age	age	Ratio	int (0-80 years)
Hypertension	hypertension	Nominal	0 or 1
Heart disease	heart_disease	Nominal	0 or 1
Smoking history	smoking_history	Nominal	No info, Never, other
BMI	bmi	Ratio	float
HbA1c level	HbA1c_level	Ratio	float
Blood glucose level	blood_glucose_level	Ratio	int (mg/dL)
Diabetes	diabetes	Nominal	0 or 1

Correlation Heatmap To understand and analyze the relationship between the variables, we developed a correlation heatmap as demonstrated in Figure 1. It indicates that the levels of blood glucose and HbA1c are the most highly correlated features with the target variable 'diabetes,' showing a 0.4 correlation ratio. These features are likely to be important predictors in the diabetes prediction model.

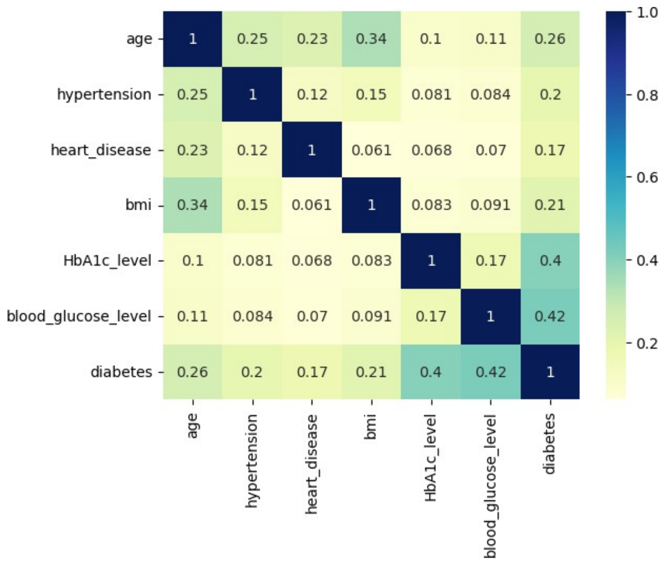

Figure 1: Correlations heatmap between all attributes

Methods

The initial stage of this study involves preprocessing the data, followed by applying a range of machine learning algorithms as a part of the experimental design. Figure 3 outlines the methodology adopted in this research.

This study attempts to evaluate metrics to assess the predictive performance both before and after implementing Synthetic Minority Oversampling Technique (SMOTE). This method created new and synthetic examples for the underrepresented class to balance the data. First, only stratified sampling was used without implementing SMOTE followed by a second analysis on data incorporating SMOTE. By comparing both of these metrics we gain insights into how this technique impacts the effectiveness of our machine learning models in predicting diabetes.

Data Preprocessing

This stage of methodology centers around preparing the dataset for analysis. Initially, it was confirmed that the dataset had no null values and constant attributes. Further preprocessing steps involved checking for redundant values, outlier removal, categorical encoding, addressing the imbalanced data and data scaling:

Redundant Values Removal: As the correlation analysis in Figure 1 indicated no correlations higher than 0.5 between attributes, it can be concluded that there are no redundant attribute values in our dataset.

Outlier Removal: After reviewing the numerical features we identified and removed the outliers to refine the dataset to 90,387 records. This step was critical in preventing any potential skewness in our models' performance due to anomalous data points, ensuring the accuracy of our model training and predictions.

Categorical Encoding: Since machine learning algorithms require numerical input and output variables, in this step variables such as 'gender' and 'smoking history' were encoded numerically through one-hot encoding, making them suitable for model input. This step helps with the integration of additional data into our predictive models and prevents misinterpretation of the categorical data as ordinal data.

Resolving the Class Imbalance: As mentioned in the data section, the dataset has a class imbalance, where class 1 is substantially less represented compared to class 0, with imbalance ratio of 0.05. To solve this problem, SMOTE technique has been used to oversample the minority class. SMOTE generates synthetic samples for the minority class (class 1) to balance the class distribution and help the model learn effectively.

Data Scaling: Standard Scaler was used as feature scaling method which was applied to standardize the range of input features to ensure that they have a mean of 0 and a standard deviation of 1. This ensures that all the features equally contribute to the model, regardless of their previous scale.

Experimental Design

The initial step in the experimental design was to split the data after the model is trained on a preprocessed dataset. The data was divided into train/validation and test sets in an 80:20 ratio. The training set is used to train the model, while the test set is used to evaluate its performance. Once the train/validation and test values are scaled thereafter a sample of this training set is then used for training.
A random sampling technique is utilized to select 20% of the original training set as a representative subset to apply machine learning algorithms. Hence, all the models are trained and tuned on a sample with the size of 14461 records.

In this study a total of six classifiers were chosen and used in two instances: first on the stratified data subset, and subsequently on a balanced subset created using SMOTE. This helped to determine the best model for diabetes prediction. The selected models include:

- **K-Nearest Neighbors (KNN)** is a classic, non-parametric algorithm that works by finding the K-nearest neighbors to a new data point in the training set and using the majority class of those neighbors to classify the new data point.
- **Na¨ıve Bayes** is a probabilistic classifier based on applying Bayes' theorem with strong (na¨ıve) independence assumptions between the features.

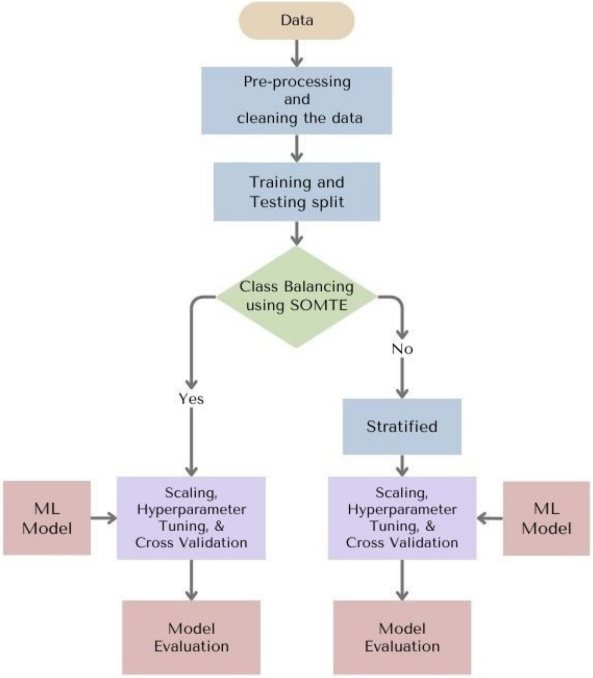
Data
Pre-processing and cleaning the data
Training and Testing split
Class Balancing using SOMTE
Yes
No
Stratified
ML Model
Scaling, Hyperparameter Tuning, & Cross Validation
Scaling, Hyperparameter Tuning, & Cross Validation
ML Model
Model Evaluation
Model Evaluation

Figure 2: Methodology Flowchart

- **Decision Trees** use a hierarchical structure of binary decisions to classify data, making them interpretable and effective for both classification and regression tasks.
- **Logistic Regression** is a linear model used for binary classification, providing probabilities of class membership based on input features.
- **Random Forest** is an ensemble method known for its high accuracy and ability to handle imbalanced data.
- **Gradient Boosting** is an ensemble technique that builds models in a sequential manner, where each subsequent model corrects errors made by the previous ones, leading to highly accurate predictions.

For each model, we carefully tuned the hyper-parameters using grid search with five-fold cross-validation. This process involves systematically testing a range of hyper-parameters for each model to find the optimal combination that maximizes performance. In this case, on the imbalanced data, we tuned the hyperparameters based on the default scoring of the GridsearchCV. For the balanced data subset, we used recall as scoring, which measures the proportion of true positives. This means our model accurately classifies individuals with diabetes as diabetic.

After completing the grid search and cross-validation steps, the results were compiled, and the top-performing models for both imbalanced and balanced subsets were identified. The models and their associated tuned hyper parameters are presented in Table 2. Ultimately, the trained model is utilized to make predictions, resulting to obtain the desired output corresponding to the patients' inputs.

The evaluation criterion of the model's performance on the test data involves the use of multiple metrics, including recall, specificity, accuracy, precision, f1-score, and area under the curve (AUC). Since this dataset is in a medical domain, recall followed by specificity was given priority. Recall (sensitivity) is crucial for identifying the true positive cases of diabetes, ensuring that individuals with the disease are correctly detected. Specificity, on the other hand, is important for minimizing false positives, which helps prevent unnecessary treatments or interventions for individuals who do not have diabetes. All the stages above are implemented using the Scikit-learn machine learning library.

Table 2: Tuned Hyperparameters for Different Models

Model	Tuned Hyperparameters
k-Nearest Neighbors	Number of neighbors
Na¨ıve Bayes	Smoothing parameter
Logistic Regression	Solver, C
Decision Tree	Criterion
Random Forest	Number of trees in the forest
Gradient Boosting	n_estimators, learning_rate, max _depth

Results

In this section, the aforementioned methodologies are applied to the dataset described earlier. The models were initially executed on a subset of the imbalanced dataset and then on a balanced subset using SMOTE. This approach helps to assess which model performs best predicting diabetes according to each metric.

Table 3: Performance Metrics of Different Models on imbalance data

Model	Accuracy	Precision	F1-Score	AUC	Recall	Specificity
k-Nearest Neighbors	0.95	0.61	0.09	0.52	0.05	1.00
Na¨ıve Bayes	0.94	0.29	0.20	0.57	0.15	0.98
Logistic Regression	0.96	1.00	0.18	0.55	0.10	1.00
Decision Tree	0.95	0.50	0.42	0.67	0.37	0.98
Random Forest	0.96	0.96	0.42	0.63	0.27	1.00
Gradient Boosting	0.96	0.82	0.43	0.65	0.29	1.00

A quick glance at the performance of different models on the subset of the imbalanced medical dataset indicates a high accuracy for all the models. However, accuracy can be misleading due to the dataset's uneven distribution between classes. Models achieving a high accuracy may be biased towards predicting the majority class (class 0), which can lead to an overlook of the positive diabetes cases or the minority class (class 1). Since it is crucial for us to prioritize recall, a closer look at this metric reflects the poor performance of all the models in correctly identifying the positive cases. This deficiency in recall to correctly identify people with diabetes is a critical factor for our medical dataset.

After analysing the models, Logistic Regression struggles with recall, while having a high precision. This means that it is potentially missing some positive cases. Random Forest and Gradient Boosting compared to other models show a balanced performance across precision, recall, and specificity. K-Nearest Neighbors performs well in specificity but has the lowest recall, meaning it might miss positive cases. Decision Tree displays a moderate performance overall. Surprisingly, it performs better than ensemble models based on recall score. Na¨ıve Bayes generally underperforms in both precision and recall.

Based on the observation of metrics for these models, the Random Forest and Gradient Boosting indicate a robust and balanced prediction across different metrics. This is mainly due to their ability to strike a good balance between recall, and specificity (figure 3). It also highlights their good performance in including the minority class, which aligns with the understanding of ensemble models to handle imbalanced datasets better than base models like logistic regression.
Figure 3 compares the recall and specificity scores among Decision Tree, Random Forest, and Gradient Boosting before SMOTE. The ensemble models have a slightly higher specificity, whereas the Decision Tree demonstrates a higher recall.

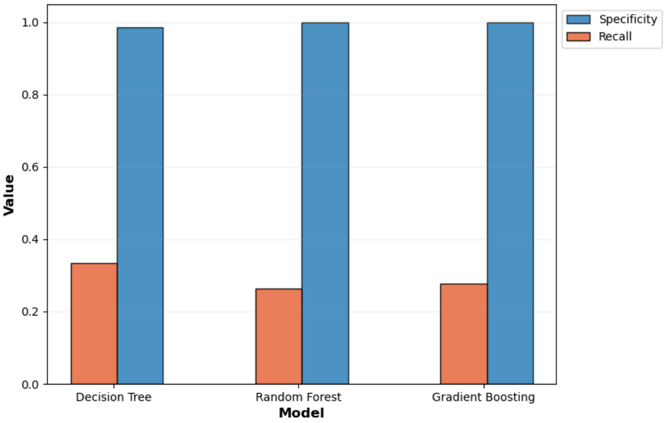

Figure 3: Bar Plot of Recall score - Comparison Before SMOTE

Additionally, to ensure effective model performance in identifying positive medical cases, while maintaining accuracy in predicting negative cases, SMOTE balancing technique was implemented to the dataset and recall was given priority to tune our hyperparameters. The results of this approach are compiled in Table 4.

The performance of all models evaluated on the balanced subset of data against the test data after recall-based tuning is illustrated in Table 4. An overall evaluation of the metrics in this table shows that Na¨ıve Bayes outperforms the other models in terms of recall (0.89). This suggests that Naive Bayes is more effective in recognizing the true positive cases. However, it has the lowest specificity across all the models, leading to a considerable amount of false positives.

Table 4: Performance Metrics of Different Models after SMOTE

Model	Accuracy	Precision	F1-Score	AUC	Recall	Specificity
k-Nearest Neighbors	0.91	0.30	0.40	0.76	0.82	0.93
Na¨ıve Bayes	0.69	0.12	0.20	0.74	0.89	0.69
Logistic Regression	0.91	0.30	0.42	0.80	0.69	0.92
Decision Tree	0.94	0.41	0.48	0.78	0.59	0.95
Random Forest	0.95	0.46	0.53	0.79	0.63	0.96
Gradient Boosting	0.96	0.66	0.60	0.76	0.53	0.99

Both Random Forest and Decision Tree show moderate recall scores and a high specificity. This increases the possibility of missing true positive cases, while making them effective in limiting false positives. Finally, the Gradient Boosting model falls short in terms of recall, although it has the best accuracy, precision, and specificity, making it the least effective in identifying true positive cases. Also, the Logistic Regression is not performing as well as other models in terms of recall and specificity.

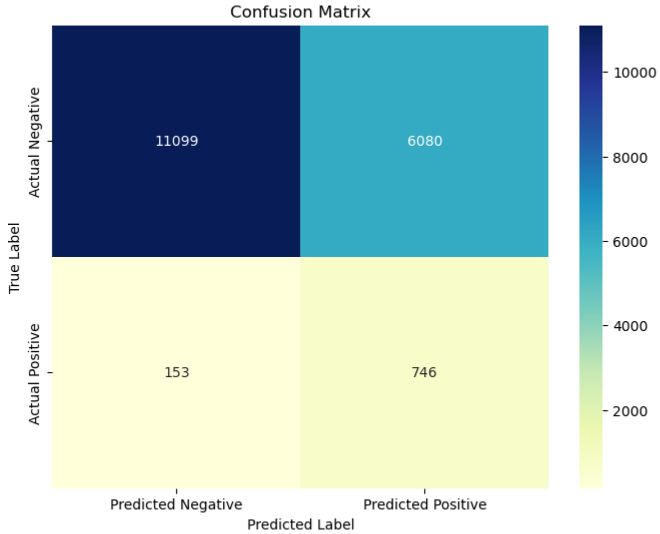

Figure 4: Naive Bayes Model confusion matrix - After balancing with SMOTE

Considering the importance of recall and specificity in medical cases, the best option with the highest recall in a balanced dataset appears to be the Na¨ıve Bayes model. The recall score in this model indicates that there are more true positives than other models, as it is demonstrated in the confusion matrix in figure 5. The K-Nearest Neighbors has a recall of 0.83 and a Specificity of 0.92, which could be an option if a balance between recall and specificity is sought.

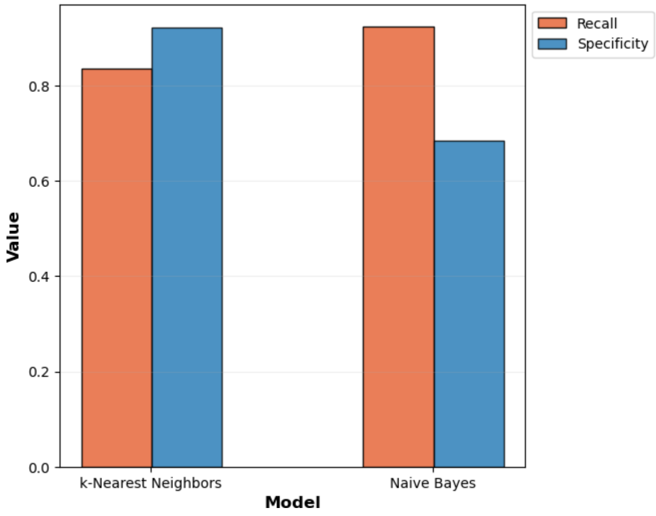

Figure 5: Bar Plot of Recall score comparison

Comparing the metrics in Table 3 (before SMOTE balancing) and Table 4 (post-SMOTE balancing) reveals some interesting patterns. Starting with the most important score for this study, recall, figure 6 illustrates that all models in Table 3 have a recall of less than 0.4 with the lowest recall of 0.05 for K-Nearest Neighbors. Table 4, on the other hand, presents a different result with improved recall scores for all the models with Na¨ıve Bayes having the highest recall of 0.89. Lastly, the majority of the models in both tables show a high specificity of 0.96 or above, indicating their effectiveness in predicting the negative class. The precision score range across the models of Table 3 is significantly greater than that of Table 4. The range for precision in table 3 starts from 0.29 in Naive Bayes being the lowest and ends with 1.00 being the highest in Logistic Regression. Table 4 indicates that Gradient Boosting has the maximum precision of 0.67, while Na¨ıve Bayes has the lowest precision of 0.11. This suggests a trade-off between recall and precision after balancing the data with SMOTE. When SMOTE is used, the models will have more exposure to the minority class (class 1), potentially resulting in more false positives from the majority class (class0), which lowers our precision score. Furthermore, the models become more sensitive to the minority class (class 1) which typically increases recall by reducing false negatives in predicting the minority class (class 1). In both metric tables, f1-score, the harmonic mean of precision and recall is comparatively lower for all the models. This suggests that recall and precision are not balanced. In addition to this, for every model in Table 3, the Area Under the Curve (AUC) is relatively low; nevertheless, Table 4 demonstrates an improvement. As illustrated in Table 3, the range of accuracy is consistently high, with a minimum value of 0.94. However, in Table 4, there is a wider range with Na¨ıve Bayes having the lowest accuracy of 0.69 and Gradient boosting having the highest accuracy of 0.96.

Figure 6 compares the recall scores among the trained models before and after SMOTE.

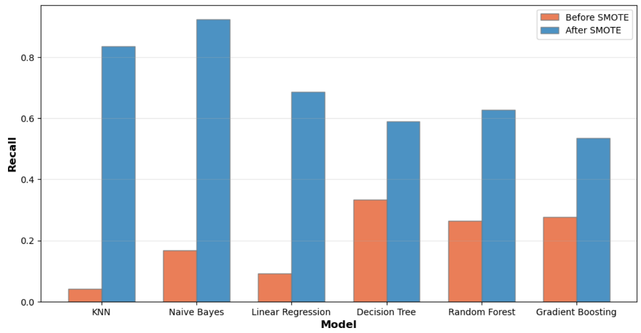

Figure 6: Bar Plot of Recall score - Comparison Before and After SMOTE

Conclusion

Our project successfully applied various classification algorithms like Logistic Regression, K-Nearest Neighbor (KNN), Naive Bayes, Decision Trees, Random forest and Gradient Boosting to predict diabetes in patients based on their medical history. Compared to all other algorithms Naive Bayes demonstrated the most promising performance in terms of recall. After Naive Bayes, KNN also performed well based on the overall balance of accuracy, recall and specificity. Our findings suggest that despite the strengths of ensemble methods like Random Forest, Naive Bayes provides a subtly better equilibrium between indicating true cases and avoiding false cases, both of which are critical for the early detection of medical conditions.

The challenges we faced included determining the most effective balancing method across all existing models, debating whether to employ SMOTE or weight balancing. We recognize the limitations of our research, and its dependence on a singular dataset that might not reflect the variety of patients with diabetes, and our focus on a selected number of classification algorithms. These restrictions pave the way for upcoming research to delve into deep learning methods and to use broader datasets, adding additional factors such as lifestyle and dietary habits. Our model will be more reliable and useful in real world situations if they undergo additional testing, which will be an important advancement towards early detection of diabetes.

References

Ataya, A. H. 2023. Early detection of Diabetes using Machine Learning Techniques. In *2023 Third International Conference on Artificial Intelligence and Smart Energy (ICAIS)*, 886–891. IEEE.

D'Souza, N.; Shah, K.; and Singh, P. 2022. Diabetes Detection Using Machine Learning Algorithms. In *2022 IEEE Bombay Section Signature Conference (IBSSC)*, 1–5.

Eid, S.; Sas, K. M.; Abcouwer, S. F.; Feldman, E. L.; Gardner, T. W.; Pennathur, S.; and Fort, P. E. 2019. New insights into the mechanisms of diabetic complications: role of lipids and lipid metabolism. *Diabetologia*, 62(9): 1539–1549. Epub 2019 Jul 25. PMID: 31346658; PMCID: PMC6679814.

Guan, Z.; Li, H.; Liu, R.; Cai, C.; Liu, Y.; Li, J.; Wang, X.; Huang, S.; Wu, L.; Liu, D.; Yu, S.; Wang, Z.; Shu, J.; Hou, X.; Yang, X.; Jia, W.; and Sheng, B. 2023. Artificial intelligence in diabetes management: Advancements, opportunities, and challenges. *Cell Rep Med*, 4(10): 101213. Epub 2023 Oct 2. PMID: 37788667; PMCID: PMC10591058.

Mustafa, M. n.d. Diabetese prediction Dataset. https://www.kaggle.com/datasets/iammustafatz/diabetes-prediction-dataset/data. Kaggle dataset.

Shampa, S. A.; Islam, M. S.; and Nesa, A. 2023. Machine Learning-based Diabetes Prediction: A Cross-Country Perspective. In *2023 International Conference on Next-Generation Computing, IoT and Machine Learning (NCIM)*, 1–6.

World Health Organization. 2023. Title of the Specific Page or Article. https://www.who.int/specificpage. Accessed: 2023-04-21.