Predictive Modeling of Diabetes

Solaris Serrano, Stephanie Reyes, Aryan Rathi, Saadi Sabyasachi, Min Nguyen, Nick Augustus, Mohammad Sadi
Department of Computer Science and Engineering
Virginia Commonwealth University

***Abstract*—Diabetes is one of the leading chronic illnesses worldwide, and one that often goes undetected until it starts to show troublesome symptoms. Late detection upends many lives in the United States, underscoring the importance of early detection and prevention in reducing both healthcare burdens and costs. Our project will apply machine learning and data science methodologies to the CDC Diabetes Health Indicators dataset to improve early detection and risk assessment. Since our initial proposal, we have refined our methodology and are designing preprocessing steps for handling class imbalance, scaling and encoding features. We are now in the process of implementation for our model. We are preparing to train our model and compare various learning algorithms for optimal coverage. Ongoing work focuses on data preparation, algorithm testing, and interface design for client use.**

I. PROBLEM STATEMENT

Diabetes is a major health issue that affects many in the United States and around the world, often without any early signs or symptoms. In 2019, the CDC reported that 37.3 million people (11.3% of the US population) had diabetes.[1] ("National and State Diabetes Trends | CDC") Currently, our methods of diagnosis rely heavily on clinical testing, which can be a barrier to entry for many people who may not be able to afford to constantly get checked for something they don't have any indication of having. This leaves many individuals undiagnosed until complications do arise, which can lead to lifestyle changes that could have been avoided with early detection. There is research that has demonstrated that machine learning classification models based on observable attributes can achieve optimistic performances in diabetes risk prediction.[2] We are in the process of testing various machine learning models and developing a tool for healthcare professionals as well as patients to help promote early detection and intervention in diabetes prone individuals.

II. INTRODUCTION

Diabetes is a leading health issue in the United States(U.S.), widely known for its impact on Americans. It is a disease that affects how the body regulates blood sugar. Specifically, type 2 diabetes has become more common in the U.S than type 1. It has negatively impacted much of the public's health. While it can be attributed to one's lifestyle habits, there is currently no cure for it. Additionally, diabetes can often go undetected and undiagnosed in many people for a period of time. Early detection of diabetes is critical as it can often lead to further health issues. Our project aims to predict the likelihood of diabetes based on one's lifestyle choices and health.

Data science has become extremely powerful in generating predictive models for different topics around the world. In this project, we will analyze data to identify the relationship between diabetes and various factors, including general health, body mass index(BMI), age, sex, physical activity, and others. We will utilize a Diabetes Health Indicators Dataset, derived from the Behavioral Risk Factor Surveillance System (BRFSS), to gather the information needed.Various models will be applied to analyze the data and identify meaningful patterns. Fine-tuning will then be used to evaluate the accuracy of our model, ensuring reliable detection of patterns related to diabetes risk. This process will help prioritize the most significant patterns to prevention and early detection.

The intention will be to spot commonalities and correlations between diabetes and lifestyle factors. While certain variables will have a greater effect on one's development of diabetes, others may have little to no effect. Using this analysis, we aim to develop an interactive application that will allow users to enter their health information and receive an estimated risk of diabetes. Through this project, we will aim to predict which variables have the most impact on the likelihood of developing diabetes with data analysis, emphasizing the importance of efforts to prevent its early stages.

III. LITERATURE SURVEY

A. Existing System

Diabetes prediction is an active area of research. Onset diabetes has been strongly linked to various factors like age, gender, body mass index (BMI), blood pressure, smoking habit, physical activity, diet etc. Early detection of diabetes can significantly assist proper treatment management, hence it is

extremely important to develop a model to predict diabetes. Various machine learning models have been explored on different data sets. Support vector machine, Bayesian network etc. have been used by different research groups. Fuzzy expert system frameworks for diabetes has been built with large scale knowledge based system by Kalpana and Kumar [A] using data from Pima Indians Diabetes Database (PIDD) of National Institute of Diabetes and Digestive and Kidney Disease (NIDDK). Fuzzy concept was used to transfer the information present into the required knowledge. In another approach, various features and machine learning classification models were studied with the same PIDD dataset [B]. Ten classification techniques such as CS-RT, C-RT, C4.5, LDA, K-NN, Naive Bayes, ID3, SVM, PLS-DA and RNDTREE were employed and their relative performance was analyzed to predict whether the patient is diabetic or not. C4.5 provided best accuracy. In 2020, comparison studies on random forest machine learning algorithm and Logistic Regression algorithm towards the prediction of diabetes was analyzed by Daghistani and Alshammari [C]. With the dataset from the Ministry of NationalGuard Health Affairs (MNGHA) hospital's database from three regions of Saudi Arabia, they achieved superior accuracy with RF algorithm at 88% [C, D].

B. Proposed System

Proposed Approach of Research

While there have been numerous studies with machine learning for diabetes prediction, the present work focuses on examining the problem from a more rigorous and comprehensive perspective. We seek to explore a diverse set of features to capture the multifactorial nature of diabetes and to quantify the uncertainty and reliability associated with model predictions.The study will include aspects of fairness, bias, and equity among different demographic and clinical subgroups to ensure robust and reliable performance. Finally, the study will explore adaptive modeling strategies to determine the most suitable predictive model for specific patient groups, thereby improving personalized prediction accuracy and clinical applicability.

IV. OUTLINE OF METHODOLOGY

A. Data Selection

The This project was driven by a key public health finding: the CDC's 2017 report showing a substantial decline in diabetes-related kidney failure between 2000 and 2014. To investigate the individual factors contributing to this trend, we used the 2015 Behavioral Risk Factor Surveillance System (BRFSS), specifically the diabetes_binary_health_indicators_BRFSS2015.csv dataset.

This dataset provided information on more than 250,000 individuals and included 21 health and demographic variables. These ranged from clinical indicators such as high blood pressure (HighBP), high cholesterol (HighChol), and BMI, to lifestyle-related factors like physical activity (PhysActivity) and smoking status (Smoker). Socioeconomic features such as income level (Income) and inability to visit a doctor due to cost (NoDocbcCost) were also incorporated, allowing us to capture external life circumstances that influence health outcomes.

The combination of clinical, behavioral, and socioeconomic variables made the BRFSS dataset well-suited for this analysis. Additionally, this dataset served as the backend data source for the prototype diabetes risk prediction and awareness application. The trained model allowed the app to estimate a user's diabetes risk score using personal health inputs such as age, BMI, physical activity, and cholesterol levels. Users were also able to view national diabetes trends filtered by state, economic bracket, or lifestyle factors through interactive dashboards built from anonymized BRFSS insights.

B. Data Preprocessing

- We implemented a full preprocessing pipeline to prepare the dataset for modeling. The data was first assessed for quality issues, including duplicates and missing values, and cleaned accordingly. All 21 features were retained because each variable demonstrated a plausible connection to diabetes risk. Special attention was given to categorical variables, including Age and Education, to ensure proper encoding and prevent the model from incorrectly interpreting categorical differences as numerical magnitudes.
- Once the data was cleaned, it was split into training (70%), validation (15%), and test (15%) sets to support model generalization rather than memorization. Numerical features such as BMI and mental health days (MentHlth) were standardized to ensure consistent scaling across all algorithms.
- We also addressed class imbalance—stemming from the larger proportion of non-diabetic individuals—using the SMOTE technique, which was applied only to the training set to prevent data leakage.
- For future app integration, the preprocessing pipeline was fully automated and included the following completed elements:
- Feature encoding and transformation automation: All user inputs in the app were passed through the same preprocessing steps used during model training to ensure consistency and accurate predictions.
- Real-time data validation: The app was programmed to flag unrealistic user inputs (e.g., BMI = 1000) and provide suggested valid ranges.
- Data privacy and security measures: All user data was processed locally or anonymized prior to model use, maintaining compliance with HIPAA-like

privacy standards.

C. Data Analysis

As we are using the BRFSS 2015 health indicator for the diabetes dataset. This dataset is used to predict and understand potential diabetes patients and their effecting factors. These datasets are divided into the training and validation sections.For analysis and model building, the training portion is being used to avoid leakage and the validating set remains untouched for final checks. We have standardized the numeric features and encode them to handle class imbalance. And using class weights for cross-validation. As we can see, there are fewer positive than negative samples, so these steps can help to keep the pipeline balanced.

At the beginning of the analysis, we try to detect how each feature can contribute to becoming a diabetes patient. How these features can differ between positive and negative cases. We analyze basic statistical tests, comparing the proportions and outcomes of positive and negative diabetes patients, and also check for false positives.

D. Feature Selection

We analyze these basic statistical tests, comparing their proportions and outcomes. Also check for false positives. By analyzing basic statistical tests and using feature selection, we filter three key variables by showing meaningful universal signals (using chi-squared/ANOVA information with FDR control). Secondly, to keep the coefficients consistent we use an L1-penalized logistic regression model . Lastly, we use a calibrated tree to confirm nonlinear values and identify any highly correlated pairs.

The dataset includes high blood pressure, cholesterol, general health, age, BMI, physical activity, and other socioeconomic markers to predict stability and redundancy. This makes the model reliable and helps reduce the number of diabetes patients by providing more accurate predictions based on education, income, age, and sex. By running an unbiased evaluation, we validate the feature set for training.

E. Model Selection and Training

To build an effective predictive framework, several supervised machine learning algorithms will be explored and compared for classifying individuals as healthy, pre-diabetic, or diabetic. The primary models under consideration include Logistic Regression, Decision Tree, Random Forest, Support Vector Machine (SVM), Gradient Boosting (XGBoost/LightGBM), and Artificial Neural Networks (ANN). The dataset, derived from the CDC's Diabetes Health Indicators (BRFSS 2015), will first be divided into training (70%), validation (15%), and testing (15%) sets to ensure generalization. The training data will be used to fit the models, while the validation set will be used for model selection and tuning. Each algorithm will be trained using standardized numerical features and one-hot encoded categorical features.

Class imbalance, common in diabetes datasets, will be handled using the SMOTE (Synthetic Minority Over-sampling Technique) applied only on the training set to prevent data leakage. During training, cross-validation will be used to ensure stability of performance and to mitigate overfitting. The selection criteria for the best-performing model will not only rely on accuracy but also emphasize recall and F1-score, as false negatives (undiagnosed diabetes) are more critical than false positives in this health context.

F. Model Evaluation

Model performance was evaluated on the multi-class BRFSS diabetes dataset with three outcome categories: healthy (0), pre-diabetic (1), and diabetic (2). The original distribution was highly imbalanced, with approximately 84% healthy, 14% diabetic, and only 2% pre-diabetic cases. To preserve this imbalance while assessing generalization, the dataset was split using stratified sampling into training (70%), validation (15%), and test (15%) sets.

Because the dataset is large, full k-fold cross-validation on all samples would be computationally expensive. Instead, we performed cross-validation on a stratified subsample of 10,000 training instances. A 3-fold StratifiedKFold scheme was used to estimate performance for candidate models (logistic regression, decision tree, Random Forest, SVM with RBF kernel, gradient boosting, and XGBoost). Among these, the Random Forest consistently achieved the best balance of weighted F1-score and macro recall, so it was selected as the primary model for deeper analysis. On the subsample, the Random Forest achieved roughly 0.83 accuracy, 0.82 weighted F1, and 0.43 macro F1, indicating strong overall performance but uneven treatment of the minority classes.

For the main evaluation, a Random Forest model with 200 trees and class-weighted loss (to partially offset class imbalance) was trained on the full training set and evaluated on the validation and test sets. On the test set, the model reached an overall accuracy of 0.84, weighted F1-score of 0.80, and macro F1-score of 0.38. Per-class metrics revealed substantial differences in performance. For the healthy class, precision and recall were both high (≈0.86 precision, ≈0.97 recall, F1 ≈0.91), showing that the model is very effective at recognizing non-diabetic individuals. For the diabetic class, performance was more modest, with precision ≈0.48, recall ≈0.16, and F1 ≈0.24, indicating that a substantial fraction of diabetic patients are misclassified as healthy. Performance on the pre-diabetic class was essentially zero (near-0 precision, recall, and F1 on both validation and test), reflecting the extreme scarcity of this label in the data and the model's tendency to absorb these cases into the majority classes.

Confusion matrices for the validation and test sets further

illustrated these patterns. Most healthy individuals were correctly assigned to the healthy category, while many diabetic and nearly all pre-diabetic samples were misclassified as healthy or, to a lesser extent, diabetic. To quantify model behavior beyond accuracy, we also computed specificity for each class in a one-vs-rest fashion. On the test set, specificity for the pre-diabetic and diabetic classes was very high (≈0.997 and ≈0.972, respectively), indicating that the model almost never predicts these labels unless it is very confident. In contrast, the specificity for the healthy class was relatively low (≈0.15), which is consistent with the high rate at which non-healthy individuals are incorrectly labeled as healthy when the roles are reversed in a one-vs-rest analysis.

Threshold-free metrics were evaluated using ROC and precision–recall curves. For the baseline Random Forest on the test set, the weighted multi-class ROC-AUC was approximately 0.79, and the weighted average precision (PR-AUC) was approximately 0.85. Class-specific ROC curves showed AUCs around 0.79–0.80 for the healthy and diabetic classes but considerably lower for the pre-diabetic class. Precision–recall curves highlighted the same pattern: the area under the PR curve was very high for healthy (AP ≈0.95), moderate for diabetic (AP ≈0.36), and extremely low for pre-diabetic (AP ≈0.03). These diagnostic plots confirm that the model is highly discriminative for the majority class and reasonably informative for diabetics, but struggles to reliably separate pre-diabetic cases from the background.

Overall, the evaluation indicates that the Random Forest provides strong aggregate performance and good discrimination between healthy and diabetic individuals, but its ability to detect the rare pre-diabetic class is extremely limited. These limitations, driven primarily by severe class imbalance, motivated the hyperparameter optimization and additional tuning procedures described in the subsequent fine-tuning section.

G. Model Fine-Tuning

After the initial comparison across classifiers, Random Forest emerged as the most promising model in terms of weighted F1-score and macro-averaged recall. However, the baseline configuration still showed clear weaknesses, particularly a very low recall for diabetic and pre-diabetic patients. To address this, we carried out a two-stage hyperparameter optimization process focused on Random Forest, with the objective of improving sensitivity to high-risk classes while preserving good overall performance.

Because the dataset is large, all tuning was performed on a stratified subsample of 8,000 instances drawn from the training set. This subsample preserved the original class proportions and made it feasible to run repeated cross-validation. For each candidate model configuration, we used 3-fold stratified cross-validation and optimized the weighted F1-score, a metric that both accounts for class imbalance and emphasizes a good balance between precision and recall across all classes.

In the first stage, we applied RandomizedSearchCV to explore the Random Forest hyperparameter space. The random search varied the number of trees, maximum tree depth, minimum number of samples required to split a node, minimum samples per leaf, and the feature sampling strategy at each split. All models were trained with class-weighted loss (class_weight="balanced") and the same preprocessing pipeline used in the baseline model. Ten randomly sampled hyperparameter combinations were evaluated, resulting in 30 training–validation fits on the 8,000-sample tuning set. The best configuration discovered by this procedure achieved a cross-validated weighted F1-score of approximately 0.82, with relatively deep trees and a moderate number of estimators. This result already matched or slightly exceeded the baseline Random Forest's performance, confirming that careful tuning of tree depth and complexity can improve the bias–variance trade-off.

In the second stage, we used Bayesian optimization via BayesSearchCV to refine the hyperparameters around the promising region identified by random search. The Bayesian search used the same tuning subset and cross-validation scheme but relied on a probabilistic model to choose new hyperparameter settings based on past evaluations, rather than sampling them uniformly at random. The search space again included the number of trees, maximum depth, minimum samples for splits and leaves, and the choice of feature subset strategy. Over ten Bayesian optimization iterations, the best model achieved a cross-validated weighted F1-score of about 0.82 on the tuning subset, comparable to the random search result but with a slightly different combination of hyperparameters. The selected configuration used 340 trees, a maximum depth of 20, a minimum of four samples required for a split, two samples per leaf, and a log2 feature sampling strategy. Given its strong and stable cross-validated performance, this BayesSearch-tuned Random Forest was selected as the final model.

To obtain the final performance estimates, the tuned Random Forest, together with the preprocessing pipeline, was retrained from scratch on the combined training and validation sets (a total of 215,628 instances) and then evaluated on the held-out test set. On this test set, the tuned model achieved an accuracy of 0.80, a weighted F1-score of 0.80, and a macro F1-score of 0.44. Compared with the baseline Random Forest, which had a macro F1 of approximately 0.38, this represents a noticeable improvement in average performance across classes. The most important gains were observed in the diabetic category: recall increased from roughly 0.16 in the baseline model to about 0.54 in the tuned model, and the diabetic F1-score rose from around 0.24 to 0.45. Performance for healthy individuals

remained strong (precision 0.91, recall 0.86, F1 0.88), while the pre-diabetic class unfortunately continued to exhibit near-zero recall and F1, reflecting the severe scarcity of this label in the data.

Threshold-independent metrics also improved slightly after fine-tuning. The weighted multi-class ROC-AUC increased from roughly 0.79 for the baseline model to about 0.80 for the tuned model, and the weighted average precision (PR-AUC) rose from approximately 0.85 to 0.86. Class-specific ROC curves showed AUC values of about 0.81 for healthy, 0.63 for pre-diabetic, and 0.82 for diabetic. Precision–recall curves confirmed very strong discriminative ability for healthy cases (area under the PR curve close to 0.96) and better ranking performance for diabetic cases (area under the PR curve around 0.41), while the pre-diabetic class remained extremely challenging (area under the PR curve near 0.03).

Overall, the fine-tuning procedure led to a Random Forest model that is more clinically useful than the baseline. By adjusting its depth, size, and node-splitting thresholds, the tuned model substantially increased sensitivity to diabetic cases and improved macro-averaged performance, at the cost of only a modest reduction in overall accuracy. At the same time, the results highlight a key limitation of the dataset: the extremely small pre-diabetic class makes it very difficult for any model to reliably detect this group. This suggests that further improvements are likely to require targeted data collection or re-design of the prediction task (for example, combining pre-diabetic and diabetic into a single “at-risk” category) in addition to algorithmic fine-tuning.

H. User Interface Development

We will create an R Shiny app to incorporate an interactive user interface that can offer an effective way to integrate user input and statistical models. The purpose of the app is to bridge the gap between the predictive model and users who require it. This allows the predictive system to be accessible to patients, healthcare professionals, and public health researchers.

Through the app, users can input relevant health indicators through form fields (e.g., BMI, blood pressure, physical activity, smoking status). Once submitted, the model will process the input and output the user’s classification as healthy, pre-diabetic, or diabetic.The application will also provide a probability score and highlight the most problematic health indicators, offering interpretability and transparency.

The app can also allow for interactive data visualizations to enhance usability. For example, users could simulate scenarios by adjusting variables such as increasing physical activity to observe how these inputs would change the predicted risk.. By combining predictive analytics with a user-friendly interface, the app will not only support early detection of diabetes but also provide insightful information for prevention and intervention for all users.

V. CONCLUSION

This project highlights the ever-growing importance of both machine learning and data science in the early detection of diabetes, a chronic condition that can progress without many detectable warning signs. By using the CDC Diabetes Health Indicators dataset, we are able to learn how health, demographic, and lifestyle factors can contribute to an individual’s risk of developing diabetes. Our system’s objective is not only to classify individuals as healthy, pre-diabetic, or diabetic but also to identify the most influential variables driving these outcomes.

REFERENCES

[1] M. Kalpana, A. Senthilkumar, Fuzzy expert system for diabetes using fuzzyverdict mechanism, Int. J. Adv. Netw. Appl. 3 (2011) 1128–1134.

[2] JK. Rajesh, V. Sangeetha, Application of data mining methods and techniques for diabetes diagnosis, Int. J. Eng. Innov. Technol. 2 (2012).

[3] T. Daghistani, R. Alshammari, Comparison of statistical logistic regression andrandom forest machine learning techniques in predicting diabetes, J. Adv. Inf.Technol. 11 (2020).

[4] V. Jaiswal, A. Negi, T. Pal, A review on current advances in machine learning based diabetes prediction, J. King Saud Univ. – Comput. Inf. Sci. 35 (2023) 101734.

[5] Burrows NR, Hora I, Geiss LS, Gregg EW, Albright A. Incidence of End-Stage Renal Disease Attributed to Diabetes Among Persons with Diagnosed Diabetes — United States and Puerto Rico, 2000–2014. MMWR Morb Mortal Wkly Rep 2017;66:1165–1170.DOI:http://dx.doi.org/10.15585/mmwr.mm6643a2.

[6] “National and State Diabetes Trends | CDC.” *CDC Archive*, 17 May2022, https://archive.cdc.gov/www_cdc_gov/diabetes/library/reports/reportcard/national-state-diabetes-trends.html? Accessed 1 October 2025.

[7] Rahman, Md Mahbubur, et al. "Diabetes prediction: a comparative study of machine learning techniques." *Healthcare* 11.6(2023):912. https://pmc.ncbi.nlm.nih.gov/articles/PMC10041290/