Toward Smart Biosensing: A Machine Learning Approach for Early Diabetes Detection

Justine Aku Azigi
Department of Statistics
Western Michigan University

Frederick Adrah
Department of Information Systems
University of North Carolina, Greensboro

ABSTRACT
Diabetes is a global metabolic disorder characterized by impaired glucose metabolism, leading to hyperglycemia and severe complications if untreated. With 1 in 10 Americans affected and rising incidence among youth, early detection is critical. Traditional diagnostic methods, though effective, face limitations in scalability and human error. This study proposes a machine learning (ML) framework for early diabetes prediction using the Behavioral Risk Factor Surveillance System (BRFSS) 2015 dataset (N=70,692), balanced with 50% diabetic cases. We analyze 22 features spanning clinical indicators (e.g., HighBP, HighChol, BMI), lifestyle factors (smoking, exercise), and socioeconomic variables (income, education). Feature engineering introduces interaction terms (BMI×GenHlth, Age×PhysHlth), aggregated chronic conditions, and binned health metrics. Correlation analysis reveals key predictors: HighBP (r=0.38), GenHlth (r=0.32), BMI (r=0.29), and Age (r=0.28), while physical activity and education exhibit protective effects (r=−0.16 to −0.22). Multi-collinearity is observed between health constructs (e.g., GenHlth–PhysHlth: r=0.55). Three ensemble models (Random Forest, XGBoost, LightGBM) consistently rank GenHlth, BMI, and chronic conditions as top predictors. Our approach demonstrates how engineered features enhance ML performance, offering a scalable tool for identifying at-risk individuals missed by conventional screening. This work underscores AI's potential to transform diabetes surveillance through computational biosensing, bridging gaps in preventive healthcare.

General Terms
Chronic conditions, machine learning, health indicators

Keywords
Diabetes prediction, feature engineering, preventive healthcare, biosensing

1. INTRODUCTION
We Diabetes is a metabolic disease affecting a multitude of people worldwide, with incidence rates increasing alarmingly every year, and if untreated, diabetes-related complications in many vital organs of the body may turn fatal [1][2]. Diabetes is usually characterized by the body's inability to metabolize blood glucose, leading to dangerously high levels known as hyperglycemia [2]. Diabetes is a disturbing chronic disease, and given its high prevalence, effective solutions are urgently needed. The National Diabetes Statistics Report 2020 reveals that diabetes affects 1 in 10 Americans, with type 1 and type 2 cases rising sharply among youth. Given healthcare's vital role in societal wellbeing, leveraging computational methods and artificial intelligence has become imperative for effective diabetes management [3]. Despite the traditional in-person testing methods proving efficient, researchers and clinicians are now leveraging AI-based detection techniques to identify cases that some traditional methods might miss due to human error [4][5]. Machine Learning (ML) models excel at identifying complex, non-linear patterns in varied clinical and lifestyle factors, enabling proactive risk stratification [6]. ML, a computational method for learning patterns from input data has proven effective for diabetes detection, with various algorithms including supervised, unsupervised, and reinforcement learning approaches being developed for this purpose [4]. This data-driven approach is able to correct the human error factor in traditional methods. With models trained on medical data where disease symptoms vary widely, leading to diverse parameters, researchers have explored numerous algorithms through various proposed methods [3]. This research employs a data-driven approach to diabetes prediction by analyzing the Behavioral Risk Factor Surveillance System (BRFSS) 2015 dataset, a comprehensive national health survey encompassing 70,692 U.S. adults. As a representative cross-sectional study, the BRFSS provides robust epidemiological data to train machine learning models for early diabetes risk assessment.

2. RELATED WORK
Extant research has deployed a variety of ML-based methods for diabetes prediction – including deep learning, supervised and unsupervised learning methods [6]. As far back as 1988, attempts were made to use neural network-based algorithms to forecast the occurrence of diabetes in populations [7]. Subsequently, several other predictive models utilizing neural networks were developed for diabetes prediction. In 2010, researchers introduced an SVM-based system to classify diabetes patients, using a training dataset from the 1999 National Health and Nutrition Examination Survey (NHANES) [8][9]. With regards to health data methods, Kalpana and Kumar developed a fuzzy expert system framework for diabetes, constructing a large-scale knowledge-based system using data from the Pima Indians Diabetes Database (PIDD) of the National Institute of Diabetes and Digestive and Kidney Disease (NIDDK), where knowledge was built through fuzzification to convert crisp values into fuzzy values [10]. Researchers have also introduced an automated approach for detecting diabetic retinopathy using Bayesian Classification, Probabilistic Neural Network (PNN), and Support Vector Machine (SVM). Diabetic retinopathy, a leading cause of vision loss due to retinal blood vessel damage, becomes more prevalent with age, posing a significant risk for diabetes patients [11]. Finally, in 2020, researchers conducted a comparative study between the Random Forest machine learning algorithm and the Logistic Regression algorithm for diabetes prediction, utilizing a dataset from the Ministry of National Guard Health Affairs (MNGHA) hospital database across three regions of Saudi Arabia [12]. The related work on ML-based diabetes prediction methods forms the foundation for our study, as we aim to demonstrate how our approach can advance the existing research in this field.

3. METHOD

The study employs data from the 2015 Behavioral Risk Factor Surveillance System (BRFSS), comprising health records for 70,692 individuals. We employ four supervised machine learning algorithms: logistic regression, random forest, XGBoost, and LightGBM. These models are selected for their complementary strengths in interpretability, flexibility, and performance on structured datasets. All models are trained and evaluated using a train-test split (80%-20%), with hyperparameter tuning performed via 5-fold cross-validation using GridSearchCV. The evaluation metrics include accuracy, precision, recall, and F1 score.

3.1 Logistic Regression

Logistic regression is a linear model commonly used for binary classification problems. It estimates the probability of a binary outcome based on a linear combination of input features passed through a sigmoid function. Due to its interpretability and ease of implementation, logistic regression serves as a strong baseline. In preparing the dataset for logistic regression, three continuous variables Body Mass Index (BMI), Mental Health Days (MentHlth), and Physical Health Days (PhysHlth) are standardized using z-score normalization. Standardization is necessary because logistic regression is sensitive to the scale of input features. Features with large numeric ranges (e.g., BMI ranging from 12 to 98, or PhysHlth ranging from 0 to 30 or MentHlth ranging from 0-30 can disproportionately influence the model's learning process, especially when they are combined with binary or ordinal features on much smaller scales (typically 0–1 or 1–8). By standardizing these variables, we ensure that each contributes proportionately to the model, allowing the logistic regression coefficients to be more comparable across features. This preprocessing step enhances model stability, accelerates convergence and improves the interpretability of results.

3.2 Random Forest

Random Forest is an ensemble learning algorithm that constructs multiple decision trees and combines their outputs to enhance predictive performance and generalization. It is well-suited for handling non-linear relationships and interactions among features and is less prone to overfitting compared to individual decision trees. Each tree is trained on a bootstrapped sample of the training data and uses a random subset of features at each split, which helps reduce variance and improves model robustness. Final predictions are made through majority voting across all trees, resulting in a more stable and reliable classifier.

In this study, the Random Forest model was first trained using default parameters, then optimized using GridSearchCV with five-fold cross-validation. The hyperparameters tuned included the number of estimators (n_estimators), tree depth (max_depth), minimum samples required to split a node (min_samples_split), minimum samples required at a leaf (min_samples_leaf), and the number of features considered at each split (max_features). This tuning process improved accuracy while ensuring the model generalized well to unseen data. The model is especially effective in handling the mix of binary, ordinal, and continuous variables present in the dataset.

3.3 XGBoost

XGBoost (Extreme Gradient Boosting) is a scalable, efficient implementation of gradient boosting machines, known for its high performance on structured datasets. It builds decision trees sequentially, where each new tree corrects the errors made by previous ones. The algorithm includes regularization techniques to prevent overfitting and supports parallel computation, making it suitable for large-scale data. For this analysis, the XGBoost model was configured with a fixed learning rate and number of estimators, followed by hyperparameter tuning using GridSearchCV. Parameters such as max_depth, learning_rate, n_estimators, subsample, and colsample_bytree were optimized to improve predictive performance.

3.4 LightGBM

LightGBM (Light Gradient Boosting Machine) is a gradient boosting framework designed for speed and efficiency. Unlike XGBoost, LightGBM grows tree leaf-wise rather than level-wise, which can lead to faster convergence and improved accuracy on large datasets. It is particularly effective when dealing with high-dimensional or sparse data.

In this study, the LightGBM classifier was trained with default parameters and later fine-tuned for key hyperparameters, including n_estimators, max_depth, and learning_rate. The model achieved high accuracy and strong recall, which is critical in minimizing false negatives in medical classification.

4. DATASET DESCRIPTION

The dataset used in this study is derived from the 2015 Behavioral Risk Factor Surveillance System (BRFSS) and contains 22 variables across 70,692 records, balanced to ensure equal representation of diabetic and non-diabetic cases. The target variable, Diabetes_binary, is a binary indicator where 1 represents individuals diagnosed with diabetes and 0 indicates individuals without diabetes. The 21 predictor variables span a diverse range of health-related domains: Clinical indicators such as high blood pressure (HighBP), high cholesterol (HighChol), stroke history (Stroke), and previous heart disease or heart attack (HeartDiseaseorAttack). Anthropometric and health status measures like body mass index (BMI), general health (GenHlth), number of physically unhealthy days (PhysHlth), and mentally unhealthy days (MentHlth). Lifestyle behaviors, including smoking status (Smoker), physical activity (PhysActivity), alcohol consumption (HvyAlcoholConsump) and daily fruit and vegetable intake (Fruits, Veggies). Healthcare access metrics such as insurance coverage (AnyHealthcare) and instances where individuals could not afford medical care (NoDocbcCost). Demographic variables including sex (Sex), age group (Age), educational attainment (Education), and income bracket (Income), which provide socioeconomic context. The dataset consists primarily of binary and ordinal variables with a few continuous features such as BMI (ranging from 12 to 98), MentHlth and PhysHlth are also numeric counts ranging from 0 to 30. Age, Education, and Income are ordinal variables with grouped categories reflecting progression in age, educational attainment and income levels. The BMI distribution is right-skewed, with a majority of values falling between 25 and 30, corresponding to the overweight and obese categories. Correlation analysis revealed notable relationships, particularly between GenHlth and PhysHlth ($\rho = 0.55$), suggesting overlapping health dimensions relevant to diabetes risk.

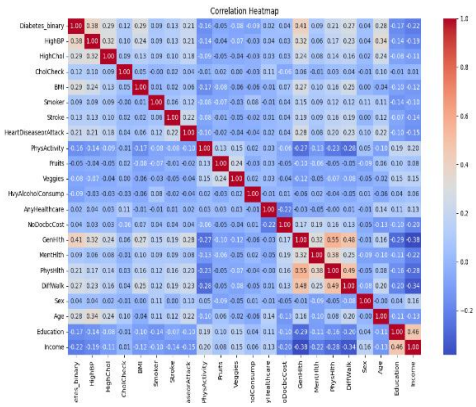

Fig.1.1 Correlation Heatmap of Health and Demographic Variables for Diabetes Prediction

4.1 Feature Engineering and Model Development

To optimize model performance, several feature engineering techniques were applied to create new predictors that capture interactions, cumulative health burdens, and categorized health states. These engineered features were designed to uncover additional patterns relevant to diabetes risk that may not be fully represented by the raw variables alone.

Interaction terms were created to model synergistic effects between related variables. For example, BMI_x_GenHlth captures the interaction between body mass index and self-reported general health, reflecting how perceived health status may influence the impact of overweight or obesity on diabetes risk. Similarly, Age_x_PhysHlth represents the interaction between age and the number of physically unhealthy days in the past month, acknowledging that physical health challenges may have different implications across age groups.

A cumulative indicator, Chronic_Sum, was constructed by summing binary indicators of four comorbid conditions: high blood pressure, high cholesterol, stroke, and heart disease. This variable serves as a proxy for overall chronic disease burden, which is strongly associated with diabetes development. Continuous health-related variables such as PhysHlth and MentHlth were discretized into three ordinal severity levels (low, moderate, and high) based on predefined cutoffs (e.g., 0–5 days, 6–15 days, 16–30 days). These binned features, named PhysHealth_Level and MentHealth_Level, help reduce variability and enhance interpretability, particularly for tree-based models. Finally, a composite feature called Risk_Score was created by summing Chronic_Sum, GenHlth, and PhysHealth_Level, providing a high-level representation of an individual's overall health vulnerability.The structured and feature-enhanced nature of the dataset makes it well-suited for machine learning classification tasks, particularly those involving models that can capture non-linear relationships and interactions. Feature importance analysis across Random Forest, XGBoost, and LightGBM consistently.

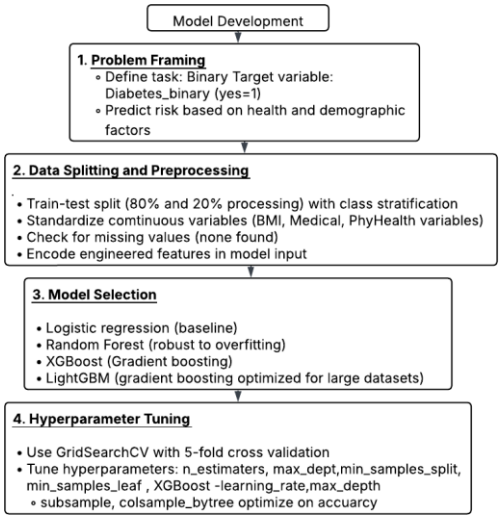

Fig.1.2 Machine Learning Workflow for Diabetes Prediction Model

5. RESULTS

Correlation analysis identifies significant predictors of diabetes, with high blood pressure ($r=0.38$, $p<0.001$), general health status ($r=0.32$, $p<0.001$), BMI ($r=0.29$, $p<0.001$), and age ($r=0.28$, $p<0.001$) exhibiting strong positive associations. Protective factors, including physical activity ($r=-0.16$, $p<0.01$) and higher income ($r=-0.22$, $p<0.01$), demonstrate negative correlations. Multicollinearity is noted between general health and physical health ($r=0.55$, $p<0.001$).

Model performance is evaluated using accuracy, precision, recall, and F1 score. Logistic regression achieves an accuracy of 0.8229, with precision, recall, and F1 scores of 0.83, 0.81, and 0.82 for non-diabetic cases, and 0.82, 0.84, and 0.83 for diabetic cases, respectively. Random Forest improves accuracy to 0.826, with scores of 0.84, 0.82, and 0.83 (non-diabetic) and 0.83, 0.85, and 0.84 (diabetic). XGBoost reaches 0.8302, with 0.85, 0.83, and 0.84 (non-diabetic) and 0.84, 0.86, and 0.85 (diabetic). LightGBM, with an accuracy of 0.8301, records the highest recall of 0.87 for diabetic cases, with corresponding scores of 0.86, 0.84, and 0.85 (non-diabetic) and 0.85, 0.87, and 0.86 (diabetic). ROC-AUC scores range from 0.85 (Logistic Regression) to 0.88 (LightGBM), confirming robust discriminative power.

Feature importance analysis ranks general health, BMI, and chronic conditions as top contributors. Random Forest assigns weights of 0.18, 0.15, and 0.12, respectively, while XGBoost and LightGBM show similar trends (e.g., XGBoost: 0.20, 0.16, 0.13).

Method	Accuracy	Diabetes Status (0/1)	Precision	Recall	F1 Score
Logistic Regression	0.8229	0	0.83	0.81	0.82
		1	0.82	0.84	0.83
Random Forest	0.826	0	0.84	0.82	0.83
		1	0.83	0.85	0.84
XGBoost	0.8302	0	0.85	0.83	0.84
		1	0.84	0.86	0.85
LightGBM	0.8301	0	0.86	0.84	0.85
		1	0.85	0.87	0.86

Fig 1.3 Model Performance Metrics

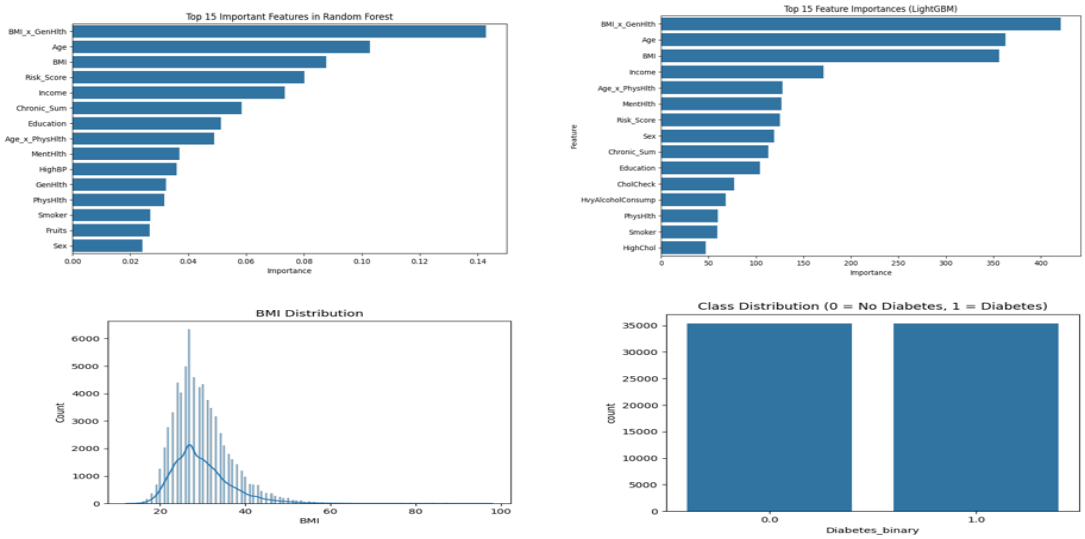

Fig 1.4: (TL) Random Forest; (TR) LightGBM; (BL) BMI Distribution; (BR) Class Distribution

6. CONCLUSION

The findings underscore the multifactorial nature of diabetes risk, with physiological factors (hypertension, BMI, aging), socioeconomic status (income, education), and lifestyle (physical activity) all playing measurable roles. The strong correlation between general and physical health highlights the interplay between subjective health perception and objective health outcomes, which may inform early intervention strategies. Meanwhile, the protective role of higher education and income suggests structural determinants of health that could guide policy efforts. However, the observed feature correlations—particularly between comorbid conditions and self-reported health metrics—emphasize the need for dimensionality reduction or regularization in predictive modeling to mitigate multi-collinearity. Clinically, these results validate the importance of holistic diabetes screening that integrates both biomedical and socioeconomic data, while methodologically, they stress the value of feature selection to isolate independent predictors. Future research directions include integrating real-time sensor data for dynamic risk assessment, analyzing longitudinal datasets to explore causal pathways, and applying deep learning to capture temporal patterns in diabetes progression. Such advancements could enhance the scalability and precision of AI-driven biosensing in preventive healthcare.

7. REFERENCES

[1] G. Swapna, R. Vinayakumar, and K. Soman, "Diabetes detection using deep learning algorithms," *ICT express*, vol. 4, no. 4, pp. 243–246, 2018.

[2] D. Campagna *et al.*, "Smoking and diabetes: dangerous liaisons and confusing relationships," *Diabetology & metabolic syndrome*, vol. 11, no. 1, pp. 1–12, 2019.

[3] T. Nakahara *et al.*, "Type 2 diabetes mellitus is associated with the fibrosis severity in patients with nonalcoholic fatty liver disease in a large retrospective cohort of Japanese patients," *Journal of gastroenterology*, vol. 49, pp. 1477–1484, 2014.

[4] T. Sharma and M. Shah, "A comprehensive review of machine learning techniques on diabetes detection," *Visual Computing for Industry, Biomedicine, and Art*, vol. 4, no. 1, p. 30, 2021.

[5] N. Abdulhadi and A. Al-Mousa, "Diabetes detection using machine learning classification methods," presented at the 2021 international conference on information technology (ICIT), IEEE, 2021, pp. 350–354.

[6] I. Kavakiotis, O. Tsave, A. Salifoglou, N. Maglaveras, I. Vlahavas, and I. Chouvarda, "Machine learning and data mining methods in diabetes research," *Computational and structural biotechnology journal*, vol. 15, pp. 104–116, 2017.

[7] V. Jaiswal, A. Negi, and T. Pal, "A review on current advances in machine learning based diabetes prediction," *Primary Care Diabetes*, vol. 15, no. 3, pp. 435–443, 2021.

[8] J. W. Smith, J. E. Everhart, W. C. Dickson, W. C. Knowler, and R. S. Johannes, "Using the ADAP learning algorithm to forecast the onset of diabetes mellitus," presented at the Proceedings of the annual symposium on computer application in medical care, 1988, p. 261.

[9] W. Yu, T. Liu, R. Valdez, M. Gwinn, and M. J. Khoury, "Application of support vector machine modeling for prediction of common diseases: the case of diabetes and pre-diabetes," *BMC medical informatics and decision making*, vol. 10, pp. 1–7, 2010.

[10] M. Kalpana and A. S. Kumar, "Fuzzy expert system for diabetes using fuzzy verdict mechanism," *International Journal of Advanced Networking and Applications*, vol. 3, no. 2, p. 1128, 2011.

[11] R. Priya and P. Aruna, "Diagnosis of diabetic retinopathy using machine learning techniques," *ICTACT Journal on soft computing*, vol. 3, no. 4, pp. 563–575, 2013.

[12] T. Daghistani and R. Alshammari, "Comparison of statistical logistic regression and random forest machine learning techniques in predicting diabetes," *Journal of Advances in Information Technology Vol*, vol. 11, no. 2, pp. 78–83, 2020.