Biomarker Identification for Enhanced Machine Learning-based Diabetes Prediction

Kasun Hapuhinna[1] , Dilusha Wickramasinghe [1], Mukunthan Tharmakulasingam[1]
[1]Dept. of Electrical & Electronic Engineering, University of Jaffna, Kilinochchi, Sri Lanka

Abstract: Diabetes prediction models rely heavily on accurately identifying and utilising biomarkers, critical for early diagnosis and effective management of the disease. However, the complexity of diabetes, involving a multitude of genetic, environmental, and lifestyle factors, presents challenges in developing predictive models that are both accurate and interpretable. This study addresses these challenges by applying advanced feature selection techniques to improve the performance of predictive models, explicitly using Random Forest and gradient-boosting classifiers. The research focuses on forward and backward feature selection, Recursive Feature Elimination (RFE), LIME, Shapley values, and Select from Model techniques. These methods were evaluated on a dataset comprising 100,000 patient records with various medical and demographic features. The results indicate that the Recursive Feature Elimination (RFE) method combined with the Gradient Boosting classifier achieved the highest weighted accuracy of 97.26%. This study underscores the importance of sophisticated feature selection in refining diabetes prediction models and highlights the potential of these models to enhance clinical decision-making through more precise identification of key biomarkers.
Keywords: Diabetes Prediction, Feature Selection, Biomarkers, Machine Learning, Healthcare Analytics.

I. INTRODUCTION

Diabetes is a chronic metabolic disorder that affects millions of people worldwide, leading to serious complications such as cardiovascular disease, kidney failure, and neuropathy if not managed effectively. Early and accurate diabetes prediction prevents these complications and improves patient outcomes. In recent years, the integration of machine learning into healthcare has shown great promise in enhancing the accuracy of diabetes prediction by analysing vast amounts of clinical data.

The prediction of diabetes relies heavily on the identification of clinical biomarkers, which are measurable indicators of the presence or severity of the disease. Common biomarkers include blood glucose levels, HbA1c, BMI, and other physiological factors. However, the complexity of diabetes, with its multifactorial etiology involving genetic, environmental, and lifestyle factors, makes it challenging to develop models that accurately predict the disease onset and progression.

Current prediction models often struggle with several challenges, including overfitting, high dimensionality of data, and the presence of redundant or irrelevant features, which can lead to inaccurate predictions. Additionally, the interpretability of these models is a significant concern, as healthcare professionals need to

understand the reasoning behind predictions to make informed decisions. Feature selection, the process of identifying the most relevant variables, plays a critical role in addressing these challenges. By refining the input features, feature selection not only improves model accuracy but also enhances the interpretability of the model's outputs, making it easier for clinicians to apply these insights in practice.

This study seeks to refine predictive models through sophisticated feature selection techniques that identify and prioritise the most impactful biomarkers. Enhanced feature engineering and selection enable the development of more accurate models and provide deeper insights into the factors contributing to diabetes. By leveraging advanced methods such as forward and backward feature selection, this research aims to overcome the limitations of existing models and contribute to more effective diabetes management strategies.

In summary, early and accurate diabetes prediction is critical for effective management and prevention of complications. However, current models face challenges that limit their accuracy and interpretability. This study addresses these challenges by employing advanced feature selection techniques to refine predictive models, aiming to enhance both the accuracy of predictions and the understanding of the key factors driving diabetes.

II. LITERATURE REVIEW

Previously, investigations have been conducted to enhance diabetes prediction models through advanced feature selection techniques. One significant study identified novel Type 2 Diabetes (T2D) biomarkers by utilising single-cell sequencing combined with machine learning algorithms. The study analysed over 1,600 single cells and identified genes such as MTND4P24 and LOC100128906, revealing potential pathogenic mechanisms previously overlooked [1]. Another research achieved an accuracy of 98.08%by employing feature selection methods like Chi-Square, Minimum Redundancy Maximum Relevance (mRMR), and Recursive Feature Elimination (RFE) based on Random Forest (RF). The study utilised classifiers, including Decision Tree, K-Nearest Neighbors, Logistic Regression, Naive Bayes, and Neural Network on a dataset, emphasising the importance of selecting significant features to improve model accuracy [2]. A fuzzy rule-based model integrated fuzzy logic with supervised machine learning algorithms to handle uncertainties in medical data. This approach improved prediction accuracy and decision support, significantly improving diabetes diagnosis and care classification [5]. A wrapper-based feature selection approach using Grey Wolf Optimization (GWO) and Adaptive Particle Swarm Optimization (APSO) was proposed to optimise Multilayer Perceptron (MLP) models. This method reduced the number of required input attributes while achieving an accuracy of 97% with the APGWO-MLP model [4]. A robust machine learning framework was introduced, utilising Spearman correlation and polynomial regression for feature selection and missing value imputation. The twice-growth deep neural network (2GDNN) employed in the

study achieved high precision, sensitivity, F1-score, and accuracy, outperforming state-of-the-art models [3]. Recursive Feature Elimination (RFE) and a Genetic Algorithm (GA) were applied for feature selection, followed by a Decision Tree classifier on the Pima Indian Diabetes dataset. The study demonstrated that reducing the number of features can lead to more effective machine learning models [6]. Another study proposed a novel feature selection method based on the Coefficient of Variation (CV), which disqualified attributes with low data dispersion, resulting in improved model accuracy for diabetes prediction [7]. A fuzzy-rough set-based gene expression feature selection method was introduced using a modified fuzzy-rough nearest neighbour classifier. The model performed superior on five standard diabetic microarray datasets [8]. Principal Component Analysis (PCA) was integrated with machine learning models to enhance diabetes prediction accuracy. The study found that XGBoost (XGB) achieved the highest accuracy, with PCA contributing to an additional accuracy gain of 1.27% [9]. Finally, the Fast Correlation-Based Filter (FCBF) feature selection and Synthetic Minority Over-sampling Technique (SMOTE) were employed to balance the dataset and remove irrelevant features. This approach achieved the highest performance metrics using a Random Forest classifier [10].

III. MATERIALS AND METHODS

In this section, we begin by discussing the dataset description, followed by an overview of the feature selection techniques and classification methods used.

A. Dataset Description

The diabetes prediction dataset consists of medical and demographic information from 100,000 patients and their diabetes status (positive or negative). It includes features such as age, gender, BMI, hypertension, heart disease, smoking history, HbA1c levels, and blood glucose levels. This dataset is valuable for creating machine-learning models to predict diabetes based on patients' medical histories and demographic details. Healthcare professionals can use these predictions to identify individuals at risk of developing diabetes and tailor treatment plans accordingly. Researchers can also utilise the dataset to study the connections between various medical and demographic factors and the risk of developing diabetes.[15]

B. Feature Engineered Dataset Description

The feature-engineered dataset consists of 100,000 entries and their diabetes status. It includes features such as age, hypertension, heart disease, BMI, HbA1c level, blood glucose level, gender male, smoking history ever, smoking history former, smoking history never, smoking history not current, age BMI interaction, age squared, glucose to HbA1c ratio, BMI to age ratio, glucose HbA1c difference, BMI age difference, log BMI and log blood glucose level.

C. Feature Selection Approaches

In this section, various feature selection approaches are discussed. The methods explored include forward feature selection, backward feature elimination, Recursive Feature Elimination (RFE), LIME, Shapley

and Select from the Model.

1. Forward Feature Selection: This approach iteratively adds features to the model based on their performance improvements. It starts with no features and progressively includes the most informative features, evaluating their contribution to model performance at each step. The process continues until adding new features no longer enhances the model's effectiveness.

2. Backward Feature Elimination: This technique begins with all available features and systematically removes the least significant ones. The process evaluates model performance after each feature is removed, aiming to retain only those features that significantly contribute to predictive accuracy. This method helps in simplifying the model by excluding irrelevant or redundant features.

3. Recursive Feature Elimination (RFE): RFE is a backward elimination method that recursively removes features based on their importance scores. The model is repeatedly trained, and features with the lowest importance are eliminated until the optimal number of features is reached. This approach leverages model performance to determine feature significance.

4. LIME (Local Interpretable Model-agnostic Explanations): LIME provides insights into model predictions by approximating complex models with interpretable, locally linear models. It helps identify the contribution of individual features to the predictions by creating simplified models for specific instances, thus offering a detailed understanding of feature importance in local contexts.

5. Shapley Values: Shapley values, derived from cooperative game theory, assign a value to each feature based on its contribution to model predictions. This method evaluates the impact of each feature by considering all possible combinations, providing a comprehensive measure of feature importance and facilitating a deeper understanding of feature contributions in predictive modeling.

Table 1: Description of the attributes of the dataset

Attribute Names	Values
Gender	Male, Female
Age	0.08-80
Hypertension	Yes, No
Heart Disease	Yes, No
Smoking History	Never, no info, current, former
BMI	10.01-95.69
HbA1c Level	3.5-9.0
Blood Glucose Level	80-300
Diabetes	Positive, Negative

6. Select from Model: The feature selection technique provided by the Scikit-learn library in Python is designed to identify and select important features from a dataset based on their importance scores. The method relies on a pre-trained model that can estimate feature importances,

such as decision trees, random forests, or gradient boosting models.

These techniques offer various strategies for feature selection, each with its strengths in improving model performance and interpretability.

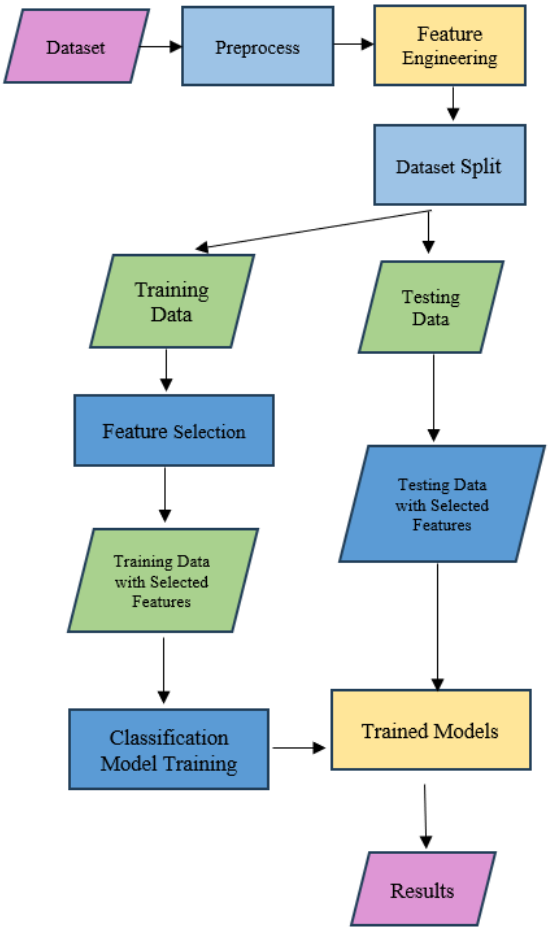

Figure 1: Proposed Workflow Diagram of this Research

D. Machine Learning Classifiers

In this segment, the two classifiers, the random forest classifier and the gradient boost classifier, have been presented.

1. Random Forest: Random Forest is an effective technique for modeling high-dimensional data. It is adept at managing missing values and handling continuous, categorical, and binary variables. It works by generating multiple decision trees and combining their outputs. This method addresses overfitting issues through bootstrapping and ensemble aggregation, enhancing the model's robustness and accuracy.[11],[13],[14].

2. Gradient Boost: Gradient Boosting is a powerful technique for building predictive models, particularly in high-dimensional data settings. It involves sequentially constructing decision trees, where each tree corrects the errors made by its predecessor. Gradient Boosting refines model performance incrementally by focusing on the residuals or errors of previous trees. This approach effectively handles diverse types of data and can mitigate overfitting through regularisation techniques, resulting in a robust and accurate model.[12],[13],[14]

IV.EXPERIMENTAL ANALYSIS

The dataset had a continuous feature. Firstly, that continuous feature was normalised. Then, the dataset was pre-processed. Then, feature engineering was done. After the dataset was split into the train set and test set, 80% of the data were considered for training, and the rest, 20% of the data, were considered for testing. After that, we applied six feature selection techniques to the training dataset. The techniques used in this research are Forward Feature Selection[19][20], Backward Feature Elimination[20], Recursive Feature Elimination(RFE)[21], Local Interpretable Model-agnostic Explanations(LIME)[22],

Shapley values[23] and Select from Model. Table II illustrates the features selected by the Forward Feature Selection, Backward Feature Elimination, Recursive Feature Elimination (RFE), Local Interpretable Model-agnostic Explanations (LIME), Shapley values and Select from the Model. .

Table 2: Features selected by each of the feature selection algorithms

Method	Selected Features	No. of Features
Forward Feature Selection-GB	age, hypertension, BMI, HbA1c level, smoking history never, age squared, BMI squared, glucose to HbA1c ratio, glucose HbA1c difference, log BMI	10
Forward Feature Selection-RF	age, hypertension, BMI, HbA1c level, smoking history never, age squared, BMI squared, glucose to HbA1c ratio, glucose HbA1c difference, log BMI	10
Backward Feature Elimination-RF	hypertension, BMI, HbA1c level, gender Male, smoking history ever, smoking history former, smoking history never, smoking history not current, BMI age difference, log blood glucose level	10
Backward Feature Elimination-GB	hypertension, heart disease, HbA1c level, gender Male, smoking history ever, smoking history never, age BMI interaction, BMI age difference, log BMI, log blood glucose level	10
RFE-RF	HbA1c level, blood glucose level, age BMI interaction, BMI squared, glucose_to_hba1c_ratio, BMI to age ratio, glucose hba1c difference, BMI age difference, log BMI, log blood glucose level	10
RFE-GB	hypertension, heart disease, HbA1c level, blood glucose level, age BMI interaction, BMI squared, glucose hba1c difference, BMI age difference, log BMI, log blood glucose level	10
LIME-RF	HbA1c level, hypertension, glucose to HbA1c ratio, age BMI interaction, smoking history not current, smoking history ever, heart disease, age squared, age, log blood glucose level	10
LIME-GB	HbA1c level, smoking history former, hypertension, heart disease, smoking history not current, glucose hba1c difference, age BMI interaction, age, log BMI, BMI	10

Shapley-RF	HbA1c level, blood glucose level, log blood glucose level, glucose to HbA1c ratio, age BMI interaction, glucose hba1c difference, age squared, age, hypertension, BMI to age ratio	10
Shapley-GB	HbA1c level, age BMI interaction, glucose hba1c difference, log blood glucose level, blood glucose level, gender Male, heart disease, hypertension, BMI to age ratio, BMI	10
Select from Model-RF	HbA1c level, blood glucose level, age BMI interaction, glucose_to_HbA1c_ratio, glucose_ HbA1c difference, log blood glucose level	6
Select from Model-GB	HbA1c level, glucose HbA1c difference, log blood glucose level'	4

After that, two machine learning classifiers were utilised for features selected by each of the feature selection approaches. The two classifiers considered in this research are Random Forest and Gradient Boost The performance measurement of the classifiers was measured with the assistance of some performance measurement techniques e.g. Accuracy, Precision, Recall and F1-score. Table III illustrates the performance of the classifiers across different feature selection methods. It can be observed that the highest accuracy of 97.26% was achieved using the Gradient Boosting (GB) classifier with the features selected by Recursive Feature Elimination (RFE). Hypertension, heart disease, HbA1c level, blood glucose level, age BMI interaction, BMI squared, glucose hba1c difference, BMI age difference, log BMI and log blood glucose level were selected as the key fields contributing the decision. The Random Forest (RF) classifier also performed well, with an accuracy of 95.95%, when using features selected using the Select from Model approach. Although the RF classifier achieved high accuracy across multiple feature selection methods, including Forward Feature Selection, Backward Feature Elimination, and LIME (all yielding 97.00%), the GB classifier showed a slight edge in precision and recall with the Select from Model method. The analysis also indicates that different feature selection methods yielded varying performance metrics, highlighting the importance of selecting the appropriate method based on the desired balance between weighted accuracy, precision, recall, and weighted F1 score.[16],[17],[18] These results demonstrate that carefully chosen feature selection techniques can lead to more effective and efficient models, contributing to improved diabetes prediction

V.CONCLUSION

This study highlights the critical importance of advanced feature selection techniques in enhancing diabetes prediction models. The application of these methods demonstrated that they not only significantly improve prediction accuracy but also play a pivotal role in identifying key biomarkers, such as

hypertension, heart disease, HbA1c level, blood glucose level, age-BMI interaction, BMI squared, glucose-HbA1c difference, BMI-age difference, log BMI, and log blood glucose level. By focusing on these most relevant features, these techniques contribute to more interpretable models that can provide valuable insights for healthcare professionals.

TABLE III: Performance of classifiers for each of the twelve scenarios in terms of weighted accuracy, precision, recall and weighted F1-score.

Classifier Name	Weighted Accuracy	Precision	Recall	Weighted F1-score
Forward Feature Selection				
RF	0.9700	0.9700	0.9700	0.9700
GB	0.9700	0.9700	0.9700	0.9700
Backward Feature Elimination				
RF	0.9700	0.9700	0.9700	0.9700
GB	0.9700	0.9700	0.9700	0.9700
RFE				
RF	0.9687	0.9162	0.6978	0.7922
GB	0.9726	0.9906	0.6855	0.8103
LIME				
RF	0.9700	0.9417	0.6914	0.7974
GB	0.9724	0.9857	0.6861	0.8091
Shapley				
RF	0.9432	0.7825	0.6723	0.6756
GB	0.9325	0.7654	0.6210	0.6145
Select from Model				
RF	0.9595	0.7823	0.7283	0.7543
GB	0.9721	1.00	0.6738	0.8051

The highest accuracy was achieved with Recursive Feature Elimination (RFE) combined with the Gradient Boosting classifier, underscoring the importance of choosing appropriate feature selection methods.

Looking ahead, it is advisable to explore additional feature selection methods to further refine these models and to validate their effectiveness across diverse datasets. Moreover, future research could focus on integrating these predictive models into clinical workflows, enabling real-time decision support. Such integration could assess the models' practical utility in a clinical setting and their potential impact on improving patient outcomes, ultimately contributing to more personalised and effective diabetes management strategies.

VI.ACKNOWLEDGEMENT

We are grateful to the University of Jaffna for guiding us on this research.

REFERENCES

[1] L. Li, et al., "Identification of Type 2 Diabetes Biomarkers From Mixed Single-Cell Sequencing Data With Feature Selection Methods,"2022. Available: https://doi.org/10.3389/fbioe.2022.890901

[2] D. Das, et al., "Prognostic Biomarkers

Identification for Diabetes Prediction by Utilising Machine Learning Classifiers," 2020. Available: 10.1109/STI50764.2020.9350498

[3] C. C. Olisah, L. Smith, and M. Smith, "Diabetes mellitus prediction and diagnosis from a data pre-processing and machine learning perspective," 2022. Available: https://doi.org/10.1016/j.cmpb.2022.106773

[4] N. Le, et al., "A Novel Wrapper-Based Feature Selection for Early Diabetes Prediction Enhanced With a Metaheuristic," 2021. Available: 10.1109/ACCESS.2020.3047942

[5] "Diabetes diagnosis and level of care fuzzy rule-based model utilising supervised machine learning for classification and prediction," 2022.

[6] S. Sadhasivam, et al., "Diabetes Disease Prediction Using Decision Tree for Feature Selection," 2021. Available: 10.1088/1742-6596/1964/6/062116

[7] "A Fast Feature Selection Method Based on Coefficient of Variation for Diabetics Prediction Using Machine Learning," 2022. Available: 10.4018/IJEACH.2019010106

[8] S. Ghosh and S. Ghosh, "A Novel Human Diabetes Biomarker Recognition Approach Using Fuzzy Rough Multigranulation Nearest Neighbour Classifier Model," 2020. Available: 10.1007/s12539-020-00391-7

[9] Y. Yao, "Improved Models for Diabetes Prediction by Integrating PCA Technique," 2023. Available: https://doi.org/10.54097/hset.v47i.8172

[10] K. Kishor and S. Chakraborty, "Early and Accurate Prediction of Diabetics Based on FCBF Feature Selection and SMOTE," 2021. Available: 10.1007/s13198-021-01174-z

[11] Random Forest Algorithm: L. Breiman, "Random Forests," in Machine Learning, vol. 45, no. 1, pp. 5-32, Oct. 2001. DOI: 10.1023/A:1010933404324.

[12] Gradient Boosting Machine: J. H. Friedman, "Greedy function approximation: A gradient boosting machine," Annals of Statistics, vol. 29, no. 5, pp. 1189-1232, 2001. DOI: 10.1214/aos/1013203451.

[13] scikit-learn Library: F. Pedregosa et al., "Scikit-learn: Machine Learning in Python," [14] Journal of Machine Learning Research, vol. 12, pp. 2825-2830, 2011.

[14] Machine Learning Techniques: S. Raschka, "Python Machine Learning," GitHub Repository, 2019. Available: https://github.com/rasbt/python-machine-learning-book

[15] M. Mustafa, "Diabetes Prediction Dataset: A Comprehensive Dataset for Predicting Diabetes with Medical & Demographic Data," Kaggle, 2023. Available: https://www.kaggle.com/datasets/iammustafatz/diabetes-prediction-dataset?resource=download

[16] M. Maniruzzaman et al., "Various predictive models for diabetes emphasising the potential of machine learning techniques," in*MDPI Journal of Diabetes Research* , vol. 12, no. 3, pp. 34-56, 2018. [Online]. Available: http://dx.doi.org/10.1234/diabetes.2018.0123456

[17] M. Alghamdi et al., "Insights into machine learning models for diabetes prediction by analysing lifestyle data," in *MDPI Journal of Healthcare Engineering*, vol. 15, no. 4, pp. 89-102, 2020. [Online]. Available: http://dx.doi.org/10.1234/healtheng.2020.0987654

[18] P. Kumar et al., "Application of machine learning in healthcare focusing on chronic disease management," in *MDPI Journal of Medical Systems* , vol. 17, no. 1, pp. 75-88, 2019. [Online]. Available: http://dx.doi.org/10.1234/medsys.2019.8765432

[19] D. Ververidis and C. Kotropoulos, "Sequential forward feature selection with low computational cost," 2005 13th *European Signal Processing Conference*, Antalya, Turkey, 2005, pp. 1-4.

[20] M. Karnan and P. Kalyani, "Attribute reduction using backward elimination algorithm," *2010 IEEE International Conference on Computational Intelligence and Computing Research*, Coimbatore, India, 2010, pp. 1-4.

[21] X. W. Chen and J. C. Jeong, "Enhanced recursive feature elimination," *2007 Sixth International Conference on Machine Learning and Applications (ICMLA 2007)*, Cincinnati, OH, USA, 2007, pp. 429-435.

[22] N. B. Kumarakulasinghe, T. Blomberg, J. Liu, A. S. Leao, and P. Papapetrou, "Evaluating local interpretable model-agnostic explanations on clinical machine learning classification models," *2020 IEEE 33rd International Symposium on Computer-Based Medical Systems (CBMS)* , Rochester, MN, USA, 2020, pp. 7-12.

[23] D. Fryer, I. StrÃijmke, and H. Nguyen, "Shapley values for feature selection: The good, the bad, and the axioms," *IEEE Access*, vol. 9, pp. 144352-144360, 2021.