Enhancing Diabetes Risk Screening for Preventive Healthcare through Statistically Validated Machine

Hanissah Mohamad @ Sulaiman[1*], Norazlina Abd Razak[1], Siti Huzaimah Husin [2], Siti Aisah Mat Junos @ Yunus[1] , Niza Mohd Idris[2], Zarina Baharudin Zamani[1]

[1]**School of Technical Foundation and Diploma Studies, Universiti Teknikal Malaysia Melaka, Hang Tuah Jaya, 76100 Durian Tunggal, Melaka, Malaysia**

[2]**Faculty of Electronic and Computer Technology and Engineering, Universiti Teknikal Malaysia Melaka, Hang Tuah Jaya, 76100 Durian Tunggal, Melaka, Malaysia**

DOI: https://doi.org/10.47772/IJRISS.2026.101000344

Received: 21 October 2026; Accepted: 26 October 2026; Published: 02 November 2026

ABSTRACT

Diabetes risk prediction using machine learning has gained increasing attention due to its potential to support early preventive healthcare screening. However, many existing models focus mainly on classification performance while giving limited attention to feature-level interpretability and training–validation consistency. This paper describes a statistically validated hybrid machine learning framework for early diabetes risk screening using a BRFSS-derived, non-invasive survey dataset. The reported analytical dataset contained 302,443 records and 22 variables, comprising the diabetes outcome and 21 candidate predictors. Statistical feature validation using Chi-Square Test, Independent Sample T-Test, Spearman Correlation and Mutual Information was followed by Random Forest-based Recursive Feature Elimination. The process retained five predictors: high blood pressure, high cholesterol, body mass index, general health and age. The Random Forest classifier achieved 10-fold cross-validation ROC-AUC of 0.8236 ± 0.0025 and F1-score of 0.6303 ± 0.0033. The observed training–validation generalization gap decreased from 5.9% in the unfiltered configuration to 1.2% in the statistically validated configuration. Nevertheless, the unfiltered model achieved higher absolute classification performance, indicating a trade-off between feature reduction, apparent overfitting and predictive discrimination. The statistical tests documented in the thesis were not sufficiently specified to establish a valid paired comparison between the two pipelines; therefore, no claim of inferential superiority is made. The framework provides a compact and transparent proof-of-concept for preliminary risk screening, not a clinically validated diagnostic tool. Independent external validation, calibration and subgroup evaluation remain necessary before practical adoption.

Index Terms- diabetes risk prediction, feature selection, interpretability, machine learning, model stability, Random Forest, statistical validation

INTRODUCTION

Diabetes mellitus is a chronic metabolic disorder associated with elevated blood glucose levels and serious long-term health complications if it is not detected and managed at an early stage [1], [2]. The increasing prevalence of diabetes has created a significant burden on healthcare systems due to long-term treatment, disease monitoring, and complications such as cardiovascular disease, kidney disease, neuropathy, and vision-related problems [1], [2]. Early identification of individuals at risk is therefore important to support preventive healthcare strategies, lifestyle modification, and timely medical screening [3], [4].

Machine learning has become an important computational approach for diabetes risk prediction because it can analyze large-scale health data and identify hidden relationships among demographic, lifestyle, and self-reported health indicators [3]-[6]. Compared with laboratory-based diagnostic procedures, survey-based prediction models are more scalable and less costly because they rely on non-invasive information that can be collected through questionnaire-based screening [10]-[12]. This makes machine learning-based risk

screening suitable for preliminary public health applications and community-level awareness systems, particularly when laboratory-based screening is limited.

Despite the growing use of machine learning in diabetes prediction, many existing studies remain strongly performance-driven. Models are commonly compared using metrics such as accuracy, precision, recall, F1-score, and ROC-AUC [3]-[6], [12]. Although these metrics are important, they do not fully address two critical requirements in healthcare-related machine learning systems: interpretability and model stability. Prior reviews have shown that machine learning models do not automatically provide better clinical prediction than simpler statistical models unless validation, transparency, and practical relevance are carefully considered [7], [8]. Therefore, a healthcare-oriented prediction framework should not only report predictive performance, but should also demonstrate that its selected features are meaningful and that its model behaviour is stable across data partitions.

A common limitation in existing diabetes prediction studies is the reliance on algorithmic feature selection alone. Techniques such as Recursive Feature Elimination, model-based feature importance, and wrapper-based selection can identify variables that improve prediction performance, but these methods do not necessarily verify whether the selected features have statistically meaningful relationships with the diabetes outcome [9], [10]. As a result, machine learning models may learn redundant, unstable, or dataset-specific patterns, especially when survey-based datasets contain correlated demographic, lifestyle, and health-related variables [9], [11]. To address this limitation, this paper examines a statistically validated feature-selection framework for early diabetes risk screening. Statistical validation is used as an evidence-based filter, while Recursive Feature Elimination further refines candidate predictors according to model-based importance. The study evaluates feature-level interpretability and training–validation consistency while recognizing that feature reduction may reduce absolute predictive discrimination.

The main contributions of this paper are as follows:

1. A two-stage feature selection framework is proposed by integrating statistical feature validation and Recursive Feature Elimination.

2. The framework supports feature-level interpretability by reducing 21 candidate predictors, from an analytical file containing 22 variables including the diabetes outcome, to five non-invasive predictors.

3. The framework reduces the observed training–validation generalization gap from 5.9% to 1.2%, while the associated decrease in absolute performance is acknowledged.

4. Cross-validation and learning-curve analysis are used to describe model variability; the scope and limitations of the statistical tests documented in the source thesis are explicitly clarified.

5. A Streamlit-based graphical user interface is developed as a proof-of-concept for non-invasive diabetes risk screening.

The remainder of this paper is organized as follows. Section II describes the research method, including dataset preparation, statistical validation, feature selection, model development, and stability evaluation. Section III presents the results and discussion. Section IV concludes the paper and highlights future research directions.

RESEARCH METHOD

This study used a structured supervised machine learning approach to develop a diabetes risk prediction framework using non-invasive survey-based health data. The workflow consisted of dataset acquisition, data preprocessing, statistical feature validation, Recursive Feature Elimination, machine learning model development, model stability evaluation, and graphical user interface implementation.

Dataset Description

The dataset was obtained from a publicly available BRFSS-derived survey extract. The processed analytical file was reported to contain 302,443 observations and 22 variables, comprising the binary diabetes outcome and 21 candidate predictors. The present analyses refer to this processed file; no independent external cohort was used. BRFSS-based feature screening and questionnaire optimization have also been studied in diabetes prediction [10].

The target variable was diabetes status, formulated as a binary classification problem. The input features included high blood pressure, high cholesterol, cholesterol check history, body mass index, smoking status, stroke history, heart disease or heart attack history, physical activity, fruit consumption, vegetable consumption, heavy alcohol consumption, healthcare access, inability to visit a doctor due to cost, general health, mental health, physical health, difficulty walking, gender, age, education, and income. The use of non-invasive survey data is suitable for early diabetes risk screening because it does not require laboratory-based testing. However, survey-based datasets may contain redundant, weakly informative, or correlated variables. Therefore, systematic feature validation was required before model training to ensure that the selected predictors were statistically meaningful and practically useful.

Data Preprocessing

Preprocessing involved inspection of duplicate records, missing or out-of-range values, and feature types. The thesis describes use of the pandas .duplicated() and .drop_duplicates() operations (keep='first') to identify and remove exact duplicate rows. Binary features were checked for valid 0/1 coding and invalid entries were assigned the corresponding modal category. The BMI variable was checked against an operational range of 10–100; values outside this range were treated as missing and replaced with the mean of valid BMI values. Ordinal and count-based features were checked against their documented allowable ranges and treated using mode- or mean-based imputation according to feature type. Treatment of missing values and outliers is an established methodological concern in diabetes risk prediction [3].

The 22-variable analytical representation comprised 21 candidate predictors and the diabetes outcome. Binary variables were treated as categorical; BMI as continuous; and general health, age, education and income as ordinal. A stratified 80:20 training-to-hold-out test partition was used to preserve the relative frequency of diabetic and non-diabetic cases. As documented in the preprocessing procedure, the analytical data were checked for duplicate rows and invalid ranges, with type-specific imputation for invalid or missing entries. Model development and internal validation were performed on the processed survey-based dataset.

Statistical Feature Validation

Statistical feature validation was introduced as the first feature screening stage. The purpose of this process was to identify features that demonstrated meaningful relationships with diabetes status before applying machine learning-based feature selection. Feature selection has been reported as an important step in diabetes prediction because it helps reduce dimensionality, remove redundant variables, and improve the reliability of predictive modelling [9], [10].

Four complementary methods were described. Chi-Square tests assessed categorical or binary predictors, and independent-sample t-tests assessed continuous predictors; the stated statistical significance criterion was $p < 0.05$. Spearman correlation was used for ordinal predictors, with the stated criteria $p < 0.05$ and $|\rho| > 0.10$. Mutual Information was used as a complementary, non-parametric ranking measure and was not treated as a p-value test. The thesis reports no fixed numerical Mutual Information cutoff or fully specified aggregation rule across the four methods. Filter- and wrapper-based feature selection strategies have previously been investigated in diabetes prediction research [9], [10]. The screening-stage subset consisted of eight candidate predictors: HighBP, HighChol, BMI, HeartDiseaseorAttack, PhysActivity, GenHlth, DiffWalk and Age. Mutual Information provided a complementary ranking of candidate features, rather than a significance-test p-value. These eight predictors were passed to RFE for further refinement.

Recursive Feature Elimination

Following statistical screening, Recursive Feature Elimination was applied with a Random Forest estimator. In the recorded experiment, the eight candidate features were reduced to five predictors: HighBP, HighChol, BMI, GenHlth and Age. HeartDiseaseorAttack, PhysActivity and DiffWalk were removed during RFE. Related diabetes-prediction research has also examined wrapper-based approaches for feature selection [9].

This two-stage selection separates univariate/statistical association from model-based feature importance. It provides a more concise set of model inputs, but it is not equivalent to a post-hoc explanation of individual predictions. The recorded reduction in feature count should therefore be interpreted as feature-level transparency rather than proof of full model explainability.

The final feature subset consisted of five core predictors:

1. High blood pressure
2. High cholesterol
3. Body mass index
4. General health
5. Age

These predictors were selected because they were statistically relevant, predictive, non-invasive, and easy to interpret. The reduced feature set also supports practical deployment because the input variables can be obtained through simple self-reporting or basic health screening.

Machine Learning Model Development

Seven classification algorithms were evaluated: Logistic Regression, Decision Tree, Random Forest, XGBoost, LightGBM, Artificial Neural Network and Naive Bayes. Hyperparameter optimization was reported using RandomizedSearchCV with stratified five-fold cross-validation. The recorded optimal Random Forest settings were max_depth = 20 and n_estimators = 200. Other reported settings were entropy criterion and max_depth = 20 for Decision Tree; C = 0.1 for Logistic Regression; learning_rate = 0.1 and n_estimators = 100 for XGBoost; n_estimators = 100 and num_leaves = 31 for LightGBM; and hidden_layer_sizes = (50, 50) for the neural network. These families of classifiers have been investigated in previous diabetes prediction research [3]–[6], [12]–[14]. Randomized hyperparameter search is described by Bergstra and Bengio [16].

Model performance was evaluated using accuracy, precision, recall, F1-score, and ROC-AUC. Since the dataset was imbalanced, ROC-AUC, recall, and F1-score were emphasized more than accuracy alone. Previous healthcare prediction studies have shown that evaluation should not rely on accuracy alone, especially when the target class is imbalanced and false negatives may have important consequences [3], [7], [8]. Random Forest was retained as the final model within the validated-feature configuration because it recorded the highest ROC-AUC and F1-score among the seven reported classifiers. Model selection was based on the recorded comparative metrics, rather than on any independently demonstrated superiority in external populations.

Model Stability Evaluation

Learning curves were used to describe the training–validation generalization gap. The recorded evaluation also included 10-fold cross-validation (separate from the five-fold tuning procedure) and reported the mean and standard deviation of accuracy, precision, recall, F1-score and ROC-AUC. The comparison was descriptive; without fold-level paired scores and the original evaluation code, formal significance of the between-pipeline differences cannot be established.

Stratified cross-validation was used to evaluate consistency across different data partitions. Mean performance and standard deviation were recorded for each metric. Cross-validation permits repeated internal evaluation across partitions, but fold-based comparisons require caution because training sets overlap [17].

The recorded statistical outputs were reviewed to clarify their interpretation. These outputs describe one-sample or reference-value accuracy assessments and a split-half stability check, rather than a documented paired comparison between the unfiltered and five-feature pipelines. Consequently, the current study reports descriptive between-pipeline differences and does not claim inferential superiority.

Graphical User Interface Development

A graphical user interface named DiabeteSense AI was developed using Streamlit as a proof-of-concept. The GUI allows users to enter non-invasive health information such as blood pressure status, cholesterol status, height, weight, general health condition, and age category. Body mass index is calculated automatically based on height and weight input. The GUI then presents a diabetes risk prediction output in a user-friendly format. The GUI was developed to demonstrate practical usability and feature-level interpretability. However, it is not intended to serve as a clinically validated diagnostic system. Further external validation and explainability analysis would be required before the system could be considered for healthcare deployment [7], [18].

RESULTS AND DISCUSSION

This section presents the results and discussion of the proposed statistically validated hybrid machine learning framework. The analysis focuses on four main aspects: the overall framework structure, feature interpretability, model performance, and model stability. The results are discussed in relation to the main objective of this paper, which is to enhance interpretability and stability in diabetes risk prediction rather than merely maximizing raw predictive accuracy.

Overview of the Proposed Framework

The proposed framework was designed to integrate statistical feature validation with machine learning-based feature refinement. The workflow begins with dataset acquisition and preprocessing, followed by statistical feature validation, Recursive Feature Elimination, machine learning model development, stability evaluation, and graphical user interface implementation.



Figure 1. Proposed statistically validated hybrid machine learning framework for diabetes risk prediction.

Figure 1 summarizes the proposed workflow, including preprocessing, statistical screening, Random Forest-based RFE, model evaluation and a Streamlit proof-of-concept. Statistical methods were used to screen candidate predictors, while Mutual Information provided complementary rankings. The model assessment is based on recorded internal evaluation, rather than external clinical validation. The GUI illustrates non-invasive risk screening and should not be interpreted as a diagnostic decision system.

Statistical Feature Validation and Feature Reduction

The analytical dataset was reported as 22 variables in total, including one diabetes outcome and 21 candidate predictors. Statistical validation produced eight reported candidate features, and RFE retained five predictors for final modelling.

Several features showed consistent statistical relevance across different validation techniques. High blood pressure, high cholesterol, body mass index, general health, and age were repeatedly identified as important predictors. These predictors are also medically plausible because they represent common risk indicators associated with diabetes risk. High blood pressure and high cholesterol reflect cardiovascular and metabolic risk, body mass index represents obesity-related risk, general health captures self-reported health condition, and age represents demographic risk.

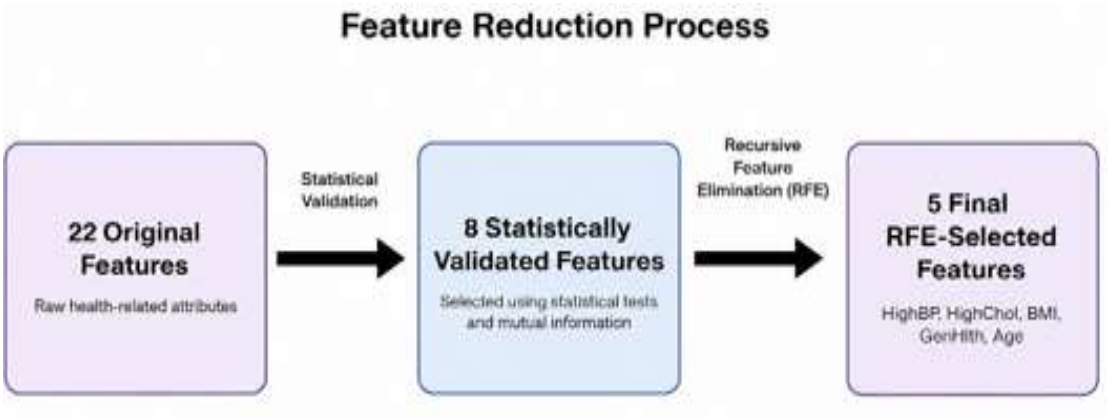

Figure 2. Feature reduction process from the original feature set to the final selected predictors.

Figure 2 summarizes the reported reduction from 22 dataset variables (including the outcome) through eight screening-stage candidates to five RFE-selected predictors: HighBP, HighChol, BMI, GenHlth and Age. The diagram displays the reported stages; it should not be interpreted as evidence of a fully specified, automated rule for combining all four statistical rankings.

This reduction is significant from an interpretability perspective. A model that depends on many input variables may be difficult to interpret, especially when several features are weak, redundant, or overlapping. By reducing the input space to five meaningful predictors, the proposed framework makes the model easier to understand and easier to explain to users. This is particularly important in healthcare-related prediction systems, where users and practitioners should be able to identify the main factors that influence a risk prediction.

The final selected features are also practical because they are non-invasive and easy to obtain. High blood pressure and high cholesterol can be self-reported or obtained from routine screening. Body mass index can be calculated using height and weight. General health and age can be collected through simple user input. Therefore, the feature reduction process not only improves model interpretability but also supports the practical usability of the framework for preliminary diabetes risk screening.

However, the interpretability achieved in this study should be understood as feature-level interpretability. The model becomes more transparent because it uses fewer and more meaningful predictors. This does not replace advanced explainable artificial intelligence methods such as SHAP, LIME, permutation importance, or partial dependence analysis. These methods should be considered in future work to provide deeper model-

level explanation. Explainable-AI approaches in healthcare provide additional methods for interpreting model behaviour [18].

Computational Efficiency of Feature Selection

The proposed statistical validation stage also improved the computational efficiency of the feature selection process. When Recursive Feature Elimination was applied directly to the original feature set, the process required 12 minutes and 12.1 seconds. After statistical validation was applied as a preliminary filtering stage, the RFE process required only 1 minute and 38.9 seconds. This represents an approximate 86.5% reduction in processing time.

Although computational efficiency is not the primary focus of this paper, this result supports the practical value of the proposed framework. RFE is a wrapper-based feature selection method and can become computationally expensive when applied to a larger number of features. By removing weak or irrelevant variables before RFE, statistical validation reduces the search space and allows the algorithm to operate more efficiently.

This finding strengthens the argument that statistical validation provides benefits beyond interpretability. It also improves the efficiency of the model development pipeline, particularly when working with large-scale healthcare survey datasets. In practical deployment, this advantage is useful when the model needs to be retrained, updated, or adapted using new data.

Machine Learning Model Performance

After feature validation and RFE, several supervised machine learning models were evaluated. The evaluated classifiers included Random Forest, Decision Tree, Logistic Regression, XGBoost, LightGBM, Artificial Neural Network, and Naive Bayes. Model performance was assessed using accuracy, precision, recall, F1-score, and ROC-AUC.

Model	Accuracy	Precision	Recall	F1-score	ROC-AUC
Random Forest	0.7274	0.5228	0.7877	0.6285	0.8223
Decision Tree	0.7275	0.5231	0.7828	0.6272	0.8220
Logistic Regression	0.7277	0.5241	0.7604	0.6205	0.8140
XGBoost	0.7245	0.5194	0.7869	0.6258	0.8208
LightGBM	0.7228	0.5173	0.7911	0.6256	0.8204
ANN	0.7241	0.5190	0.7861	0.6252	0.8190
Naive Bayes	0.7199	0.5146	0.7614	0.6141	0.8088

Table 1. Model Performance Comparison

Table 1 compares the performance of the seven machine learning classifiers. The results show that all models produced relatively close accuracy values, ranging from 0.7199 to 0.7277. Logistic Regression achieved the highest accuracy value of 0.7277, but its recall and ROC-AUC were lower than those of Random Forest. This indicates that accuracy alone is not sufficient for selecting the final model, especially in a healthcare screening context where identifying at-risk individuals is important.

Random Forest achieved the highest ROC-AUC value of 0.8223 and the highest F1-score of 0.6285, while also maintaining strong recall at 0.7877. This indicates that Random Forest provided the best balance

between discriminative ability and positive case detection. Although LightGBM achieved slightly higher recall than Random Forest, its precision, F1-score, and ROC-AUC were slightly lower. Therefore, Random Forest was selected as the final model because it provided the most balanced performance across the evaluation metrics.

The result should not be interpreted as a claim that Random Forest is universally superior for diabetes prediction. Rather, within the context of the statistically validated feature subset used in this study, Random Forest demonstrated the strongest balance between recall, F1-score, ROC-AUC, and stability. This interpretation is more appropriate for healthcare-oriented machine learning, where model reliability and generalization are as important as raw classification performance.

Cross-Validation Stability

The Random Forest classifier was assessed using reported 10-fold cross-validation after model development, separately from the stratified five-fold procedure described for randomized hyperparameter tuning. Table 2 provides the means and standard deviations of the recorded performance metrics.

Metric	**Mean ± Standard Deviation**
Accuracy	0.7282 ± 0.0025
Precision	0.5237 ± 0.0029
Recall	0.7913 ± 0.0050
F1-score	0.6303 ± 0.0033
ROC-AUC	0.8236 ± 0.0025

Table 2. Random Forest 10-Fold Cross-Validation Performance

Table 2 presents the 10-fold cross-validation performance of the Random Forest classifier using the statistically validated feature subset. The results show that the model achieved consistent performance across different folds, as indicated by the low standard deviation values for all evaluation metrics. Accuracy recorded a mean value of 0.7282 ± 0.0025, while ROC-AUC achieved 0.8236 ± 0.0025, indicating stable discriminative ability across different data partitions.

The recall value of 0.7913 ± 0.0050 is particularly important in the context of early diabetes risk screening because the model should be sensitive enough to identify individuals who may be at risk. The F1-score of 0.6303 ± 0.0033 also suggests a reasonable balance between precision and recall. Although the precision value was lower than recall, this trade-off is acceptable in a screening-oriented framework where reducing missed at-risk cases is more important than maximizing precision alone.

The small standard deviations show limited variation among the reported cross-validation folds for the five-feature Random Forest configuration. These values describe internal consistency within the study dataset but do not, by themselves, establish external generalizability or the absence of data leakage.

Learning Curve and Generalization Gap

Learning-curve results were examined using the difference between training and validation scores at the reported training-set volume. A narrower gap indicates closer agreement between these scores but does not, by itself, establish superior discrimination, calibration or performance in a new population.

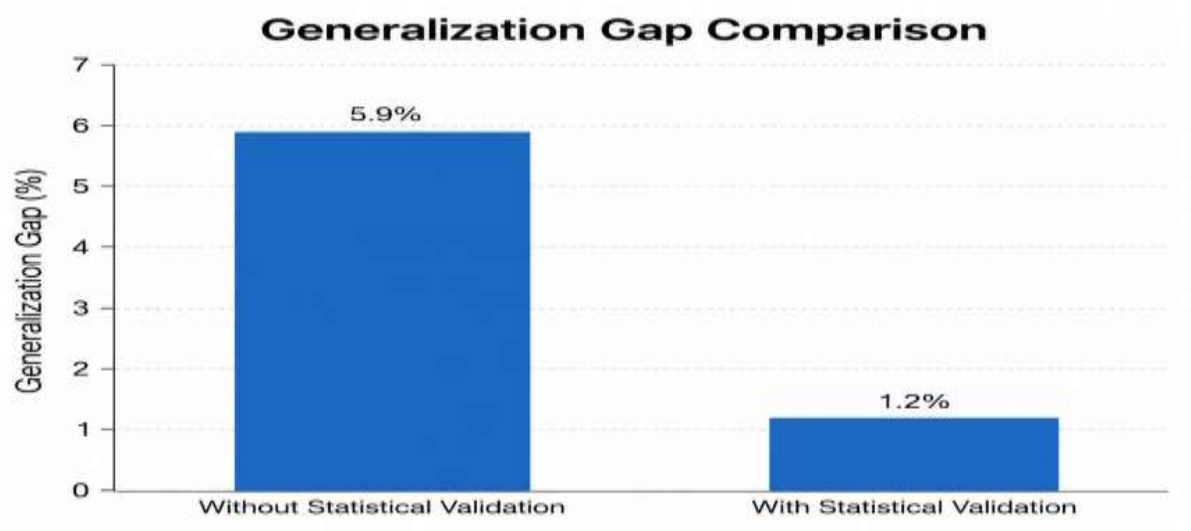

Figure 3. Comparison of generalization gap before and after statistical feature validation.

Figure 3 shows that the reported Random Forest training–validation gap was 5.9% without statistical screening (93.6% training and 87.7% validation) and 1.2% with screening (76.7% training and 75.5% validation). These values describe the observed relationship between training and validation scores under the documented experimental settings.

The smaller gap after feature reduction suggests less apparent overfitting in this internal evaluation. However, the fall in both training and validation scores also indicates an important loss of absolute predictive performance. Consequently, gap reduction must not be presented as proof that the screened model is uniformly more accurate or more generalizable.

The main contribution is therefore a more compact predictor set with lower recorded cross-validation variability and a narrower training–validation gap. In contrast, the unfiltered Random Forest configuration achieved higher absolute ROC-AUC and accuracy. The resulting trade-off between dimensionality, observed stability and predictive discrimination should be considered when choosing a screening model.

Baseline Comparison and Statistical Testing Clarification

The baseline configuration was the Random Forest classifier using the original candidate predictors without the additional statistical validation stage. The proposed configuration used the eight screening-stage candidates followed by Random Forest-based RFE to retain five final predictors. Table 3 provides the documented 10-fold cross-validation comparison. The baseline had higher average accuracy and ROC-AUC, whereas the five-feature configuration showed lower reported fold-to-fold variability and a narrower learning-curve gap.

Configuration	**10-fold Accuracy (mean ± SD)**	**10-fold ROC-AUC (mean ± SD)**	**Generalization gap**
Unfiltered RF baseline	0.8498 ± 0.0067	0.9423 ± 0.0072	5.9%
Statistical validation + RFE + RF	0.7282 ± 0.0025	0.8236 ± 0.0025	1.2%

Table 3. Descriptive Baseline and Proposed Pipeline Comparison

The baseline and proposed results in Table 3 are compared descriptively. The available original statistical outputs concerned classification accuracy against a fixed reference value and a split-half stability assessment. They do not provide documented paired scores for a direct inferential comparison of the two pipelines. Accordingly, the previously reported paired-test significance claims have been withdrawn.

A valid statistical comparison of the two pipelines would require evaluation on appropriately matched partitions or paired predictions and a testing procedure that accounts for the resulting dependence. Such paired data were not available for verification. The study therefore does not claim a statistically significant between-pipeline difference. Statistical test selection for classifier comparisons should account for study design and potential Type I errors [17].

Thus, the evidence for the framework is limited to feature-count reduction, recorded computational time, internal cross-validation summaries and the descriptive training–validation gap. These findings can motivate further assessment, but not establish statistical superiority or clinical utility.

GUI Implementation

The final Random Forest model was implemented in a Streamlit-based graphical user interface known as DiabeteSense AI. The interface was developed as a proof-of-concept to demonstrate how the statistically validated feature subset can be translated into a practical non-invasive diabetes risk screening tool. The GUI allows users to enter simple health-related inputs, including high blood pressure status, high cholesterol status, height, weight, general health condition, and age category. Body mass index is calculated automatically from height and weight. The system then generates a diabetes risk prediction output in a user-friendly format using the selected predictors identified by the proposed framework.

The implementation supports practical interpretability because the prediction process is based on a small number of familiar and meaningful health indicators. Users are not required to provide a large number of complex variables, which makes the system more accessible for preliminary screening purposes. This also demonstrates the practical value of reducing the original feature space to five core predictors. However, DiabeteSense AI should be interpreted only as a proof-of-concept system and not as a clinically validated diagnostic tool. The current implementation is intended to demonstrate usability and framework applicability. Further external validation using local clinical datasets, usability testing, ethical review, calibration analysis, and explainability assessment would be required before the system can be considered for real-world healthcare deployment.

OVERALL DISCUSSION

The study contributes a two-stage feature-screening approach focused on feature-level interpretability and descriptive internal model consistency. The recorded results show substantial feature reduction and lower variability across the reported validation folds, while also revealing a reduction in absolute predictive performance compared with the unfiltered baseline.

From the interpretability perspective, the recorded selection reduced 21 candidate predictors to five model inputs, within an analytical dataset containing 22 variables including the diabetes outcome. The retained inputs were high blood pressure, high cholesterol, BMI, general health and age. This supports a more concise model-input representation, although feature reduction does not by itself explain individual Random Forest decisions [7], [18].

From the internal stability perspective, the training–validation gap decreased from 5.9% to 1.2%, and the five-feature configuration showed lower cross-validation standard deviations. However, the unfiltered Random Forest achieved higher absolute accuracy and ROC-AUC in the thesis tables. A narrower gap and lower fold-to-fold variation therefore indicate a more limited form of observed consistency, not verified superiority in a new population. Another important finding is the distinction between statistical relevance and predictive usefulness. Statistical validation identifies features that have meaningful relationships with diabetes status, while Recursive Feature Elimination identifies the subset of features that contributes most effectively to model performance. Combining these two stages makes the feature selection process more defensible than relying solely on algorithmic feature ranking. This supports the broader role of feature selection in reducing dimensionality, minimizing redundancy, and improving model reliability in diabetes prediction [9], [10], [15].

Several limitations should be acknowledged. The study used an internally evaluated BRFSS-derived survey dataset rather than a separate, representative external population. Independent validation in a Malaysian population remains a future direction. The archived research materials do not fully specify the source-extraction audit trail, exact combination rule for statistical feature rankings, complete randomized tuning search space, or the placement of feature selection and imputation inside cross-validation folds; therefore, full code-level reproducibility and freedom from leakage could not be independently established. Calibration, decision-curve analysis, demographic subgroup comparisons and formal between-pipeline confidence intervals were not reported. Furthermore, feature reduction was not accompanied by model-level explanations such as SHAP or permutation importance. These limits constrain claims about generalizability, calibration, clinical utility and fairness [8], [18]. Overall, the framework should be viewed as an internally evaluated, non-invasive proof-of-concept for preliminary diabetes risk screening. It is not a diagnostic replacement or clinically validated decision system. Its observed value lies in a smaller input set and recorded internal performance characteristics that warrant further independent validation.

CONCLUSION

This study described a statistically screened, RFE-refined machine learning framework for preliminary diabetes risk screening. The framework applied Chi-Square tests, independent-sample t-tests, Spearman correlation and Mutual Information for initial feature screening, followed by Random Forest-based RFE and classifier comparison.

The analytical dataset was reported to comprise 302,443 records and 22 variables (one outcome and 21 candidate predictors). Eight candidate predictors were reported after statistical screening, and five were retained after RFE: high blood pressure, high cholesterol, BMI, general health and age. The final Random Forest configuration achieved a reported 10-fold ROC-AUC of 0.8236 ± 0.0025 and F1-score of 0.6303 ± 0.0033. The learning-curve gap decreased from 5.9% to 1.2%, while the unfiltered baseline continued to show higher absolute accuracy and ROC-AUC. These findings document a trade-off between dimensionality, internal variability and predictive discrimination rather than conclusive superiority of the screened model. The study supports the potential of combining statistical screening and RFE for reducing feature dimensionality in survey-based prediction. However, it does not establish external generalizability, calibration, clinical net benefit or demographic fairness. Future work should reproduce the full pipeline from code with fold-specific preprocessing and selection, undertake external validation using a representative independent population, and investigate calibration, decision-curve performance, subgroup evaluation and model-level explanations before real-world screening use.

ACKNOWLEDGMENT

The authors would like to express their appreciation to the Faculty of Electronic and Computer Engineering, Universiti Teknikal Malaysia Melaka, for providing academic support, facilities, and a conducive research environment throughout the completion of this study.

Ethical Consideration

This study used a publicly available and anonymized secondary dataset. No direct human participant recruitment or clinical intervention was conducted by the authors.

Data Availability Statement

The dataset used in this study was obtained from a publicly available BRFSS-based diabetes health indicators dataset.

Conflict of Interest

The authors declare that there is no conflict of interest regarding the publication of this paper.

REFERENCES

1. International Diabetes Federation, "Facts & Figures: Diabetes around the world in 2024," 2024. [Online]. Available: https://idf.org/about-diabetes/diabetes-facts-figures/
2. World Health Organization, "Diabetes," Fact sheet, Oct. 05, 2023. [Online]. Available: https://www.who.int/news-room/fact-sheets/detail/diabetes
3. M. Maniruzzaman, M. J. Rahman, M. Al-MehediHasan, H. S. Suri, M. M. Abedin, A. El-Baz, and J. S. Suri, "Accurate diabetes risk stratification using machine learning: Role of missing value and outliers," Journal of Medical Systems, vol. 42, Art. no. 92, 2018, doi: 10.1007/s10916-018-0940-7.
4. 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, doi: 10.1016/j.csbj.2016.12.005.
5. J. J. Khanam and S. Y. Foo, "A comparison of machine learning algorithms for diabetes prediction," ICT Express, vol. 7, no. 4, pp. 432–439, 2021, doi: 10.1016/j.icte.2021.02.004.
6. L. Kopitar, P. Kocbek, L. Cilar, A. Sheikh, and G. Stiglic, "Early detection of type 2 diabetes mellitus using machine learning-based prediction models," *Scientific Reports*, vol. 10, Art. no. 11981, 2020, doi: 10.1038/s41598-020-68771-z.
7. E. Christodoulou, J. Ma, G. S. Collins, E. W. Steyerberg, J. Y. Verbakel, and B. Van Calster, "A systematic review shows no performance benefit of machine learning over logistic regression for clinical prediction models," *Journal of Clinical Epidemiology*, vol. 110, pp. 12–22, 2019, doi: 10.1016/j.jclinepi.2019.02.004.
8. R. D. Riley, K. I. E. Snell, J. Ensor, D. L. Burke, F. E. Harrell, K. G. M. Moons, and G. S. Collins, "Minimum sample size for developing a multivariable prediction model: Part II—binary and time-to-event outcomes," Statistics in Medicine, vol. 38, no. 7, pp. 1276–1296, 2019, doi: 10.1002/sim.7992.
9. J. Kaliappan et al., "Analyzing classification and feature selection strategies for diabetes prediction across diverse diabetes datasets," Frontiers in Artificial Intelligence, vol. 7, Art. no. 1421751, 2024, doi: 10.3389/frai.2024.1421751.
10. L. Jiang, Z. Yang, G. Liu, Z. Xia, G. Yang, H. Gong, J. Wang, and L. Wang, "A feature optimization study based on a diabetes risk questionnaire," Frontiers in Public Health, vol. 12, Art. no. 1328353, 2024, doi: 10.3389/fpubh.2024.1328353.
11. Z. Xie, O. Nikolayeva, J. Luo, and D. Li, "Building risk prediction models for type 2 diabetes using machine learning techniques," Preventing Chronic Disease, vol. 16, Art. no. 190109, 2019, doi: 10.5888/pcd16.190109.
12. E. Dritsas and M. Trigka, "Data-driven machine-learning methods for diabetes risk prediction," Sensors, vol. 22, no. 14, Art. no. 5304, 2022, doi: 10.3390/s22145304.
13. R. F. Albadri and S. M. Awad, "A diabetes prediction model using hybrid machine learning algorithm," Mathematical Modelling of Engineering Problems, vol. 11, no. 8, pp. 2119–2126, 2024, doi: 10.18280/mmep.110813.
14. P. Alagumariappan, M. Sathyamoorthy, R. K. Dhanaraj, et al., "Optimized hybrid machine learning framework for early diabetes prediction using electrogastrograms," *Scientific Reports*, vol. 15, Art. no. 8875, 2025, doi: 10.1038/s41598-025-93495-3.
15. B. Gregorutti, B. Michel, and P. Saint-Pierre, "Correlation and variable importance in random forests," *Statistics and Computing*, vol. 27, no. 3, pp. 659–678, 2017, doi: 10.1007/s11222-016-9646-1.
16. J. Bergstra and Y. Bengio, "Random search for hyper-parameter optimization," *Journal of Machine Learning Research*, vol. 13, pp. 281–305, 2012.
17. T. G. Dietterich, "Approximate statistical tests for comparing supervised classification learning algorithms," *Neural Computation*, vol. 10, no. 7, pp. 1895–1923, 1998, doi: 10.1162/089976698300017197
18. Z. Sadeghi, R. Alizadehsani, M. A. Cifci, et al., "A brief review of explainable artificial intelligence in healthcare," *arXiv preprint arXiv:2304.01543*, 2023.