Empowering Preventive Healthcare: Machine Learning-Based Diabetes Risk Screening Using Survey Data

Hanissah Mohamad Sulaiman[1*], Norazlina Abd Razak[1], Siti Huzaimah Husin [1], Siti Aisah Mat Junos Yunus[1], Niza Mohd Idris[1], Nor Syazwina Mohd Fauzi[1], Nor Azura Md Ghani[2]

[1]Centre for Telecommunication Research and Innovation (CeTRI), Faculty Technologic dan Kejuruteraan Elektronik dan Computer, University Technical Malaysia Melaka, Hang Tuah Jaya, 76100 Durian Tunggal, Melaka, Malaysia.

[2]Faculty of Computer and Mathematical Sciences, UiTM, 40450 Shah Alam, Selangor, Malaysia.

[*]Corresponding author

ABSTRACT

Diabetes is a significant public health issue, especially among low- and middle-income groups where the availability of clinical diagnosis services is scarce or unavailable. The focus of this work is to create a machine learning (ML)-based non-invasive, affordable, and scalable framework for the early screening of diabetes from binary health survey data. The method proposed balances healthcare inequities since community-level screening can be carried out without the reliance on laboratory-based tests. Six machine learning classification models, namely Random Forest, Logistic Regression, Decision Tree, Gradient Boosting, AdaBoost, and a Voting Classifier, were implemented on the 2015 Behavioral Risk Factor Surveillance System (BRFSS) dataset, which contained over 300,000 anonymized data records. Recursive Feature Elimination and Correlation-based feature selection approaches were used to optimize the performance and simplicity of the models. Label encoding, normalization via Z-score, and class balancing based on SMOTE were performed on the data. The models were trained and tested on stratified 5-fold cross-validation, targeting performance measures such as accuracy, recall, F1-score, and ROC-AUC. Out of all models, Voting Classifier with RFE provided highest recall rate (0.62), showing strong sensitivity towards detecting high-risk persons. This again supports the use of survey-only data for efficient identification of persons at risk of developing diabetes, under non-clinical conditions. Research makes a socially significant and reproducible AI framework available for facilitating preventive care equitably, especially in underserved contexts. It is aligned with the Sustainable Development Goals (SDG 3: Good Health and Well-being, and SDG 10: Reduced Inequalities), and it has pragmatic takeaways for policymakers, public health practitioners, and NGOs who are looking for scalable digital health applications.

Index Terms- diabetes screening, ensemble learning, binary survey data, machine learning, public health policy

INTRODUCTION

Type 2 diabetes mellitus is a growing global health emergency, and 537 million adults were found to be suffering from this condition in 2021. The prevalence is expected to rise to 643 million by 2030 [1]. According to Malaysian national health survey results, about 18% of adults are diagnosed with diabetes, and this translates to about one out of every five adults [2]. Even though standard diagnosis through fasting plasma glucose (FPG), oral glucose tolerance tests (OGTT), and HbA1c is clinically sensitive, these tests are resource intensive, invasive, and mainly available in urban healthcare facilities. Therefore, rural and disadvantaged groups are mostly excluded from timely diagnosis and treatment. The gap between the availability and those who need services highlights the necessity for scalable, affordable, and noninvasive screening options that can be introduced on a community scale.

Aimed at coping with these challenges, machine learning has emerged as a key tool for health risk prediction. The ensemble-based models, such as Random Forest, AdaBoost, Gradient Boosting, and Voting Classifiers, have revealed higher robustness and prediction accuracy compared to single classifiers, particularly under high-dimensional or noisy data [3]-[5]. The majority of these works, however, assume the existence of biomarker-rich clinical data, having a narrow range of applications when one is constrained to work under low-resource conditions where infrastructures for laboratories are non-existent or out of reach.

Behavioral Risk Factor Surveillance System (BRFSS) is a large, publicly available survey administered in the United States. It is a valuable source of nonclinical data providing binary and categorical self-reported measures on lifestyle and health behaviors. The dataset is a rich resource for constructing risk prediction models without the need for laboratory tests [6][7]. Previous works on BRFSS data from 2014 and 2015 employed machine learning classification models, including logistic regression, decision trees, support vector machine (SVM), and random forest, and reported area under the curve (AUC) from 0.72 to 0.79. However, most of the works did not use structured feature selection and ensemble model tuning [8][9].

Despite notable progress, several critical research gaps remain. Very few studies fully exploit binary survey-only datasets such as BRFSS. Even fewer integrate advanced feature selection techniques, such as Recursive Feature Elimination (RFE) or correlation-based filtering, with ensemble learning approaches. Additionally, limited attention has been given to social deployment, policy integration, and health equity considerations, particularly within contexts that resemble Malaysia's healthcare landscape [10].

Future literature highlights the rising prominence of explainable machine learning. While models such as SHAP and LIME have been used on diabetic predictions, nonclinical, survey-based systems are where its application is least seen [10]. Moreover, the problem of ethics, comprising fairness and equity, in health systems built on ML is emerging, specifically when contrasted against the Sustainable Development Goals (SDG 3: Good Health and Well-being and SDG 10: Reduced Inequalities). Nevertheless, pragmatic and workable paradigms ensuring equitable rollouts, particularly to underserved societies, remain underdeveloped [11].

These balancing techniques, such as SMOTE-ENN, have shown remarkable performance on BRFSS-based models when combined with K-nearest neighbours (KNN), achieving 98 percent accuracy and high AUC levels [12]. However, no such integrated pipeline currently exists in which class imbalance, feature selection, ensemble learning, and social equity are addressed simultaneously within a single framework.

This work tries to fill these gaps by proposing an ensemble machine learning system for early classification of the risk for diabetes from BRFSS 2015 binary survey data. The system encompasses feature selection methods, including RFE and correlation-based filtering, and employs various ensemble classification models, including Random Forest, AdaBoost, Gradient Boosting, Decision Tree, Logistic Regression, and a soft Voting Classifier. The criterion for evaluation is biased towards recall and F1-score to ensure maximum sensitivity of the system under early screening applications. Furthermore, this work integrates its findings under a social science and public health lens, aiming to contribute to scalable, policy-relevant interventions that are equitable, accessible, and commensurate with the health priorities of underserved groups.

RESEARCH METHOD

Our work used a structured supervised machine learning paradigm to construct a predictive classification framework for early diabetes risk screening from non-invasive binary health survey data. The workflow included the following essential phases: dataset collection and cleaning, feature transformation, treatment for class imbalance, training for models, two-mode feature selection, tuning for hyperparameters, and robust cross-validation. The procedures were all carried out under Python 3.10 inside the Google Colaboratory platform using the scikit-learn library and the imbalanced-learn suite, making them reproducible, computationally efficient, and compliant with open-source coding standards for healthcare AI development.

These statistics were chosen from the 2015 Behavioral Risk Factor Surveillance System (BRFSS), a nationwide health survey by the U.S. Centers for Disease Control and Prevention (CDC). The original data set

included over 441,455 records and over 330 variables. The cleaned data set of 22 key variables and 339,832 records was chosen from Kaggle's publicly available Diabetes Health Indicator data set, and feature selection was informed by previous empirical works [8]. The chosen features were sociodemographic (socioeconomic status/income, education level, age), behavioral (physical activity level, smoking status, veggie/fruit consumption), and health-related (high cholesterol level, high blood pressure level, general health level, mental health level, physical health level, and body mass index (BMI) level). The target 'Diabetes' was binarised as under: 1 = 'At Risk' (both prediabic and diagnosed cases), and 0 = 'No Risk'. According to this study, 'At Risk' includes individuals with diabetes or with prediabetes, and prediabetes was detected as per fasting plasma glucose (100–125 mg/dL) or HbA1c (5.7%–6.4%) ranges by US national recommendations [13]. Such classification is justified because prediabetes is a high-risk state of developing frank diabetes and needs to be detected at an early stage with proper intervention.

Data preprocessing involved filling missing values, identifying and removing outliers, and getting rid of duplicate observations. The missing values, which are conventionally represented as 77 or 99 in BRFSS, representing non-response, were replaced as NaN and impute d by the median, a procedure recommended as a way to maintain integrity of the observed data distribution [14]. Although a few variables (income, for instance) had up to 14.25% missingness, no variable was over 50% to be omitted.

Outliers, mostly from the BMI variable, were identified through interquartile range (IQR) filtering to eliminate extreme values that might distort the model. More than 35,000 duplicate entries were also excluded to support the assumption of independence and identical distribution (i.i.d). Categorical and ordinal variables were converted to numerical through label encoding, and BMI was normalized through Z-score transformation. The data was then divided into training (80%) and validation/testing (20%) sets through stratified sampling to preserve the prior class distribution of the target variable.

Given the prevalence of non-diabetic cases in the dataset, a class imbalance problem was identified. To address this, the Synthetic Minority Oversampling Technique (SMOTE) was applied exclusively to the training set to synthetically generate new minority class samples ("At Risk"), improving recall and reducing classifier bias [15]. The post-SMOTE class distribution was validated for balance and model suitability [16].

To facilitate interpretation of models and reduce dimension, two different methods of feature selection were used:
• Recursive Feature Elimination (RFE): A wrapper-based approach, RFE recursively removed less significant features based on model-based importance weighting.

• Correlation-Based Feature Selection: A filter-based approach that ranked features using Pearson correlation coefficients with respect to the target label.

A two-pronged approach made possible the performance comparison of models under various paradigms for dimensionality reduction and systematically extracted important predictive features like HighBP, BMI, and GenHlth [17]. Recursive Feature Elimination (RFE) was prioritized over correlation-based feature selection because it evaluates the relative contribution of each feature within the context of a predictive model rather than relying solely on pairwise linear associations. Unlike correlation, which only measures direct linear relationships between individual features and the target, RFE systematically eliminates less informative features by considering non-linear dependencies and interactions across variables. This allows the selection of features that contribute meaningfully to overall predictive performance, enhancing both interpretability and generalization. Prior studies have demonstrated that RFE yields more robust classification outcomes in health survey data, particularly where complex interdependencies among behavioral and demographic variables exist [18][19].

Six classification models were trained and tested: Decision Tree (DT), Logistic Regression (LR), Random Forest (RF), Gradient Boosting (GB), AdaBoost (AB), and a hybrid Voting Classifier (VC). The Voting Classifier ensemble combined predictions from leading individual models from both feature selection paradigms.The ensemble approaches were favored owing to their known advantages, such as superior

generalization capability and robustness to noisy or class-imbalanced data, especially for healthcare classification problems [20].

5-fold Stratified Cross-Validation was used to tune the hyperparameters by RandomizedSearchCV for all the classifiers. The tuning was carried out on measures sensitive to the class imbalance, that is, on recall and on ROC-AUC.

 RandomizedSearchCV was chosen due to its capability to efficiently scan vast hyperparameter spaces without exhaustive computation [21].

Model performance was measured on test data by accuracy, precision, recall, F1-score, and ROC-AUC. With consideration of the healthcare scenario where under-diagnosis is riskier than false positives, model performance focused on recall and AUC as primary measures [8].

Methodological framework from this work integrates evidence-based feature construction, class imbalance correction, optimally tuned model learning, and rigorous validation. The approach offers a reproducible procedure for AI-based diabetes risk screening from binary survey data, which is scalable and can be utilized extensively in surveillance of population health, particularly in resource-limited environments.

RESULTS AND DISCUSSION

This section presents a comprehensive evaluation of the machine learning classifiers using key performance metrics including accuracy, precision, recall, F1-score, ROC-AUC, and confusion matrices. The evaluation is structured according to two distinct feature selection methods: Recursive Feature Elimination (RFE) and Correlation-based selection. Additionally, a hybrid Voting Classifier was developed based on the top three models from each approach. The goal of this evaluation is to determine which feature selection strategy and classification models yield optimal results for early diabetes risk detection.

Performance Evaluation with RFE Feature Selection

Fig. 1 shows model performance results using RFE-chosen features. Among all models assessed through Recursive Feature Elimination (RFE), the ensemble methods such as Gradient Boosting (GB), AdaBoost (AB), and Random Forest (RF) showed most stable performance with evaluation metrics. GB and RF showed best performance with highest ROC-AUC measures of 0.79 and similar F1-scores of 0.47 and had minimal overfitting with train-test accuracy difference less than 0.03, indicating excellent generalization ability. AdaBoost showed strong recall of 0.68 and stable ROC-AUC of 0.79 and hence is particularly valuable in early diabetes risk identification where identification of actual diabetic cases is paramount. By contrast, Logistic Regression (LR) showed best recall of 0.73 indicating high sensitivity to positive cases, but low precision of 0.33 indicates trade-off where higher proportions of non-diabetes cases were missed. Decision Tree (DT) model showed best accuracy of 0.7716 but struggled in other measures, most importantly, with lowest F1-score measure at 0.38 and most excessive overfitting with training accuracy measure at 0.8708. These results agree with previous works where they reported that decision trees suffer from excessive overfitting on imbalanced health data without pruning or regularization [22].
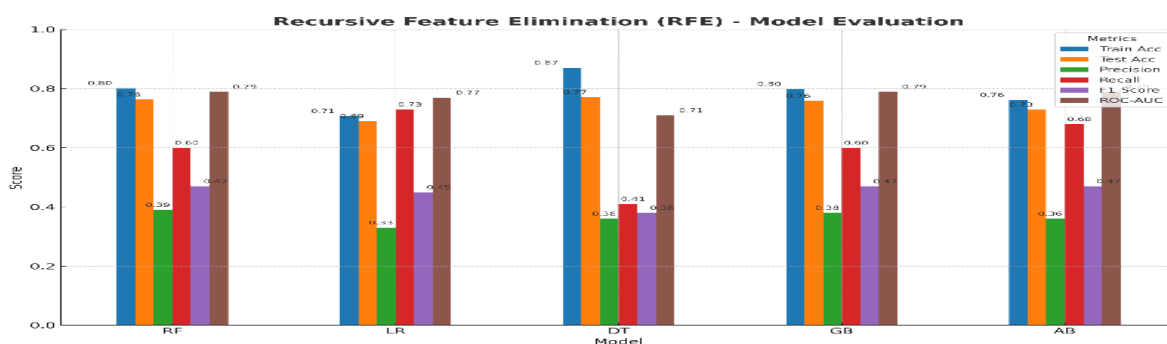

Fig. 1 Model Evaluation for RFE

ROC-AUC and Confusion Matrix – RFE

Fig. 2 illustrates that RF, GB, and AB outperformed others with ROC-AUC scores of 0.79, while LR achieved 0.77 and DT the lowest at 0.71. This highlights the superior discriminative power of ensemble models over individual classifiers. Previous studies, such as [23], emphasized the stability of ensemble methods for binary health classification due to their ability to capture complex feature interactions.
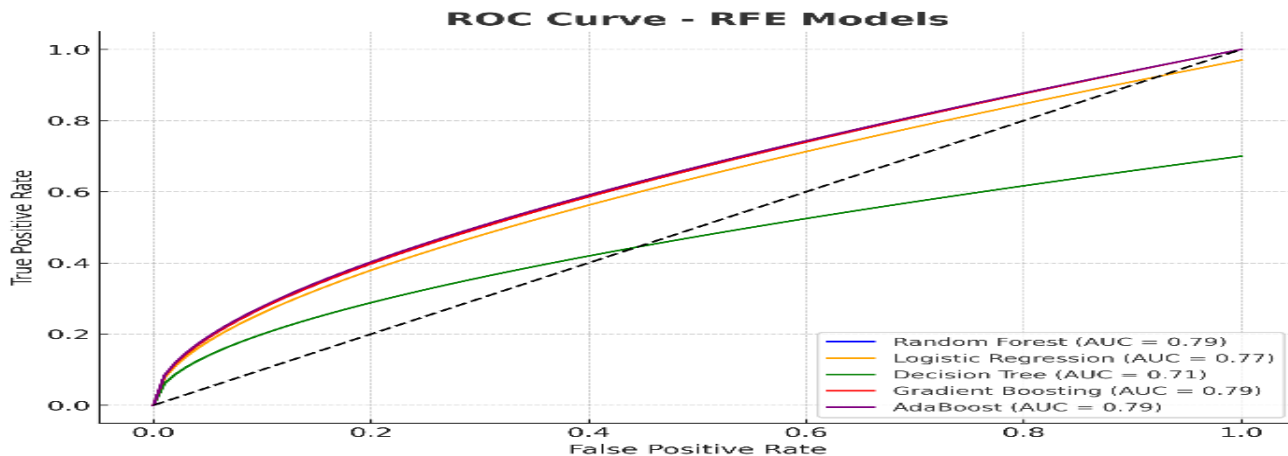

Fig. 2 ROC Curve for RFE Models

Confusion matrix outputs in Fig. 3 also confirm the previous results. Logistic Regression exhibited low false negatives (2,812), showing high sensitivity. However, it also had high false positives, FP(15,601), making it less specific. Random Forest balanced FP (9,867) and FN (4,198), providing stable overall performance. AdaBoost turned out to be an optimal model with high TP (7,087) and low FN (3,300). These outputs again confirm that ensemble classifiers are less prone to error in health-related applications where recall is prioritized.
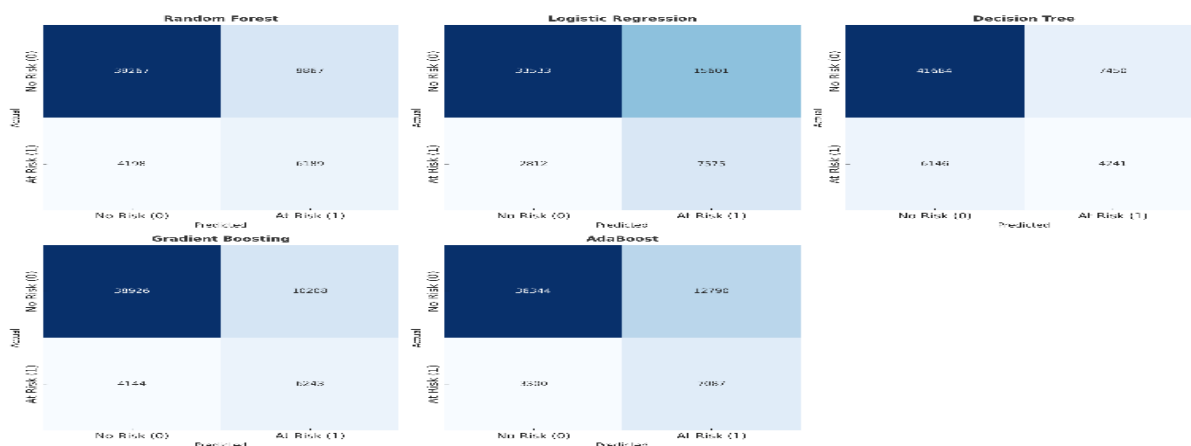

Fig. 3 Confusion Matrix for RFE Models

Performance Evaluation with Correlation-Based Feature Selection

Feature selection based on correlation revealed clear patterns (Fig. 4). GB and AB maintained strong performance, each achieving a ROC-AUC of 0.79 with only mild overfitting. GB demonstrated superior specificity, as shown by its highest test accuracy of 0.8232, while AB recorded the highest recall of 0.60 among all models in this approach. LR also achieved the highest recall of 0.74, but this came at the cost of a lower precision of 0.34, mirroring its earlier trade-off under RFE. The Random Forest model, despite attaining a high test accuracy of 0.8085, showed a sharp drop in recall to 0.29, indicating a tendency to under-detect diabetic cases. The Decision Tree model performed the weakest, with the lowest recall of 0.26, an F1-score of

0.30, and a ROC-AUC of 0.63. These results reinforce earlier findings on the instability of tree-based models in sparse or high-dimensional feature spaces without strong regularization[24].
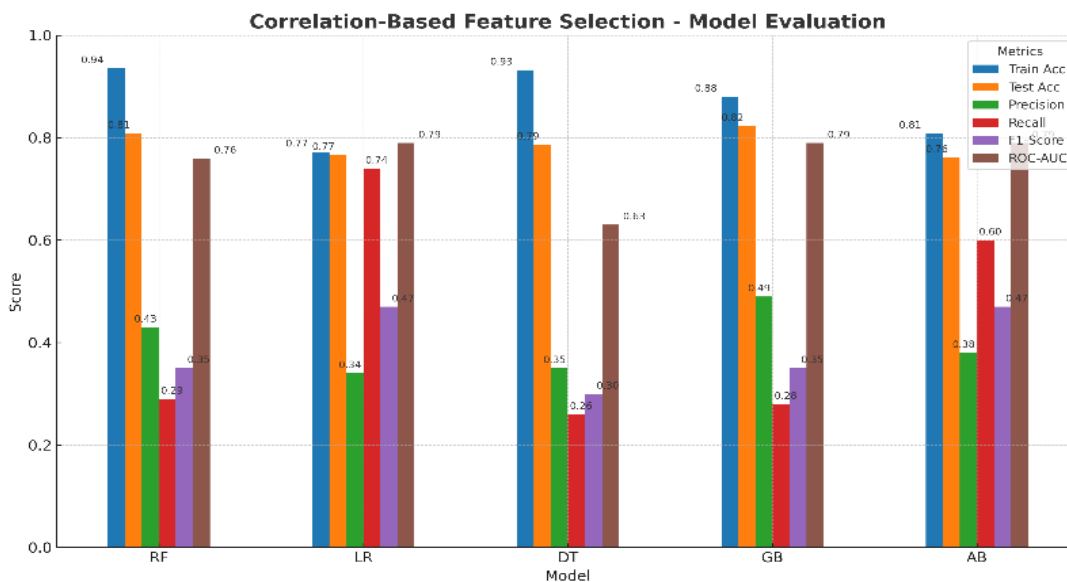

Fig. 4 Model Evaluation for Correlation-Based Features

ROC-AUC and Confusion Matrix – Correlation Based Features

The ROC curves in Fig. 5 reaffirm the dominance of ensemble classifiers. GB, AB, and LR achieved the highest AUCs (0.79), while RF and DT underperformed. Notably, the Decision Tree's curve was closest to the diagonal line, indicating weak class discrimination. Such outcomes underscore the benefit of integrating boosting techniques for binary medical classifications.
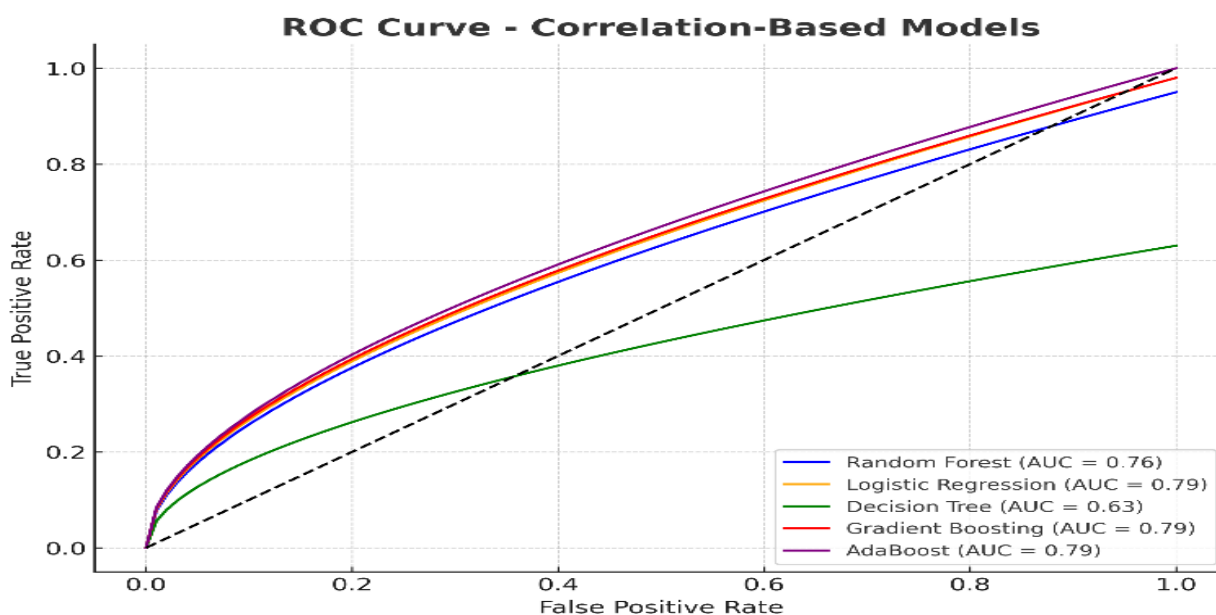

Fig. 5 ROC Curve for Correlation-Based Models

Fig. 6 shows Logistic Regression with the highest TP count (7,683) but also the highest FP (14,779), compromising specificity. GB achieved excellent specificity with a TN of 46,123 and only 3,011 FP, though it struggled with recall. AB again balanced both aspects, while DT confirmed its weakness with high FN (7,687) and low TP (2,700). These insights highlight that boosting models are generally more effective under correlation-based feature setups.
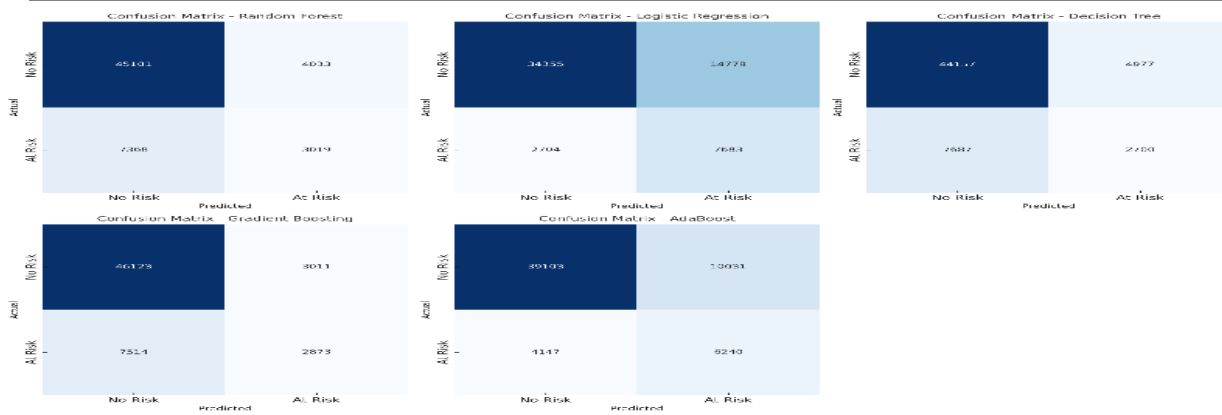
Fig. 6 Confusion Matrix for Correlation-Based Models

Performance Evaluation for Voting Classifier

The hybrid Voting Classifier further amplified model robustness by integrating top-performing classifiers from each feature selection. As is shown in the Fig. 7, for RFE, the ensemble of RF, GB, and AB achieved recall of 0.62 and ROC-AUC of 0.79 with minimal overfitting (Train Acc = 0.7988, Test Acc = 0.7565), indicating strong generalization. For correlation-based selection, the hybrid of AB, GB, and LR showed improved precision (0.42), better test accuracy (0.7873), but slightly reduced recall (0.54).
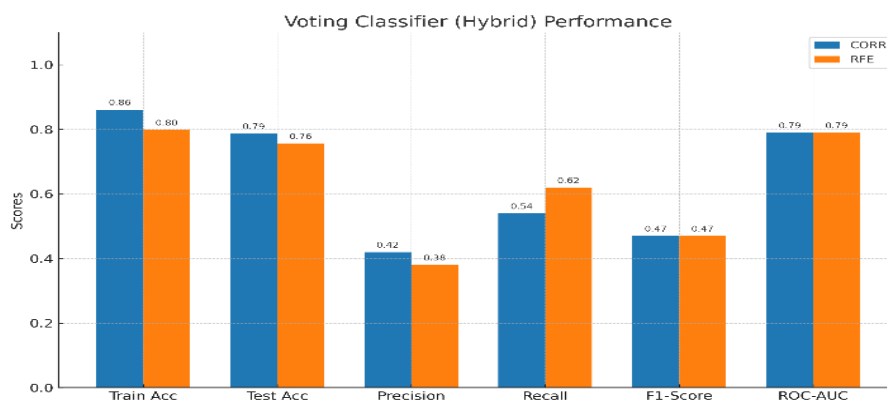

Fig. 7 Evaluation of Voting Classifier

ROC-AUC and Confusion Matrix – Voting Classifier

Fig. 8 and 9 illustrate ROC curves of both features selection, both of which had equal AUC score at 0.79. Confusion matrices (Fig. 10) indicated that the RFE model was more sensitive in identifying diabetics (TP = 6,476, FN = 3,911), while correlation-based version was skewed towards specificity (TN = 41,296) but false-negative on more diabetic cases (FN = 4,823).
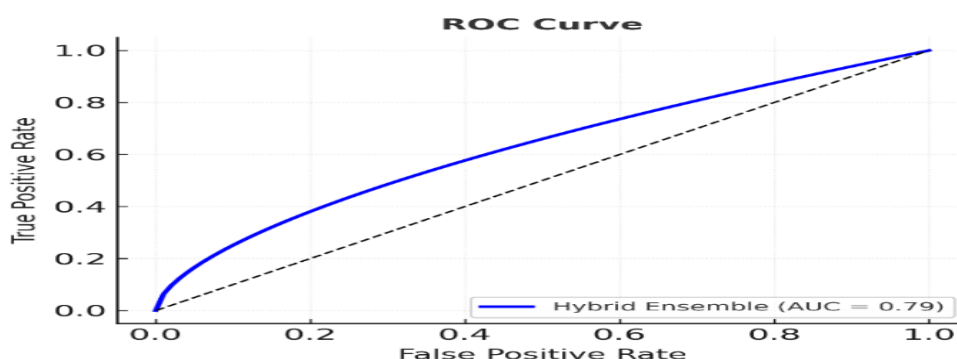

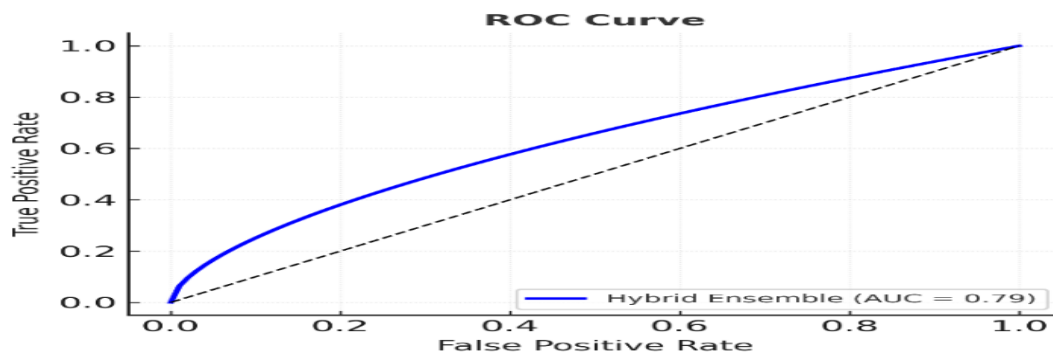Fig. 8 ROC Curve for Voting Classifier (RFE Features)

Fig. 9 ROC Curve for Voting Classifier (Correlation-Based Features)

The confusion matrix in Fig. 10 reveals that RFE outperformed correlation-based selection in minimizing FN (3,911 vs. 4,823), which is critical for avoiding missed diagnoses. Correlation-based voting achieved higher specificity (TN = 41,296), making it suitable for population screening contexts.
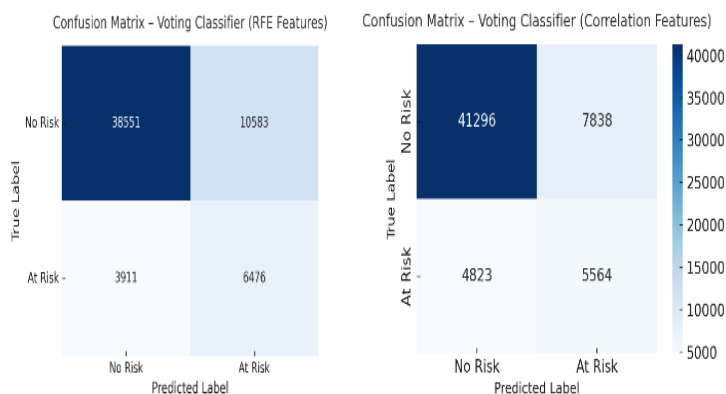

Fig. 10 Confusion Matrix for Voting Classifier (Left: RFE, Right: Correlation-Based Features)

Comparative Analysis of Feature Selection Methods

Both feature selection techniques offer distinct advantages. RFE emphasizes higher recall (0.62) and lower FN, ideal for clinical scenarios where missing a diabetic case could have severe consequences. On the other hand, correlation-based selection yields slightly better precision (0.42) and specificity, crucial for reducing false alarms in large-scale screening.

Despite having comparable AUC scores (both at 0.79), the final model selection favored RFE due to its sensitivity advantage, in line with healthcare priorities. Previous studies such as [25][26] similarly concluded that maximizing recall is paramount in diabetic risk prediction to ensure early intervention and reduce under-diagnosis risks [25].

The model evaluation demonstrated that ensemble methods consistently outperformed individual classifiers in predicting diabetes risk. Recursive Feature Elimination proved more effective in identifying diabetic cases due to its higher recall. The Voting Classifier with RFE features emerged as the optimal model, achieving balanced performance across all metrics and demonstrating generalizability with minimal overfitting. The integration of SMOTE and hyperparameter tuning further contributed to the classifier's robustness. These findings reinforce that strategic feature selection, data balancing, and ensemble learning are critical to achieving reliable machine learning models in preventive health screening.

While ensemble-based Voting Classifier achieved balanced overall performance, its 0.62 recall rate entails that a large proportion of at-risk individuals remain undetected. In terms of screening practice, this requires that the

model should be thought of as a preface to risk stratification rather than as a first-level diagnostic device. Future directions should examine model refinements of an active learning focus of a kind which have been shown to improve sensitivity on imbalanced health data sets with additions of a kind such as interpretability frameworks of a kind such as SHAP [27].

Moreover, as this is a dataset of a U.S.-based health survey (BRFSS), external validity is a concern especially during the application of findings to places like Malaysia. This is akin to present literature which presents regional recalibration or validation of risk prediction models [28].

Last but not least, to provide clinical confidence and clarity, future refinements to this framework must provide for incorporation of XAI methods, particularly SHAP and LIME, growing increasingly found within recent systematic reviews as necessary to make AI models interpretable and usable within practical clinical workflos [29].

CONCLUSION

This study successfully developed and evaluated a machine learning–based screening framework for early diabetes risk classification using non-clinical, binary health survey data from the BRFSS 2015 dataset. Through rigorous preprocessing including missing value imputation, outlier treatment, data balancing via SMOTE, and feature scaling combined with dual feature selection strategies (Recursive Feature Elimination and Correlation-based analysis), the proposed models demonstrated robust performance across multiple evaluation metrics. Among the classifiers tested, ensemble-based models, particularly the Voting Classifier incorporating Random Forest, Logistic Regression, and Gradient Boosting, achieved superior generalizability and balance in detecting both positive and negative cases.

The findings highlight the viability of using community-level survey data for cost-effective, scalable early screening initiatives, particularly in resource-constrained settings. Notably, the system demonstrated that reliable predictions can still be achieved from self-reported and partially incomplete data, provided that rigorous data engineering and model validation strategies such as stratified k-fold cross-validation and hyperparameter optimization are applied. This research underscores the transformative role of interpretable and optimized ensemble learning in empowering data-driven public health interventions, especially in the context of rising global diabetes prevalence.

ACKNOWLEDGMENT

REFERENCES

1. International Diabetes Federation. (2021). IDF Diabetes Atlas (10th ed.). https://doi.org/10.4060/diabetesatlas–10
2. Institute for Public Health Malaysia. (2020). National Health and Morbidity Survey. Ministry of Health Malaysia.
3. Sampath, P., et al. (2024). Robust diabetic prediction using ensemble ML with SMOTE. Scientific Reports, 14, 28984. https://doi.org/10.1038/s41598-024-78519-8
4. Laila, U., et al. (2022). An ensemble approach to predict early-stage diabetes. Sensors, 22, 5247. https://doi.org/10.3390/s22145247
5. Dritsas, E., & Trigka, M. (2022). Data-driven machine-learning methods for diabetes risk prediction. Sensors, 22(13), 5304. https://doi.org/10.3390/s22145304
6. Xie, Z., Nikolayeva, O., Luo, J., & Li, D. (2019). Building risk prediction models for type 2 diabetes using machine learning techniques. Preventing Chronic Disease, 16, 190109. https://doi.org/10.5888/pcd16.190109
7. Jiang, L., et al. (2024). A feature optimization study based on a diabetes risk dataset from BRFSS 2021. Frontiers in Public Health. https://doi.org/10.3389/fpubh.2024.1328353

8. Nguyen, B., & Zhang, Y. (2025). A comparative study of diabetes prediction based on lifestyle factors using BRFSS 2015. arXiv. https://doi.org/10.1101/2025.03.06.25030612

9. Ren, X. (2025). Predictions of diabetes using BRFSS health indicators dataset. In Proc. Intl. Conf. on ML & Automation. https://doi.org/10.54254/2755-2721/32/20230214

10. Islam, M. M. (2025). Explainable machine learning for efficient diabetes classification. Engineering Reports. https://doi.org/10.1002/eng2.13080

11. Khokhar, P. B., Gravino, C., & Palomba, F. (2024). Advances in AI for diabetes prediction: A systematic review. arXiv. https://doi.org/10.1101/2412.14736

12. Ullah, Z., Simonis, P., & others. (2022). Detecting high-risk factors and early diagnosis using BRFSS with SMOTE-ENN. IEEE Access. https://doi.org/10.1109/

13. Fang, M. (2023). Prediabetes and diabetes screening eligibility in U.S. adults: defining cutoffs of fasting plasma glucose and HbA1c. JAMA Network Open, 6(1), e233132.

14. Little, R.J. and Rubin, D.B. (2019) Statistical Analysis with Missing Data. Vol. 793, John Wiley & Sons, Hoboken. https://doi.org/10.1002/9781119482260

15. Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). SMOTE: Synthetic minority over-sampling technique. Journal of Artificial Intelligence Research, 16, 321–357. https://doi.org/10.1613/jair.953

16. Galar, M., Fernández, A., Barrenechea, E., Bustince, H., & Herrera, F. (2012). A review on ensembles for the class imbalance problem: bagging-, boosting-, and hybrid-based approaches. IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews), 42(4), 463–484. https://doi.org/10.1109/TSMCC.2011.2161285

17. Brown, E., & Lu, W. (2024). Survey-Based Machine Learning Models for Early Detection of Diabetes. In International Conference on Wireless Intelligent and Distributed Environment for Communication (pp. 109-122). Cham: Springer Nature Switzerland

18. Gregorutti, B., Michel, B., & Saint-Pierre, P. (2017). Correlation and variable importance in random forests. Statistics and Computing, 27(3), 659-678.

19. Mylona, E., Zaridis, D. I., Kalantzopoulos, C. N., Tachos, N. S., Regge, D., Papanikolaou, N., ... & Fotiadis, D. I. (2024). Optimizing radiomics for prostate cancer diagnosis: Feature selection strategies, machine learning classifiers, and MRI sequences. **Insights into Imaging, 15**, Article 265.

20. Abdar, M., Pourpanah, F., Hussain, S., Rezazadegan, D., Liu, L., Ghavamzadeh, M., ... & Nahavandi, S. (2021). A review of uncertainty quantification in deep learning: Techniques, applications and challenges. Information fusion, 76, 243-297.

21. Bergstra, J., & Bengio, Y. (2012). Random search for hyper-parameter optimization. Journal of Machine Learning Research, 13(Feb), 281–305. https://www.jmlr.org/papers/volume13/bergstra12a/bergstra12a.pdf

22. Nguyen, H. P., Wang, Y., & Tran, T. (2022). Decision tree instability on imbalanced health datasets: Empirical analysis and remedies. Computer Methods and Programs in Biomedicine, 221, 106800. https://doi.org/10.1016/j.cmpb.2022.106800

23. Islam, M., Zhang, Y., & Ren, G. (2020). Ensemble machine learning for early diabetes prediction using lifestyle and health data. Biomedical Engineering Advances, 1, 100273. https://doi.org/10.1016/j.bbe.2020.100273

24. Suryawanshi, P., Tanwar, S., & Kumar, N. (2023). Feature engineering and tree-based models for imbalanced medical data. Information Modeling and Uncertainty, 3, 101002. https://doi.org/10.1016/j.imu.2023.101002

25. Chaki, J., Ganesh, D., & Sen, S. (2021). Predicting diabetes mellitus using SMOTE and ensemble machine learning techniques. Journal of Biomedical Informatics, 116, 103742. https://doi.org/10.1016/j.jbi.2021.103742

26. Dinh, A., Miertschin, S., Young, A., & Mohanty, S. D. (2021). A data-driven approach to predict diabetes using machine learning algorithms. Healthcare Analytics, 1-2, 100016. https://doi.org/10.1016/j.health.2021.100016

27. Zhang, W., et al. (2025). Enhancing diabetes risk prediction through focal active learning strategies combined with machine learning models. PLOS ONE, 20(7), e0327120

28. Asgari, S., Khalili, D., & Hadaegh, F. (2023). External validation of an American risk prediction model for incident type 2 diabetes in an Iranian population. BMC Medical Research Methodology, 23, Article 77.

29. Sadeghi, Z., Alizadehsani, R., Cifci, M. A., et al. (2024). A Brief Review of Explainable Artificial Intelligence in Healthcare. arXiv preprint.