International Conference on Finance, Economics, Management, Accounting and Informatics

"Digital Transformation and Sustainable Business: Challenges and Opportunities for Higher Education Research and Development"

Lifestyle-Based Early Detection of Diabetes: A Machine Learning Approach

Egga Asoka[1], Sulistiyanto[1*], Sony Oktapriandi[1], Yulia Hapsari[1]
[1] *Informatics Management Sriwijaya State Polytechnic*
* *sulistiyanto@polsri.ac.id*

Abstract

Early detection of diabetes based on lifestyle factors plays a vital role in preventing long-term complications. This study proposes a machine learning classification approach using an optimized XGBoost algorithm to identify diabetes status from ten lifestyle-related variables. The original dataset, consisting of three class labels, was simplified into two categories (Non-Diabetic and Diabetic) by merging classes, followed by a Euclidean distance analysis to compute the centroid gap. After feature selection and data scaling, two strategies were applied to address class imbalance: SMOTE and the scale_pos_weight parameter. Experimental results revealed that the scale_pos_weight method achieved superior performance, yielding an accuracy of 0.68 and a recall of 0.73 for the minority class. The model also reached a weighted F1-score of 0.73, indicating high sensitivity toward at-risk individuals. These findings highlight the effectiveness of combining lifestyle-based features and appropriate imbalance handling techniques for robust and reliable early detection of diabetes using machine learning.

Keywords: Diabetes Detection, Lifestyle, XGBoost, Class Imbalance, Scale_Pos_Weight, Machine Learning.

Introduction

Type 2 diabetes mellitus (T2DM) has become a global health crisis, with the number of affected individuals rising from 108 million in 1980 to over 422 million by 2014 (Owess, Owda, Owda, & Massad, 2024). This rapid increase, especially in low- and middle-income countries, has led to millions of deaths and severe complications annually. Early identification of high-risk individuals is crucial for preventing or delaying the onset of T2DM, as timely lifestyle interventions can significantly reduce progression to diabetes (Li et al., 2025). Indeed, growing evidence shows that lifestyle modifications (improving diet, increasing physical activity, weight management, etc.) can prevent or postpone diabetes in those with prediabetes or other risk factors (Almutairi, Abbod, & Hunaiti, 2025). This underscores the importance of developing screening tools that leverage lifestyle factors for early detection.

Well-established risk factors for T2DM include an unhealthy diet, physical inactivity, obesity, and smoking. In addition to these traditional lifestyle risks, researchers have highlighted the role of sleep and fatigue in diabetes risk. Fatigue and poor sleep quality are among the most common complaints of people with diabetes, often appearing as early signs of the disease (Lien, Jiang, Tsai, Hwang, & Lin, 2020). These factors are closely related to glycemic control and the development of complications, suggesting that metrics of sleep quality and chronic fatigue could serve as valuable predictors in diabetes risk models. Harnessing such non-invasive lifestyle indicators is especially attractive for low-cost, population-level screening. Traditional screening methods typically rely on biochemical tests (e.g. blood glucose levels), which may be impractical in resource-limited settings or for large-scale prevention programs (Birk et al., 2021). By contrast, lifestyle-based screening using data on diet, exercise, sleep, and other habits can be done via questionnaires or wearable sensors, offering a convenient alternative for identifying at risk individuals.

Machine learning (ML) techniques have emerged as powerful tools to detect complex patterns in health data and improve the accuracy of disease risk prediction. In the context of diabetes, numerous studies have successfully applied ML algorithms including Support Vector Machines (SVM), neural networks, decision trees, and ensemble methods to predict diabetes onset or diagnose undetected cases (Almutairi et al., 2025; Li et al., 2025). These data-driven models can incorporate multidimensional inputs and capture non-linear relationships between lifestyle factors and disease risk, potentially outperforming traditional risk scores. SVM in particular has been noted for its strong classification performance in medical applications, often achieving high sensitivity in identifying positive cases (Riveros Perez & Avella-Molano, 2025). Given that

T2DM is largely a lifestyle-dependent disease (especially for type 2, which is insulin-independent but strongly influenced by behavioral factors)(Ganie & Malik, 2022), applying ML to lifestyle data is a promising approach for early detection.

This study presents an experimental machine learning framework for the early detection of type 2 diabetes using a set of ten lifestyle and health-related variables: body mass index (BMI), smoking status, physical activity, fruit and vegetable consumption, alcohol consumption, general health status, physical and mental health, and difficulty walking. An optimized Extreme Gradient Boosting (XGBoost) classifier was developed to classify individuals as diabetic or non-diabetic. To address class imbalance in the dataset, two strategies, Synthetic Minority Oversampling Technique (SMOTE) and the scale_pos_weight parameter were compared. The dataset was preprocessed through feature scaling and centroid-based class simplification, followed by model training and hyperparameter tuning. The proposed model demonstrated promising results, particularly when using scale_pos_weight, highlighting the feasibility of lifestyle-based, non-invasive screening tools for diabetes risk assessment, especially in resource-limited settings.

Literature Review

A. Lifestyle Factors as Predictors of Diabetes

Epidemiological research firmly establishes that lifestyle factors play a pivotal role in diabetes risk (Almutairi et al., 2025; Li et al., 2025). Unhealthy dietary patterns, sedentary behavior, smoking, and alcohol consumption have been consistently associated with increased diabetes risk. For instance, (Almutairi et al., 2025) modeled diabetes trends in Saudi Arabia and identified smoking, obesity, and low physical activity as the most influential risk factors for the disease. Unhealthy diet patterns are similarly critical; (Birk et al., 2021) demonstrated that a comprehensive diet quality metric can help flag individuals with prediabetes. In their study, a novel Global Diet Quality Score (GDQS) was computed from food-frequency questionnaires in rural India, and when combined with simple covariates like age and tobacco use, it enabled a machine-learning model to classify prediabetes with an area under the ROC curve (AUC) of about 0.72. This finding is notable because it validates that diet-focused lifestyle questionnaires, without any laboratory tests, can achieve reasonable accuracy in identifying high-risk individuals. Additionally, the authors stressed that diet quality is a strong risk factor for T2DM development and that low-cost dietary assessment tools could be leveraged for diabetes screening in resource-limited settings.

Using a hybrid approach that combined statistical analysis and feature selection with machine learning, Lien *et al.* built an SVM-based model to predict diabetic nephropathy presence from ten selected variables (including fatigue and sleep measures) with about 74% accuracy. Although this study focused on a complication of diabetes, its implications extend to early detection: it highlights fatigue and sleep disturbance as early indicators correlated with glycemic dysregulation. These results support incorporating sleep quality metrics (e.g. insomnia, sleep duration, snoring frequency) and fatigue scores into diabetes risk prediction models. Poor sleep has been linked in other research to increased insulin resistance and higher incidence of T2DM, so it is biologically plausible that sleep-related variables improve predictive performance (Lien et al., 2020).

B. Machine Learning Approaches for Diabetes Prediction

A variety of ML models have been applied to predict diabetes using non-invasive health data. Traditional statistical models (like logistic regression) are often used as baselines, but more flexible supervised learning methods can capture complex interactions among lifestyle factors. Support Vector Machine (SVM) classifiers, in particular, have shown promise in this domain. (Riveros Perez & Avella-Molano, 2025) conducted a large study using U.S. NHANES data (2007–2018) to predict prevalent type 2 diabetes based solely on self-reported lifestyle and anthropometric variables. They compared five algorithms (logistic regression, SVM, random forest, XGBoost, CatBoost) and found that ensemble trees achieved the highest overall accuracy (~85% with AUC ≈0.82). XGBoost emerged as one of the top-performing models, achieving an AUC of 0.82 and high specificity. Similarly, Owess et al. (2024) applied ML on a Palestinian population dataset and showed that integrating age, BMI, diet, and physical activity into a Random Forest model achieved up to 98.4% accuracy for detecting hyperglycemia.

Several other studies echo the efficacy of ML using lifestyle inputs. (Owess et al., 2024) developed ML classification models on a dataset of Palestinian adults from a STEPS non-communicable disease risk-factor survey. Their model incorporated factors like age, body mass index (BMI), diet (eating habits), physical activity level, and comorbid conditions to predict "raised blood sugar," a proxy for undiagnosed diabetes or prediabetes. Among various algorithms tested, an ensemble based on random forests achieved an impressive 98.4% accuracy in detecting individuals with hyperglycemia. This extremely high accuracy may partly reflect the inclusion of a biochemical measure (fasting blood sugar) in the feature set, but it nevertheless demonstrates how combining lifestyle and basic clinical data with ML can yield highly accurate screening tools.

In a similar vein, (Ganie & Malik, 2022) proposed an ensemble learning framework for early T2DM prediction using lifestyle indicators in an Indian population. By applying techniques like bagging, boosting, and feature engineering on a balanced dataset, their best model (a bagged decision tree) achieved over 99% accuracy and 95.8% recall (sensitivity) in classifying diabetes vs. non-diabetes. Such performance, approaching perfection, suggests that in certain datasets the patterns in lifestyle and personal health data are sufficiently distinct to allow near-complete discrimination of diabetic vs. healthy individuals. However, it's worth noting that extremely high accuracies often indicate either a very diagnostically powerful feature (for example, a glucose measurement) or potential overfitting; thus, results like 99% should be interpreted with caution pending external validation.

These findings are consistent with prior research on metabolic syndrome detection using hybrid machine learning models, particularly the study by author (Asoka, 2025), which combined K-Means clustering and XGBoost to improve hypertension classification. The study highlighted that key metabolic indicators such as systolic and diastolic blood pressure, BMI, and fasting glucose are effective in early risk stratification for diabetes and hypertension. Although the previous work emphasized clinical features, it underscores the growing relevance of integrating behavioral and metabolic data to improve machine learning-based screening systems (Asoka, 2025). The current study expands this direction by focusing on lifestyle features alone such as sleep quality, fatigue, diet quality, and activity level thus exploring a more accessible and non-invasive predictive approach.

The present study builds on this foundation by applying an optimized XGBoost classifier trained on ten lifestyle and health indicators: BMI, smoking, physical activity, fruit and vegetable consumption, alcohol use, general health status, physical health days, mental health days, and difficulty walking. Special emphasis is placed on addressing class imbalance often overlooked in prior work through comparative analysis of SMOTE and scale_pos_weight. The model is further enhanced via hyperparameter tuning to achieve robust, real-world performance.

C. Towards Lifestyle-Based Early Detection

The collective findings from recent literature strongly support the feasibility of early diabetes detection grounded in lifestyle data. Multiple independent studies have demonstrated that even without clinical lab tests, individuals at risk for diabetes can be identified with moderate to high accuracy by analyzing patterns in their daily habits and symptoms. High-impact journals in digital health and public health (e.g. *BMJ Open*, *BMC Public Health*, *PLOS ONE*, *Frontiers*) have published works validating diet-only or lifestyle-only models (Birk et al., 2021; Riveros Perez & Avella-Molano, 2025). Key lifestyle features repeatedly identified include dietary quality scores, physical activity frequency or duration, smoking status, alcohol consumption, sleep duration/quality, and self-reported fatigue or energy levels(Lien et al., 2020).

Many of these factors are interrelated (for instance, poor sleep can lead to fatigue and overeating), and ML algorithms are well-suited to untangling such interdependencies. Ensemble models combining several algorithms (or combining lifestyle indices with basic health indicators like BMI or blood pressure) often achieve the best results, as seen in various studies(Ganie & Malik, 2022). Importantly, the use of interpretable ML techniques (such as SHAP values or rule-based models) is gaining traction, which helps in quantifying the contribution of each lifestyle factor to the prediction. This is valuable for public health messaging, for example, showing that in a given model, physical inactivity and high-fat diet were the top contributors to predicted risk can reinforce those targets for intervention.

In conclusion, the literature indicates a clear trend towards leveraging lifestyle data for diabetes risk prediction, powered by machine learning methods including SVM. While results vary across populations, most

studies report that adding lifestyle factors improves predictive power, and in some cases, lifestyle-based models can approach the accuracy of models that include clinical data. This opens up opportunities for non-invasive, cost-effective early screening programs. The challenge moving forward is to integrate findings from these studies to design an optimized prediction model that generalizes well. Such a model that the focus of our ongoing work would take as input an individual's lifestyle profile (diet, exercise, sleep, smoking, etc.) and output an estimate of diabetes risk. The ultimate aim is to deploy it as an early warning system: identifying high-risk persons who can then be guided through preventive measures or confirmatory diagnostics. The research reviewed here provides a strong foundation for this approach, showing that "lifestyle-based early detection of diabetes" is not only plausible but already yielding encouraging results in diverse settings.

Despite promising results in prior studies, most machine learning models for diabetes detection still rely heavily on biochemical or clinical features. Few have explored whether lifestyle indicators alone, without blood-based biomarkers can provide sufficient predictive accuracy. This study aims to fill this gap by evaluating the feasibility of using a minimal set of lifestyle features in a support vector machine model for early diabetes detection.

Methods

This study employs a supervised machine learning approach to predict diabetes status using only lifestyle and health-related indicators. The methodological framework consists of five key stages: dataset acquisition, data preprocessing, class imbalance handling, model development, and evaluation, as illustrated in Figure 1.

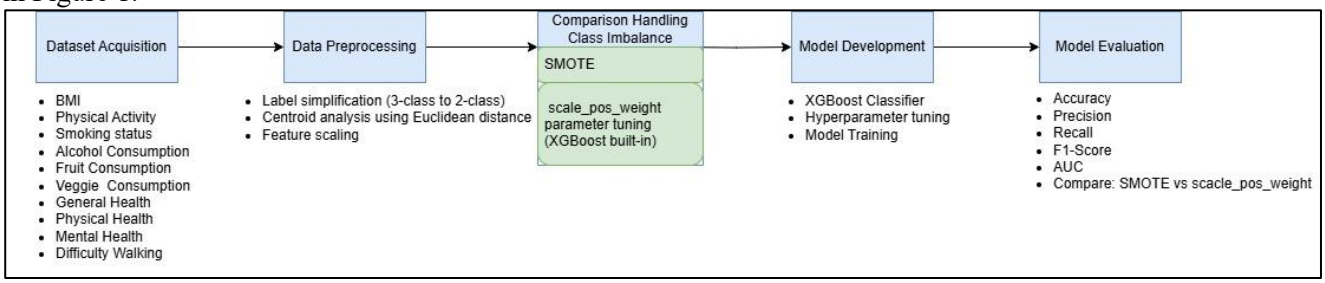

Figure 1. Technical Approach

The dataset includes ten lifestyle and health-related variables: Body Mass Index (BMI), smoking status, physical activity, fruit consumption, vegetable consumption, alcohol consumption, general health status, physical health days, mental health days, and difficulty walking. These attributes were selected based on their relevance to behavioral and metabolic risk, and their potential for non-invasive, large-scale screening.

In the data preprocessing phase, the original target variable consisted of three classes (healthy, prediabetes, and diabetes). To simplify the classification task and enhance detection sensitivity, the labels were restructured into a binary format—"Not Diabetes" and "Diabetes"—by merging the prediabetes and diabetes classes. This relabeling was supported by a centroid distance analysis using Euclidean metrics, which revealed a closer proximity between prediabetes and diabetes groups.

All numerical features were normalized using standard scaling to ensure consistent value ranges and support optimal model convergence. No dimensionality reduction was applied, as all features were retained for interpretability and completeness.

To mitigate the class imbalance observed in the dataset, two techniques were employed for comparative purposes. The first was Synthetic Minority Oversampling Technique (SMOTE), which synthetically generates new instances of the minority class. The second was the scale_pos_weight parameter available within the XGBoost classifier, which adjusts the contribution of minority samples during training. Both approaches were implemented independently to examine their effect on model performance.

In the model development phase, an Extreme Gradient Boosting (XGBoost) classifier was constructed due to its robustness in handling structured tabular data and imbalanced distributions. Hyperparameter optimization was performed using randomized search combined with cross-validation, targeting key parameters such as learning rate, maximum depth, number of estimators, and subsampling ratio. After tuning,

the model was trained using the preprocessed and balanced datasets prepared via both SMOTE and scale_pos_weight.

Subsequently, the trained models were prepared for evaluation using classification performance metrics. However, performance results and comparisons are discussed in the following section.

Results and Discussion

This study evaluated two approaches for handling class imbalance in lifestyle-based diabetes prediction using an optimized XGBoost classifier: Synthetic Minority Oversampling Technique (SMOTE) and the built-in scale_pos_weight parameter. Both models were trained and tuned under identical experimental settings to ensure fairness in comparison. The primary aim was to assess which technique better supports the early identification of high-risk individuals in an imbalanced dataset dominated by non-diabetic cases.

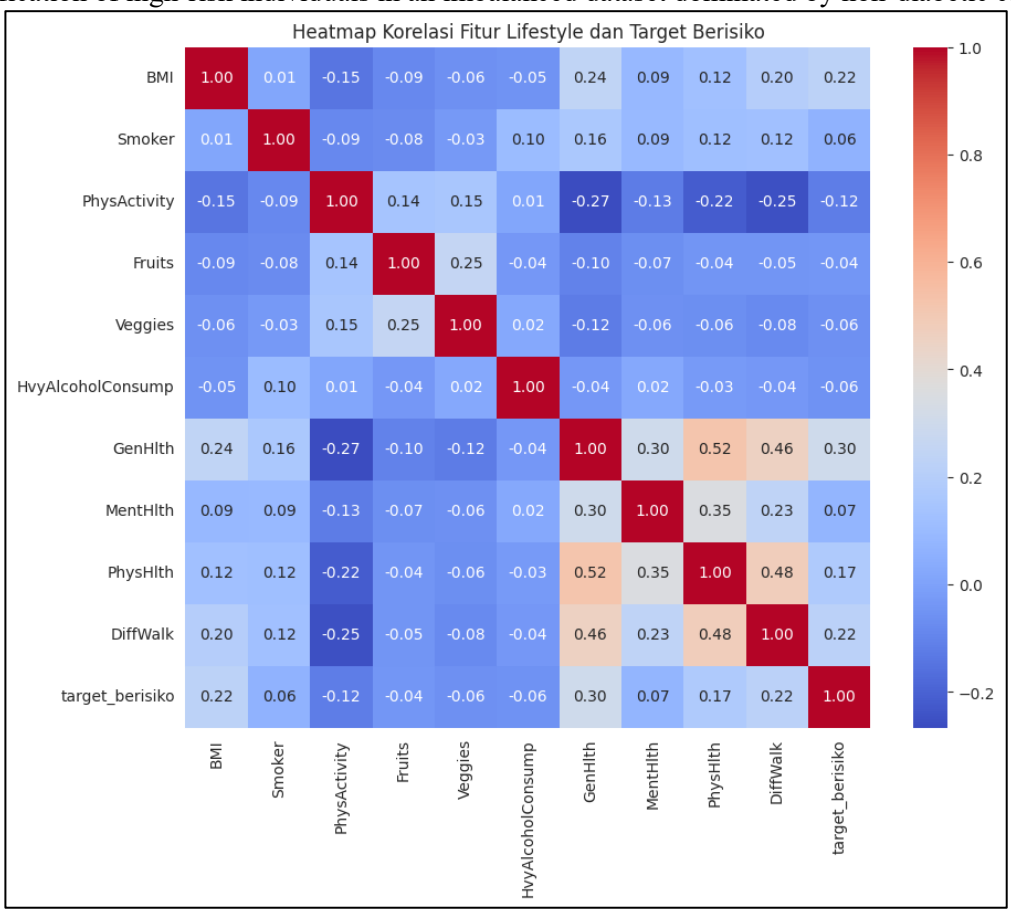

Figure 1. Heatmap Correlation

To better understand the relationships between the ten lifestyle-related features and the diabetes label, a heatmap was constructed in Figure 1. The correlation heatmap reveals that features such as general health, physical health, difficulty walking, and BMI had relatively stronger positive associations with diabetes risk. In contrast, fruit and vegetable consumption showed minimal correlation. These findings reaffirm existing literature indicating that perceived health status and physical limitations are valuable predictors in non-invasive screening contexts.

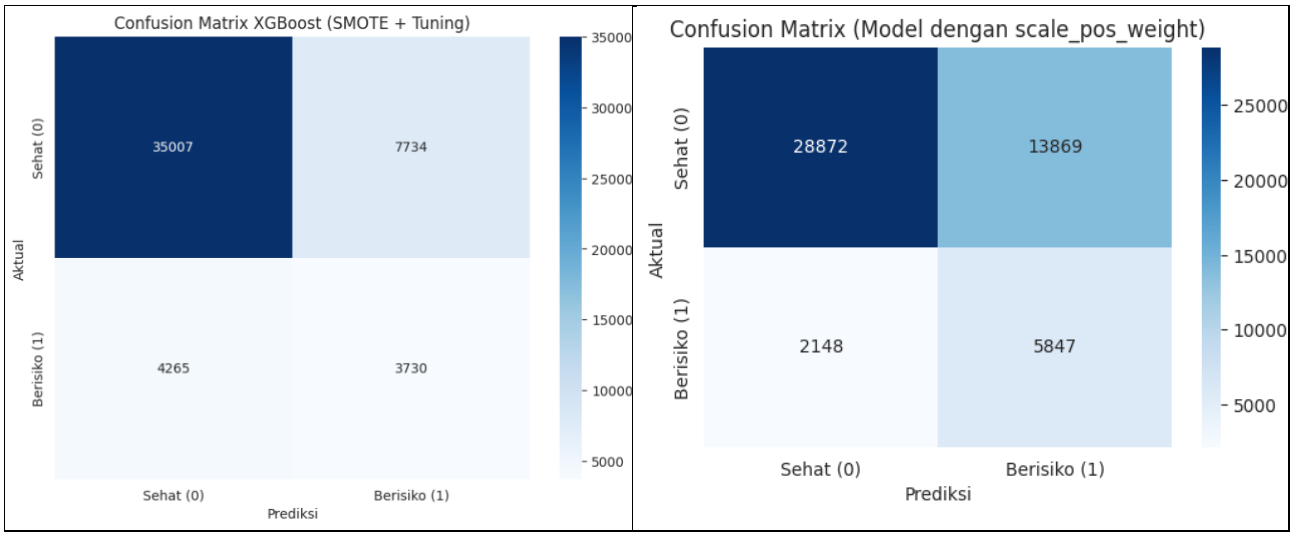

Figure 2. Confusion Matrix of XGBoost Model using SMOTE and scale_pos_weight

Next, both models were evaluated using standard classification metrics including accuracy, precision, recall, and F1-score. The confusion matrix for the model using SMOTE in Figure 2 showed that it correctly predicted 3,730 at-risk individuals but misclassified 4,265 positive cases as healthy. Although the model achieved a high accuracy of 0.85 and a weighted F1-score of 0.80, it demonstrated poor sensitivity to the positive class, with a recall of only 0.11 and an F1-score of 0.18 for class 1. These results indicate that SMOTE, despite boosting general accuracy, failed to adequately capture the minority class, which is problematic in preventive health applications where missing high-risk individuals can have serious consequences.

Conversely, the model trained using scale_pos_weight, Figure 3 captured 5,847 true positives and only 2,148 false negatives, resulting in a recall of 0.73 and a positive class F1-score of 0.42. While its overall accuracy was lower at 0.68 and precision for class 1 was modest (0.30), this trade-off was justified by its stronger detection capability. In the context of public health, such a model is preferable because it prioritizes identifying at-risk individuals who can then be referred for confirmatory testing and intervention.

Table 1. Comparison of SMOTE and scale_pos_weight Models

Metric	XGBoost + SMOTE	XGBoost + scale_pos_weight
Accuracy	0.85	0.68
Precision (Class 1)	0.55	0.30
Recall (Class 1)	0.11	0.73
F1-Score (Class 1)	0.18	0.42
Macro F1-Score	0.55	0.60
Weighted F1-Score	0.80	0.73

A summary of the comparison between both models is presented in Table 1. It is evident that although SMOTE appears more accurate at the surface level, the scale_pos_weight strategy demonstrates superior performance in classifying the positive class, highlighting its value in screening-oriented scenarios.

Taken together, these results demonstrate that the use of scale_pos_weight offers a more balanced and clinically meaningful approach for early diabetes detection using lifestyle data. This method proved more sensitive to at-risk individuals without relying on oversampling, making it better suited for deployment in real-world, large-scale, and resource-constrained screening systems.

Conclusion

This study demonstrates the feasibility of using lifestyle and general health-related indicators for early diabetes detection through a machine learning framework based on XGBoost. By incorporating ten non-invasive features including BMI, physical activity, diet, general health, and mobility limitations, the model successfully identified individuals at risk for diabetes without relying on clinical or biochemical tests.

A key focus of the study was evaluating two different strategies to address class imbalance: SMOTE and the scale_pos_weight parameter. While SMOTE yielded a higher overall accuracy (0.85), it significantly underperformed in detecting the minority class, with a recall of only 0.11. In contrast, the model using scale_pos_weight achieved a recall of 0.73 and a better F1-score for the positive class, making it more effective in capturing high-risk individuals.

These results highlight that in screening-oriented applications where sensitivity to positive cases is critical, scale_pos_weight offers a more balanced and clinically useful approach. The study affirms that lifestyle-based data, when coupled with well-tuned ensemble models and appropriate imbalance handling, can serve as a low-cost, scalable alternative for diabetes risk screening in public health settings.

Future work may explore integrating explainable AI techniques to improve interpretability for clinical users and expand the feature set with longitudinal or sensor-based lifestyle data for enhanced precision.

References

Abbas, H. T., Alic, L., Erraguntla, M., Ji, J. X., Abdul-Ghani, M., Abbasi, Q. H., & Qaraqe, M. K. (2019). Predicting long-term type 2 diabetes with support vector machine using oral glucose tolerance test. *PLoS ONE*, *14*(12), 1–11. https://doi.org/10.1371/journal.pone.0219636

Almutairi, E., Abbod, M., & Hunaiti, Z. (2025). Prediction of Diabetes Using Statistical and Machine Learning Modelling Techniques. *Algorithms*, *18*(3). https://doi.org/10.3390/a18030145

Asoka, E. (2025). Machine Learning Models for Metabolic Syndrome Identification with Explainable AI, *6*(3), 1159–1172.

Birk, N., Matsuzaki, M., Fung, T. T., Li, Y., Batis, C., Stampfer, M. J., … Lake, E. (2021). Exploration of Machine Learning and Statistical Techniques in Development of a Low-Cost Screening Method Featuring the Global Diet Quality Score for Detecting Prediabetes in Rural India. *Journal of Nutrition*, *151*, 110S-118S. https://doi.org/10.1093/jn/nxab281

Choi, S. B., Kim, W. J., Yoo, T. K., Park, J. S., Chung, J. W., Lee, Y. H., … Kim, D. W. (2014). Screening for prediabetes using machine learning models. *Computational and Mathematical Methods in Medicine*, *2014*. https://doi.org/10.1155/2014/618976

Ganie, S. M., & Malik, M. B. (2022). An ensemble Machine Learning approach for predicting Type-II diabetes mellitus based on lifestyle indicators. *Healthcare Analytics*, *2*(November), 100092. https://doi.org/10.1016/j.health.2022.100092

Li, X., Ding, F., Zhang, L., Zhao, S., Hu, Z., Ma, Z., … Zhao, Y. (2025). Interpretable machine learning method to predict the risk of pre-diabetes using a national-wide cross-sectional data: evidence from CHNS. *BMC Public Health*, *25*(1). https://doi.org/10.1186/s12889-025-22419-7

Lien, A. S. Y., Jiang, Y. Der, Tsai, J. L., Hwang, J. S., & Lin, W. C. (2020). Prediction of diabetic nephropathy from the relationship between fatigue, sleep and quality of life. *Applied Sciences (Switzerland)*, *10*(9). https://doi.org/10.3390/app10093282

Owess, M. M., Owda, A. Y., Owda, M., & Massad, S. (2024). Supervised Machine Learning-Based Models for Predicting Raised Blood Sugar. *International Journal of Environmental Research and Public Health*, *21*(7). https://doi.org/10.3390/ijerph21070840

Riveros Perez, E., & Avella-Molano, B. (2025). Learning from the machine: is diabetes in adults predicted by lifestyle variables? A retrospective predictive modelling study of NHANES 2007-2018. *BMJ Open*, *15*(3), 1–10. https://doi.org/10.1136/bmjopen-2024-096595