Predicting Type-2 Diabetes Risk among Adults in Ethiopia: A Machine Learning Approach

Yonas Sileshi
HiLCoE School of Computer Science and Technology, Addis Ababa, Ethiopia
yonassb@hilcoeschool.com

Eyob Nigussie
Addis Ababa University, Addis Ababa, Ethiopia
eyob.alemu@aau.edu.et

Abstract

This study aims to develop a machine learning model to predict the risk of Type 2 diabetes (T2D) among Ethiopian adults, using demographic, behavioral, physical, and biochemical data. The research responds to the increasing diabetes burden in Ethiopia, utilizing data from the EPHI NCD STEPS survey, which includes 9,800 instances. Seven machine learning models—Decision Tree, Random Forest, Logistic Regression, SVM, KNN, ANN, and XGBoost were employed to assess predictive accuracy, incorporating appropriate preprocessing, class balancing, and evaluation techniques. The Random Forest model consistently outperformed others, achieving an accuracy of 87.79% without resampling. When using SMOTE and 10-fold cross-validation, its accuracy peaked at 93.70%, with precision and recall also at 93.69%, and an F1-score of 93.69%. Furthermore, it yielded an impressive ROC-AUC score of 98.92%, establishing it as the most effective classifier. The findings demonstrate the potential of machine learning to enhance early diabetes risk detection in Ethiopia, highlighting key lifestyle predictors such as alcohol use, fruit and vegetable intake, waist-hip ratio, and BMI. The study advocates for the integration of machine learning-driven screening tools into national health programs, proposing a scalable, cost-effective strategy for the early diagnosis and prevention of Type 2 diabetes.

Keywords: Machine Learning, Type 2 Diabetes, Classification, Imbalanced Data, Chronic Disease, Diabetes Risk

1. Introduction

Type 2 Diabetes Mellitus (T2D) is one of the fastest-growing global public health challenges of the 21st century and a major contributor to the burden of non-communicable diseases (NCDs). It is characterized by insulin resistance and impaired insulin secretion, leading to chronic hyperglycemia that can result in severe complications such as cardiovascular disease, kidney failure, neuropathy, and blindness if not properly managed [1]. According to the International Diabetes Federation (IDF), approximately 588.7 million adults aged 20–79 were living with diabetes in 2024, a figure projected to rise to 853 million by 2050 [2]. Type 2 diabetes accounts for nearly 90% of all diabetes cases worldwide [2].

Ethiopia is experiencing a rapidly increasing diabetes burden, driven by factors such as rapid urbanization, lifestyle changes, and limited access to preventive healthcare. Recent estimates indicate that about 3.2 million Ethiopians are affected by diabetes, with a large proportion remaining undiagnosed [3]. The IDF reports a national adult diabetes prevalence of 4.4% [2], while the WHO estimates a prevalence of 3.8%, reflecting a rising trend [4]. In response, Ethiopia has adopted World Health Organization strategies for NCD prevention, implementing policies and guidelines targeting shared behavioral, metabolic, and environmental risk factors [5], [6]. Despite these efforts, early identification of individuals at high risk remains limited.

The motivation for this study stems from the absence of robust, data-driven predictive tools for early T2D risk assessment in Ethiopia. While previous studies have documented prevalence and associated risk factors, few have integrated demographic, behavioral, and clinical variables into machine learning(ML) based predictive frameworks using nationally representative data. Early risk prediction is critical for enabling timely intervention, reducing complications, and lowering long-term healthcare costs.

Accordingly, this study aims to develop and evaluate ML models for predicting T2D risk among Ethiopian adults. The specific objectives are to identify key risk factors, compare multiple ML algorithms based on accuracy, sensitivity, and specificity, and determine the most reliable models for practical application. The study also examines how lifestyle, demographic, and clinical factors jointly influence diabetes risk.

This research utilizes data from a national NCD prevalence survey of Ethiopian adults, incorporating demographic characteristics, lifestyle behaviors, and selected clinical indicators such as body mass index, waist-to-hip ratio, cholesterol level, and hypertension. The scope excludes genetic factors, detailed family history, longitudinal follow-up, and intervention-based analyses. Nevertheless, the findings provide important insights into the potential of ML-based risk prediction to support early detection and evidence-based decision-making within the Ethiopian healthcare system.

2. Related Work

Recent years have seen growing adoption of machine learning techniques for diabetes risk prediction and clinical decision support. In Ethiopia, Oumer Abdulkadir Ebrahim and Getachew Derbew applied supervised ML models in the Afar region and reported high accuracy using Random Forest (93.8%) [7]. While this study demonstrates the feasibility of ML-based diabetes prediction in an Ethiopian context, its limited sample size (2,239 cases) and regional scope restrict generalizability to the national population.

Studies from other regions, particularly South Asia, further confirm the effectiveness of ML approaches. Research from Bangladesh reported strong performance using Random Forest, Logistic Regression, and Support Vector Machines, with Random Forest achieving up to 94.6% accuracy [8], while logistic regression showed superior performance when lifestyle and dietary factors were emphasized [9]. However, these studies rely on region-specific datasets and sometimes include redundant features (e.g., height, weight, and BMI simultaneously), which may introduce multicollinearity and reduce interpretability.

Widely used benchmark datasets such as the Pima Indian Diabetes Database have also been employed to compare ML algorithms, including Decision Tree, Random Forest, Support Vector Machine, and K-Nearest Neighbors [10]. Although these studies provide useful algorithmic benchmarks, they often exclude key lifestyle and hereditary risk factors such as smoking, alcohol use, and family history. Similar studies from Nigeria and other regions have reported very high accuracy values, occasionally exceeding 95%, but concerns regarding data quality, feature redundancy, and unrealistic attribute distributions limit the reliability of these findings.

More recent work has attempted to improve predictive performance by addressing class imbalance and feature selection. For example, Ahmed I. ElSeddawy et al. [11] applied oversampling and undersampling techniques such as SMOTE and Tomek links, while Selim Buyrukoglu and Ayhan Akbas [12] proposed hybrid feature selection combined with ANN classifiers. Although these approaches achieved impressive results, they often relied on narrow datasets and primarily framed diabetes prediction as a binary classification problem, neglecting the prediabetic stage that is critical for early intervention.

Overall, existing studies reveal several gaps: limited focus on African populations, insufficient modeling of prediabetes, inadequate handling of class imbalance, and a lack of integrated behavioral, demographic, and biochemical features. This study addresses these gaps by developing an ML-based T2D risk prediction model tailored to Ethiopian adults, incorporating diverse risk factors and explicitly handling class imbalance to support early and accurate detection.

3. Methodology

This study employed an experimental research design to develop and evaluate machine learning (ML) models for the prediction of Type 2 Diabetes (T2D) risk. The data were obtained from the Ethiopian Public Health Institute (EPHI) via the World Health Organization (WHO)–sanctioned Non-Communicable Disease (NCD) STEPS Survey, which was executed across nine regional states and two administrative cities. The initial dataset encompassed 10,029 records with 433 variables. Subsequent to the processes of data cleansing and preprocessing, 9,800 valid instances were preserved for analytical purposes.

The preprocessing phase entailed the elimination of non-informative records and the exclusion of attributes exhibiting over 20% missing values. Missing numerical values were addressed through the application of the K-Nearest Neighbors (KNN) imputation method, while categorical variables were imputed utilizing mode values. Feature engineering was employed to extract clinically significant variables, which included Hypertension, Body Mass Index (BMI), Waist-to-Hip Ratio (WHR), and the target variable, Diabetes Risk. The Diabetes Risk label was derived from fasting blood glucose levels in accordance with the diagnostic criteria established by the American Diabetes Association (ADA) [13].

To validate the reliability of the generated labels, a Fleiss' Kappa analysis was performed on a sample of 100 randomly chosen records that were annotated by three domain experts, yielding a perfect agreement coefficient ($\kappa = 1.0$, $p < 0.001$). Feature selection was executed through Spearman's rank correlation analysis to identify pertinent predictors and to achieve dimensionality reduction. Features exhibiting weak or negative correlations with the target variable were systematically excluded from consideration. Despite the clinical relevance of sex and family history of diabetes, these variables were omitted due to their weak associations within the context of this dataset, which aligns with findings presented in the EPHI national report [14].

The comprehensive final feature set is encapsulated in Table 1. All experimental procedures were conducted utilizing Python through the Google Colab platform, employing Pandas and NumPy for data processing, scikit-learn for modeling and evaluation, TensorFlow and Keras for constructing neural networks, and XGBoost for implementing gradient boosting. Visualization and exploratory data analyses were carried out utilizing Matplotlib and Seaborn.

Table 1: Selected Features

Features Code	**Name of Features**	**Feature Type**	**Range/Description**
Label Class	Diabetes_Risk	Categorical	(Levels: 012)
b5	Fasting blood glucose	Numerical	Continuous (transformed)
a1	Alcohol use	Categorical	(Levels:12)
agecat	Age category	Categorical	(Levels: 15-29 30-44 45-59 60-69)
c6	Ethnic background	Categorical	(Levels: 123...11 88)
d1	Fruit usage	Categorical	(Levels: 012...7 77)
d3	Vegetable usage	Categorical	(Levels: 012...7 77)
b17	HDL Cholesterol	Numerical	Continuous (transformed)
WHR	Waist-to-hip ratio	Numerical	Continuous (transformed)
hypertension	Hypertension status	Categorical	(Levels: 0 1)
c7	Marital status	Categorical	(Levels: 12...588)
BMI	Body Mass Index	Numerical	Continuous (transformed)

4.Experimentation and Results

4.1 Experiments

To evaluate the predictive performance of machine learning models for Type 2 Diabetes (T2D) risk, a total of Eighty-four base experiments were conducted using seven classifiers (DT, LR, RF, KNN, XGB, SVM, ANN), three resampling methods (none, SMOTE, ADASYN), two train–test splits (80/20, 70/30), and two feature sets, with 10-fold cross-validation yielding 840 validation runs.

Two scenarios: (i) include blood glucose (B5) and (ii) exclude B5, using only non-invasive features. Performance metrics: accuracy, precision, recall, F1, ROC-AUC (one-vs-rest for multiclass) **[15],[16**], and confusion matrices. The data were highly imbalanced (8,592 healthy, 939 pre-diabetic, 269 diabetic); SMOTE and ADASYN were applied only to training sets. The class distributions before and after resampling are summarized in Table 2.

Table 2: Training Sample Distribution Before and After Resampling

Diabetes Risk	**Initial Data (100%)**	**No Resampling (80%)**	**No Resampling (70%)**	**SMOTE (80%)**	**SMOTE (70%)**	**ADASYN (80%)**	**ADASYN (70%)**
Healthy	8592	6,874	6,015	6,874	6,015	6,874	6,015
Pre-Diabetes	939	751	657	6,874	6,015	6,981	6,135
Diabetes	269	215	188	6,874	6,015	6,911	6,059
Total	**9800**	**7,840**	**6,860**	**20,622**	**18,045**	**20,766**	**18,209**

4.1.1 Experiment 1: Derived Label Class including Blood Glucose Level (B5)

This experiment evaluated model performance when the feature used to derive the Diabetes Risk label (B5) was included. Random Forest (RF) consistently outperformed other models across most configurations. The highest test accuracy (99.46%) was achieved by RF without resampling using a 70/30 split. As shown in Table 3, Cross-validated results further confirmed the robustness of RF and XGBoost under SMOTE and ADASYN, achieving near-perfect ROC-AUC values. These findings indicate that including blood glucose leads to near-ceiling performance, as expected for a feature directly related to the outcome.

Table 3: Top 3 Models from Cross-Validated Results

Rank	**Model**	**Strategy**	CV Accuracy	CV Precision	CV Recall	**CV F1**	CV-ROC AUC
1	RF	SMOTE(80/20)	0.9925	0.9926	0.9925	0.9934	0.9998
2	XGBoost	SMOTE(80/20)	0.9937	0.9937	0.9937	0.9937	0.9996

3	RF	ADASYN(80/20)	0.9934	0.9934	0.9934	0.9934	0.9999

4.1.2. Experiment 2: Derived Label Class excluding Blood Glucose Level (B5)

To assess predictive performance using only non-invasive features, B5 was excluded. Overall performance declined across all models, highlighting the importance of blood glucose in T2D prediction. Without resampling, Random Forest achieved the best results, with a maximum accuracy of 87.79% and ROC-AUC of 0.6847. After applying SMOTE and ADASYN, test-set accuracy slightly decreased; however, cross-validation performance improved substantially. As shown in Table 4, Random Forest with SMOTE achieved the highest cross-validated accuracy (93.70%), F1-score (93.69%), and ROC-AUC (0.9892), demonstrating that resampling techniques significantly enhance generalization when blood glucose is excluded.

Table 4: Top 3 Cross-Validated Model Performance Results

Split	Strategy	Model	Accuracy	Precision	Recall	F1 Score	ROC AUC
70/30	SMOTE	Random Forest	**0.9370**	0.9369	0.9370	0.9369	0.9892
70/30	ADASYN	Random Forest	0.9370	0.9369	0.9370	0.9368	0.9885
80/20	SMOTE	Random Forest	0.9357	0.9355	0.9357	0.9354	0.9887

5. Discussion

This study evaluated machine learning (ML) approaches for predicting Type 2 Diabetes (T2D) risk among Ethiopian adults using two experimental scenarios. Although including blood glucose level (B5) resulted in near-perfect performance, it introduced data leakage because B5 directly defines the outcome label. Therefore, the discussion focuses on Experiment 2, which excludes B5 and provides a more realistic assessment of early T2D risk prediction using non-invasive features.

The main objective of this study was to answer the research questions outlined in the first Section. This section explains how each question was addressed in this study.

RQ1: Which features are the most important in predicting type 2 diabetes in Ethiopian adults, and how do they jointly impact the performance of the prediction model?

Feature importance analysis using the best-performing Random Forest (RF) model (Figure 1) identified BMI, waist-to-hip ratio (WHR), alcohol consumption, and fruit and vegetable intake as the strongest predictors. Demographic and clinical factors such as age, marital status, HDL cholesterol, and ethnic background also contributed meaningfully. These findings confirm that T2D risk arises from a combination of behavioral, socioeconomic, and clinical factors, highlighting the importance of integrated prevention strategies targeting modifiable behaviors alongside broader social determinants of health.

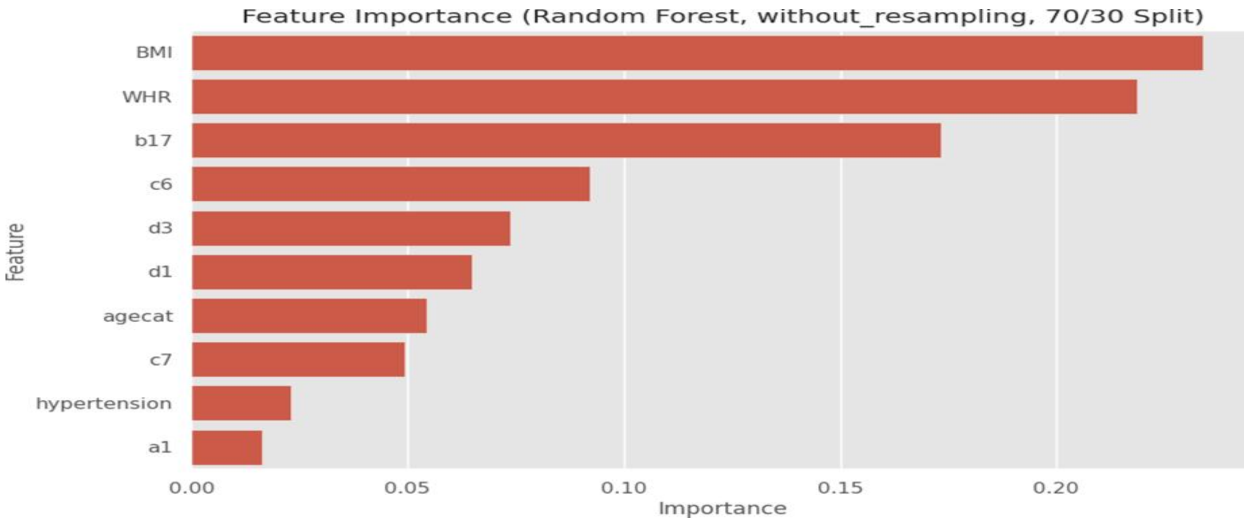

Figure 1: Feature Importance

RQ2: How do machine learning models compare to predict the risk of type 2 diabetes, and which model performs best in key evaluation metrics?

Across all configurations, Random Forest consistently outperformed other models. Without resampling, RF achieved the highest accuracy (87.79%) and F1-score (82.95%) on the 70/30 split, while SVM and linear models showed weaker discrimination despite similar accuracy. After applying SMOTE and ADASYN, cross-validated performance improved substantially, with RF + SMOTE achieving 93.70% accuracy and a ROC-AUC of 0.9892 (Figure 2).

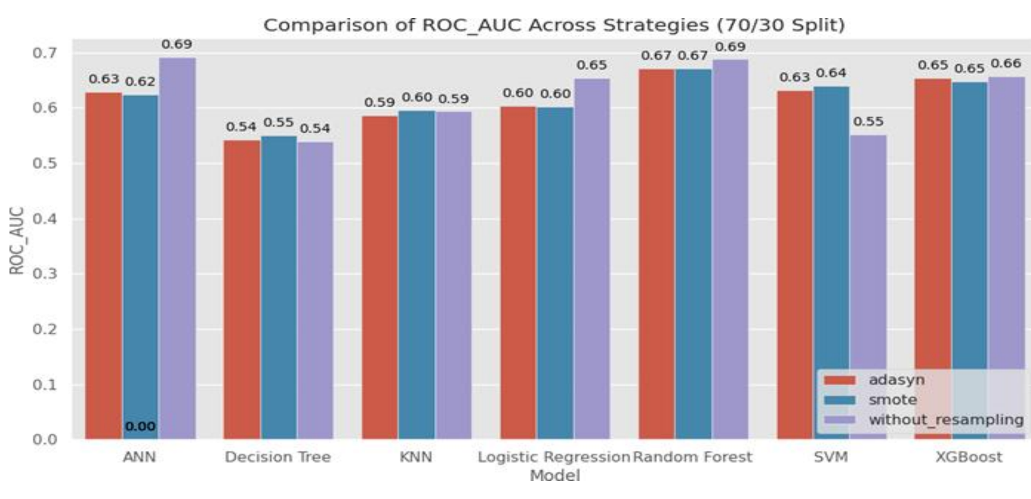

Figure 2: Performance comparison of models using ROC AUC across three training strategies (Without Resampling, SMOTE, ADASYN) on the 70/30 split.

Confusion matrix analysis (Figure 3) shows that most misclassifications occur between

Non-Diabetes and Pre-Diabetes, reflecting overlapping clinical characteristics and the progressive nature of T2D. The RF model with SMOTE and a 70/30 split demonstrated better class separation than the 80/20 split. Poorer-performing models such as Logistic Regression and SVM exhibited higher confusion among minority classes due to limited ability to capture non-linear feature interactions, whereas ensemble models better handled heterogeneous risk patterns.

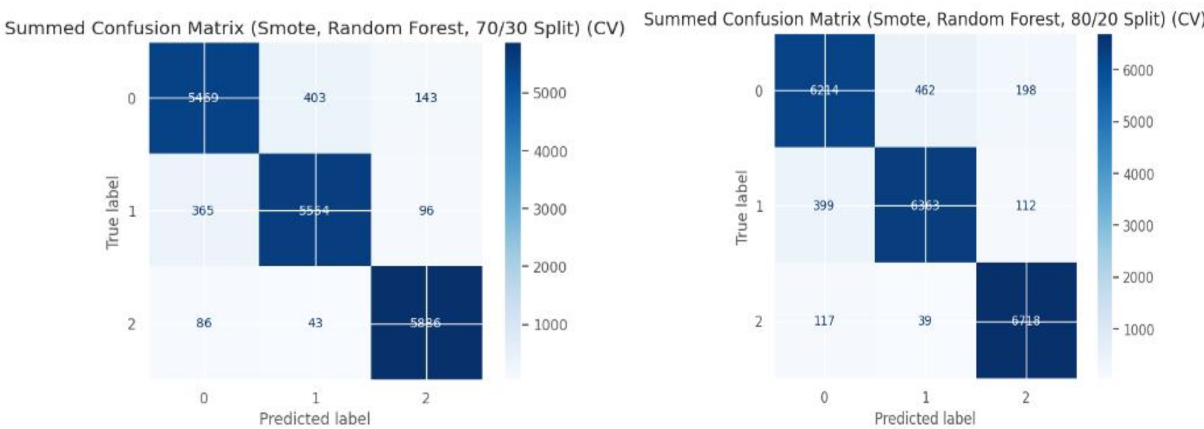

Figure 3: Confusion matrices of the two best-performing SMOTE-based Random Forest models evaluated using cross-validation

RQ3: How interpretable and clinically relevant are the predictions generated by the machine learning models for use in real-world healthcare settings in Ethiopia?

Model interpretability assessed using SHAP values (Figure 4) showed that BMI, WHR, ethnic background, alcohol use, fruit intake, and HDL cholesterol were key contributors to T2D risk. Higher BMI and WHR increased risk, while healthier dietary behaviors reduced it. Importantly, strong predictive performance was maintained without blood glucose measurements, supporting the feasibility of deploying interpretable ML models for early screening in resource-limited Ethiopian healthcare settings.

Comparison with Related Work Compared to prior studies, this research advances the literature by focusing on a nationally representative Ethiopian dataset and explicitly addressing class imbalance through SMOTE and ADASYN. Unlike studies that reported high accuracy without balancing or relied on binary classification, this work adopted a multiclass framework (non-diabetic, pre-diabetic, diabetic), enabling earlier and more actionable intervention. The substantial improvement in RF performance after resampling—from an AUC of 0.6847 to 0.9892—highlights the importance of imbalance handling for reliable prediction of underrepresented classes, particularly pre-diabetes.

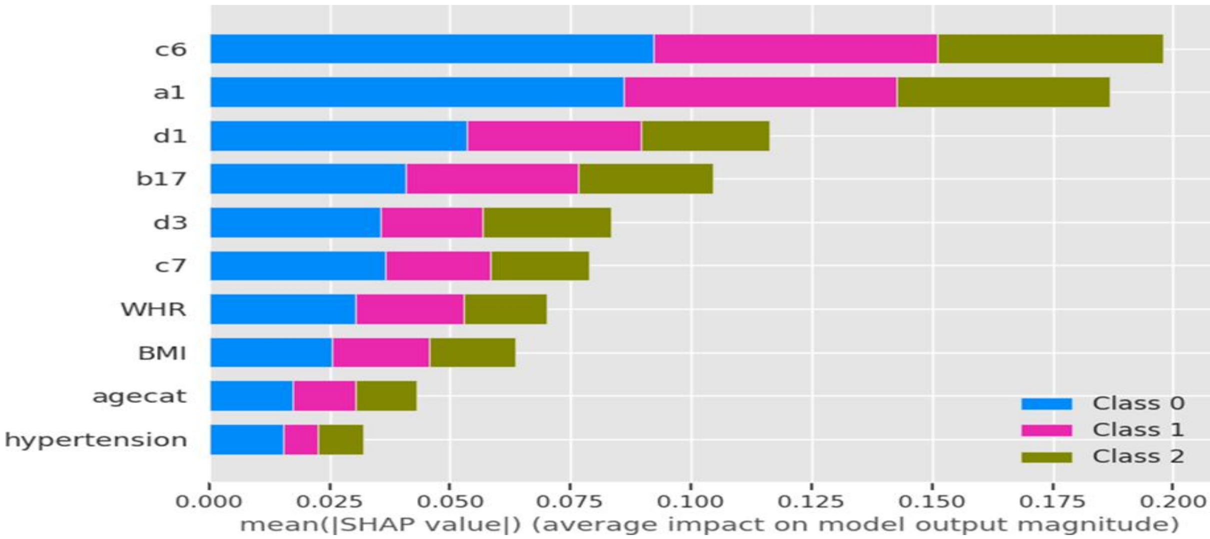

Figure 4: SHAP plot

Overall, the findings validate and extend previous research by demonstrating that balanced, interpretable ML models can achieve strong and generalizable performance even when relying on non-invasive predictors.

In summary, this study demonstrates that combining resampling techniques with cross-validation significantly enhances ML model performance on imbalanced health data. By integrating behavioral, demographic, socioeconomic, and clinical factors using a large, nationally representative dataset, the proposed approach achieves robust predictive performance and strong interpretability. These findings underscore the potential of localized, data-driven ML models to support early T2D detection and guide preventive public health interventions in Ethiopia and similar low-resource settings.

6. Conclusion

The study demonstrates through its findings that Random Forest, among other machine learning models, is highly effective in predicting the risk of Type 2 diabetes for adults in Ethiopia. The research examines a wide variety of lifestyle and social factors alongside demographic characteristics to identify their connections to the risk of diabetes. This study applied a complete set of features that align with epidemiological research and provide valuable information that is relevant to Ethiopia.

Based on study findings, we recommend strengthening public health education on hereditary diabetes risks and encouraging patients to share family medical histories to improve prediction and data quality. We also recommend that the government, through EPHI and the Ministry of Health, implement large-scale, low-cost, non-invasive diabetes screening programs, such as adapting the 60-second ADA risk test for local use. Furthermore, enhancing national NCD data tools to include diabetes diagnosis and screening history will improve disease tracking and support targeted health interventions.

Future research should focus on improving national health datasets, such as the EPHI

NCD STEPS survey, including clear indicators of Type 2 diabetes and NCD status to improve data reliability and research efficiency. Developing an Ethiopian-specific and non-invasive diabetes risk assessment tool—similar to the 60 second Type 2 diabetes risk test of the American Diabetes Association (ADA) [23] and the Indian Diabetes Risk Score (IDRS) [24] tests—will support early screening using factors such as age, BMI, and family history. Collaboration with the Ethiopian Diabetes Association is vital for creating accessible tools and mobile platforms that promote early detection, while integrating advanced machine learning and deep learning methods will strengthen model performance and generalizability.

References

[1] World Health Organization, "Noncommunicable diseases," [Online]. Available: https://www.who.int/news-room/fact-sheets/detail/noncommunicable-diseases. Accessed on June 18, 2025.

[2] International Diabetes Federation, "Idf diabetes atlas, 11th edition," https://diabetesatlas.org, 2025, accessed on June 18, 2025.

[3] M. Tesfaye, G. Shifferaw, and B. Birru, "Prevalence and factors associated with diabetes mellitus in ethiopia: A systematic review," Ethiopian Journal of Health Development, vol. 35, no. 1, p. 24–33, 2021.

[4] World Health Organization, "Global report on diabetes," [Online]. Available: https://www.who.int/publications/i/item/9789241565257. Accessed on June 18, 2025.

[5] Federal Democratic Republic of Ethiopia Ministry of Health, Guidelines on Clinical and Programmatic Management of Major Non-Communicable Diseases. Addis Ababa, 2016.

[6] F. Shiferaw, M. Letebo, and A. Misganaw, "Non-communicable diseases in Ethiopia: Dis- ease burden, gaps in health care delivery and strategic directions," Ethiop. J. Health Dev, vol. 32, no. 3, 2018.

[7] O. A. Ebrahim and G. Derbew, "Application of supervised machine learning algorithms for classification and prediction of type-2 diabetes disease status in afar regional state, northeastern Ethiopia," Scientific Reports, vol. 13, no. 7779, pp. 1–12, 2023.

[8] M. Pushpo, "Predicting diabetes using machine learning: A comparative study of super-vised classification models," Master's thesis, Brac University, Dhaka, Bangladesh, August 2023, B.Sc. in Mathematics, Summer 2023.

[9] M. R. Haque and I. Alharbi, "A dataset-specific machine learning study for predicting diabetes (type-2) in a developing country context," Indian Journal of Science and Tech- nology, vol. 15, no. 38, pp. 1932–1940, 2022.

[10] R. Krishnamoorthi, S. Joshi, H. Z. Almarzouki, P. K. Shukla, A. Rizwan, C. Kalpana, and B. Tiwari, "A novel diabetes healthcare disease prediction framework using machine learning techniques," Journal of Healthcare Engineering, vol. 2022, pp. 1–10, 2022.

[11] A. I. ElSeddawy, F. K. Karim, A. M. Hussein, and D. S. Khafaga, “Predictive analysis of diabetes-risk with class imbalance,” Computational Intelligence and Neuroscience, vol. 2022, pp. 1–16, 2022.

[12] S. Buyrukoglu and A. Akbas, “Machine learning based early prediction of type 2 diabetes: A new hybrid feature selection approach using correlation matrix with heatmap and sfs,” Balkan Journal of Electrical amp; Computer Engineering, vol. 10, no. 2, pp. 110–116, April 2022.

[13] American Diabetes Association, “Diagnosis of diabetes,” [Online]. Available: https://www.diabetes.org/a1c/diagnosis. Accessed on June 13, 2025.

[14] Ethiopian Public Health Institute, “Ethiopia steps report on risk factors for non-communicable diseases and prevalence of selected ncds,” Ethiopian Public Health Institute, Addis Ababa, Dec. 2016, page 93.

[15] M. Kuhn and K. Johnson, Applied Predictive Modeling. New York, NY, USA: Springer, 2013.

[16] M. Sokolova and G. Lapalme, “A systematic analysis of performance measures for classification tasks,” Information Processing & Management, vol. 45, no. 4, pp. 427–437, 2009.

[17] K. Abnoosian, R. Farnoosh, and M. H. Behzadi, “Prediction of diabetes disease using an ensemble of machine learning multi-classifier models,” BMC Bioinformatics, vol. 24, no. 337, 2023.

[18] M. R. A. Gowthami S., Venkata Siva Reddy, “Type 2 diabetes mellitus: Early detection using machine learning classification,” International Journal of Advanced Computer Sci- ence and Applications, vol. 14, no. 6, 2023.

[19] H. Deberneh and S. Kim, “Prediction of type 2 diabetes based on machine learning al- gorithm,” International Journal of Environmental Research and Public Health, vol. 18, no. 6, p. 1–12, 2021-03.

[20] I. J. Kakoly, M. R. Hoque, and N. Hasan, “Data-driven diabetes risk factor prediction us- ing machine learning algorithms with feature selection technique,” Sustainability, vol. 15, no. 4930, 2023.

[21] 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, 2020.

[22] B. Ozturk, T. Lawton, S. Smith, and I. Habli, “Predicting progression of type 2 diabetes using primary care data with the help of machine learning,” in Caring is Sharing – Exploiting the Value in Data for Health and Innovation. IOS Press, 2023, pp. 38–42, European Federation for Medical Informatics (EFMI).

[23] in learning. 2008 IEEE International Joint Conference on Neural Networks (IEEE World Congress on Computational.

[24] CADI Research, “The indian diabetes risk score,” [Online]. Available:

https://cadiresearch.org/topic/diabetes-indians/the-indian-diabetes-risk-score. Accessed on June 13, 2025.