Breaking Language Barriers in Healthcare: A Multilingual AI for Early-Stage Diabetes Risk Prediction

Authors: *Mikre Getu Mihrete, Hana Mekonen Tamiru, Abeshu Kebede Kelbesa, Eden Habtetsion Gebremedhin, Yared Zenebe Zewde*

Did you know that 1 in 3 adults worldwide has prediabetes a condition where blood sugar levels are higher than normal but not yet diabetic? Worse, **90% of them don't even know it** [1] [2]**.** Left unchecked, prediabetes can progress to Type 2 diabetes, leading to severe complications like heart disease, kidney failure, and nerve damage. Imagine facing a potential health crisis, but the tools designed to warn you are only available in a language you don't understand. This linguistic barrier is a critical gap, hindering non-English speakers from understanding their health risks. This is especially critical in regions like Ethiopia, where linguistic diversity creates healthcare accessibility gaps.

Our project aims to bridge this gap by developing a multilingual system that uses machine learning to predict early-stage diabetes risk in **10 languages** (Afar, Afan Oromo, Amharic, Arabic, Chinese, English, French, Hindi, Somali, and Tigrigna), making screening **accessible, interpretable, and actionable** for underserved populations. By integrating individual health data and demographics, we're enhancing accessibility, awareness, and enabling early intervention, regardless of language. This project directly contributes to the **Sustainable Development Goals (SDGs)**: to **SDG 3 (Good Health and Well-Being)** by facilitating early detection and prevention of diabetes, thus reducing healthcare burdens; and to **SDG 10 (Reduced Inequalities)** by ensuring vital health information is accessible to non-English speakers through multilingual support. Our goal is to empower underserved communities with personalized health insights in their native languages, fostering a more inclusive and equitable future in digital health.

I. The Problem: Critical Limitations in Existing Diabetes Risk Predictors

Current widely used diabetes risk calculators, such as the Finnish Diabetes Risk Score (FINDRISC) FINDRISC (Finnish Diabetes Risk Score) and QDiabetes (QDiabetes-2018), encounter significant limitations that hinder effective global prevention:

- ✓ **Language Exclusion:** Most existing tools are primarily available in English. This critical oversight disregards the needs of **over 1 billion global non-native English speakers** [3], creating a profound barrier to early detection and health literacy. In regions like Ethiopia, with its rich linguistic diversity, this exclusion severely compounds healthcare accessibility gaps.

- ✓ **Lack of Cultural Relevance:** Beyond mere translation, the efficacy of health assessments can falter if symptoms and health concepts are not culturally adapted. For instance, a direct translation of a medical term like "polyuria" (excessive urination) may not resonate or be accurately interpreted by a patient in languages like Amharic or Afan Oromo, potentially leading to miscommunication and missed diagnoses. Existing tools often lack this crucial cultural nuance, making them less relatable and actionable for diverse populations. Figure 1 effectively illustrates a key limitation in existing diabetes risk prediction tools.

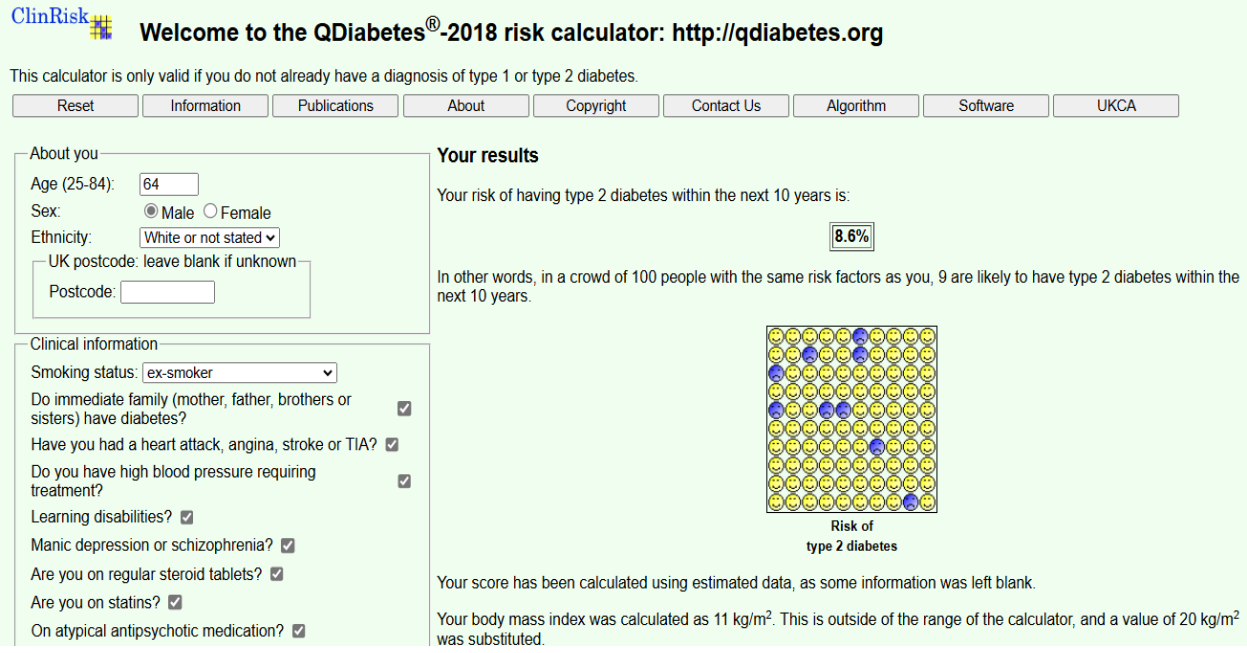

Figure 1: Language Barrier in Diabetes Risk Prediction – The QDiabetes Calculator Interface

This screenshot of the QDiabetes®-2018 risk calculator (QDiabetes-2018) demonstrates an English-only interface, highlighting how a lack of multilingual support can create significant language barriers in healthcare, particularly for non-native English speakers seeking to assess their diabetes risk.

Table 1: Language coverage of existing tools vs. our system

Feature / Tool	***QDiabetes®*** *(QDiabetes-2018)*	***FINDRISC®*** *(FINDRISC (Finnish Diabetes Risk Score))*	***Our Multilingual AI System***
Total Languages Supported	1 (English Only)	30 Global Languages	10 Local & Global Languages
African Languages Included	None	1 (Arabic)	Multiple (e.g., Afar, Afan Oromo, Amharic, Arabic, Somali, Tigrigna)
Cultural Adaptation	✗ Primarily UK-centric	✗ Western-centric	✓ Local context & demographics

To directly confront these disparities, our project introduces a groundbreaking **machine learning-powered web application** designed to democratize early-stage diabetes risk prediction:

- ✓ **Risk Prediction Without Lab Tests:** Our AI system accurately predicts risk using readily available **symptom and demographic data**, eliminating the need for expensive or inaccessible laboratory tests.
- ✓ **Expansive Multilingual Support:** We break down linguistic barriers by supporting **10 diverse languages**, including five crucial Ethiopian languages (Afar, Afan Oromo, Amharic, Somali, Tigrigna) and five widely spoken global languages (Arabic, Chinese, English, French, Hindi).
- ✓ **Culturally Adapted Symptom Descriptions:** Recognizing that health understanding is deeply rooted in culture, our system provides **culturally adapted symptom descriptions**, ensuring clarity, relevance, and accurate interpretation for all users.

Effective early-stage diabetes risk prediction is fundamental for improving global health outcomes. Our proposed Multilingual AI System directly addresses the critical linguistic and cultural gaps in existing tools, which is further highlighted in our comparative analysis, as shown in Table 1.

II. The Process: Our Methodology

1. Data Collection & Preprocessing

We utilized the UCI Early-Stage Diabetes Risk Prediction Dataset, which consists of 520 patient records with 17 features, including demographic features (age and gender) and symptom-related features (polyuria, polydipsia, sudden weight loss, weakness, polyphagia, genital thrush, visual blurring, itching, irritability, delayed healing, partial paresis, muscle stiffness, alopecia and obesity) and a target label indicating early-stage diabetes risk (Positive or Negative) [4]. Because the dataset relies on simple, physician-approved questions rather than costly laboratory tests, it is particularly well suited for symptom-based screening in low resource environments. The key challenges we faced included language expansion and cultural adaptation. We translated symptom questionnaires into ten languages using professional translators and localized terms to ensure they resonated with users.

Raw data, regardless of its richness, requires careful preparation before it can effectively train a machine. Our process began with comprehensive data preprocessing, during which we meticulously cleaned and encoded all responses; this included mapping "Yes" or "No" answers to binary 1/0 values and converting target categorical field into appropriate numeric codes. A thorough quality check confirmed a remarkably clean dataset, notably devoid of any missing values. This ensured that all valuable data points could be fully utilized in our subsequent analyses.

2. Exploratory Data Analysis

Following preprocessing, we conducted Exploratory Data Analysis (EDA). This involved examining the age distribution, which revealed bimodal peaks around 40 and 60 years, and visualizing the target variable to ensure class balance. We also scaled the 'Age' feature using ***StandardScaler*** to ensure it contributed equally to our models, preventing it from dominating other

features due to its larger numerical range. We applied the **Chi-squared ($\chi 2$) test** primarily for feature selection.

3. Model Training

To build a robust early-stage diabetes risk predictor, we trained a diverse suite of ten supervised classifiers on our preprocessed dataset. We used an 80% train / 20% test split, stratified to preserve class balance, and balanced the training set *RandomOverSampler* to address the initial class imbalance (256 positives and 160 negatives in the training split became 256 positives, and 256 negatives after oversampling). Our lineup spanned linear models (Logistic Regression [5], Linear Discriminant Analysis [6]), probabilistic methods (Gaussian Naive Bayes [7]), distance-based learners (K-Nearest Neighbors [8]), tree-based algorithms (Decision Tree [9], Random Forest [10], AdaBoost [11]), a margin-based classifier (Linear SVM [12]), a gradient-boosted ensemble (XGBoost [13]), and a small neural network (Multilayer Perceptron [14]). Each model ingested label-encoded, binary symptom flags plus a standardized Age feature. We first wrapped model training in a stratified 10-fold cross-validation loop to gauge out-of-sample performance and identify promising candidates.

4. Hyperparameter Tuning

For models with tunable knobs, we performed systematic searches to unlock their best performance. For models with tunable knobs, we performed systematic searches to unlock their best performance. We primarily used ***RandomizedSearchCV*** for models with larger parameter grids, such as Random Forest and XGBoost, to efficiently explore a wide range of hyperparameter combinations. Crucially, ***RandomizedSearchCV*** internally employs cross-validation (by default, 5-fold) for each trial to ensure fair and robust evaluation of different hyperparameter sets. This optimization involved tweaking parameters like ***n_estimators, max_depth***, and ***max_features*** for the Random Forest Classifier. Each search was nested inside this cross-validation framework to ensure fair, robust hyperparameter selection and prevent overfitting.

5. Model Evaluation

After tuning, we evaluated each model on the held-out test set using a battery of metrics: **Accuracy** (overall correctness), **Precision** (correct positive predictions), **Recall** (coverage of true positives), **F1-Score** (harmonic mean of precision & recall), and **ROC AUC** (ranking ability across thresholds). To visualize comparative strengths and weaknesses, we plotted confusion matrices highlighting true/false positives and negatives, and overlaid all models' ROC curves in a single chart. This multi-metric, visual approach allowed us to identify the model that not only scored highest by AUC but also maintained balanced precision and recall, critical for real-world screening where false negatives carry serious health risks.

III. Implementation: From Theory to Practice

The implementation process was systematically executed in the following sequential phases: data preprocessing, Exploratory Data Analysis (EDA), and model deployment.

1. Data Preprocessing and Exploratory Data Analysis

A thorough initial check revealed a remarkably clean dataset, notably devoid of any missing values. Furthermore, a rigorous analysis of the 'Age' variable confirmed the absence of extreme outliers, ensuring that all valuable data points could be fully utilized. This pristine initial state significantly streamlined our preprocessing efforts.

Out[3]:

	Age	Gender	Polyuria	Polydipsia	sudden weight loss	weakness	Polyphagia	Genital thrush	visual blurring	Itching	Irritability	delayed healing	partial paresis	muscle stiffness	Alopecia	Obesity
0	40	Male	No	Yes	No	Yes	No	No	No	Yes	No	Yes	No	Yes	Yes	Yes
1	58	Male	No	No	No	Yes	No	No	Yes	No	No	No	Yes	No	Yes	No
2	41	Male	Yes	No	No	Yes	Yes	No	No	Yes	No	Yes	No	Yes	Yes	No
3	45	Male	No	No	Yes	Yes	Yes	Yes	No	Yes	No	Yes	No	No	No	No
4	60	Male	Yes	Yes	Yes	Yes	Yes	No	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes

Figure 2: Sample of the Original Dataset

The symptoms and demographic details in our original dataset were in categorical text formats, such as 'Yes'/'No' and 'Male'/'Female'. To prepare this data for machine learning algorithms, which primarily operate on numerical inputs, we performed ***binary encoding***. We systematically

translated these categorical responses into a numerical form: 'Yes' was mapped to 1, 'No' to 0; 'Male' was mapped to 1, and 'Female' to 0. Similarly, the target class ('Positive' for diabetes risk, 'Negative' for no risk) was converted to 1 and 0, respectively. This binary encoding is an essential step, transforming human-readable descriptors into a machine-interpretable format.

Out[21]:

	Age	Gender	Polyuria	Polydipsia	sudden weight loss	weakness	Polyphagia	Genital thrush	visual blurring	Itching	Irritability	delayed healing	partial paresis	muscle stiffness	Alopecia	Obesity
0	40	1	0	1	0	1	0	0	0	1	0	1	0	1	1	1
1	58	1	0	0	0	1	0	0	1	0	0	0	1	0	1	0
2	41	1	1	0	0	1	1	0	0	1	0	1	0	1	1	0
3	45	1	0	0	1	1	1	1	0	1	0	1	0	0	0	0
4	60	1	1	1	1	1	1	0	1	1	1	1	1	1	1	1

Figure 3: Sample of the Dataset After Categorical Transformation

Our Exploratory Data Analysis (EDA) provided vital insights by visualizing key trends within the dataset. For instance, the age distribution prominently displayed bimodal peaks around 40 and 60 years, suggesting that these specific age groups might exhibit increased susceptibility to early diabetes symptoms, as shown in Figure 4.

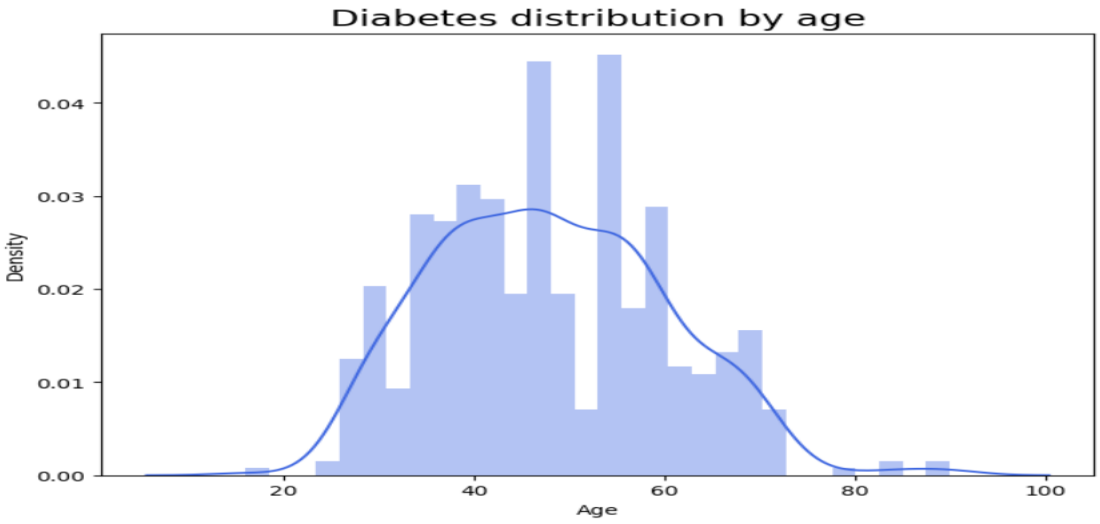

Figure 4: Diabetes distribution by age

Furthermore, through the Chi-squared ($\chi 2$) test, we successfully identified the top predictive symptoms. As illustrated in Figure 5 (Feature Importance Score) and visually confirmed in Figure 6 (Feature Importance Visualization), polydipsia (excessive thirst), polyuria (frequent urination), and sudden weight loss emerged as the strongest indicators of diabetes risk within the dataset.

	column	scores
3	polydipsia	120.785515
2	polyuria	116.184593
4	sudden weight loss	57.749309
12	partial paresis	55.314286
1	gender	38.747637
10	irritability	35.334127
6	polyphagia	33.198418
14	alopecia	24.402793
0	age	18.845767
8	visual blurring	18.124571
5	weakness	12.724262
7	genital thrush	4.914009
13	muscle stiffness	4.875000
15	obesity	2.250284
11	delayed healing	0.620188
9	itching	0.047826

Figure 5: Feature importance score

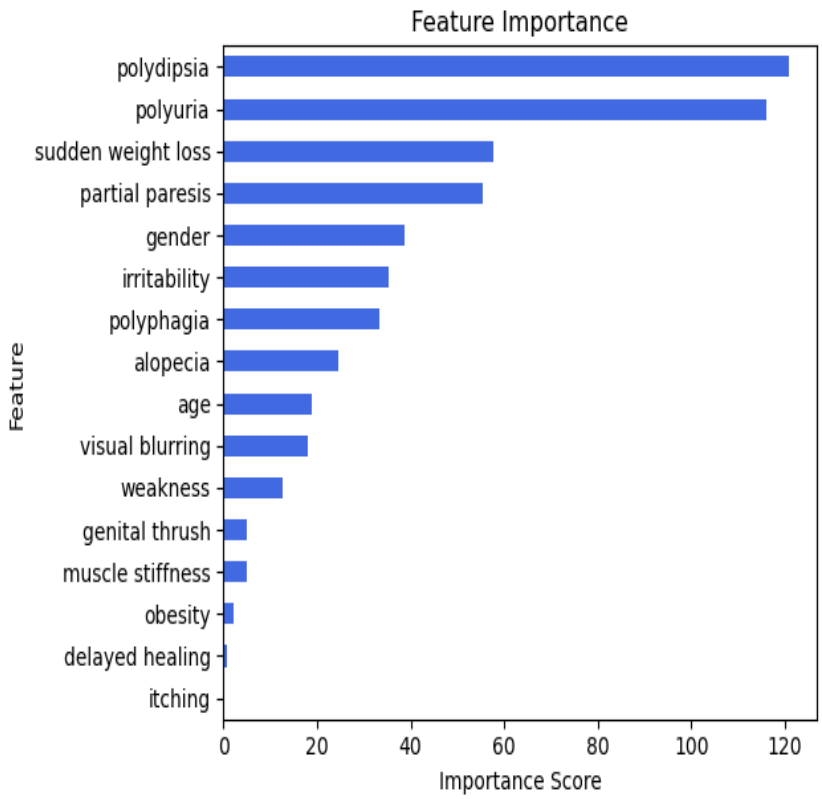

Figure 6: Feature importance visualization

2. Model Training and Results

Our approach involved extensive **hyperparameter tuning** using **RandomizedSearchCV** to significantly optimize our model's performance. For our Random Forest, we meticulously refined key parameters: the ***n_estimators*** were set to ***1100***, determining the number of decision trees in the forest; the ***max_depth*** was set to ***16***, controlling the maximum depth of each tree; and ***min_samples_split*** was refined to ***10***, specifying the minimum number of samples required to split an internal node. These adjustments were crucial for maximizing the model's predictive capabilities.

We tested over ten algorithms, as shown in Table 2, with the Random Forest Classifier (RFC) consistently **emerging** as our top-performing model, achieving an impressive **97.12%** accuracy and a high ROC AUC score of **99.77%** on the training set during cross-validation. After hyperparameter tuning and evaluation on the held-out test set, the final tuned Random Forest Classifier achieved a remarkable **99.04%** accuracy and an AUC of **100%**.

Table 2: Performance Metrics of Trained Models

Model	*Accuracy*	*Precision*	*Recall*	*F1 Score*	*ROC AUC*
LR	91.11%	91.18%	91.11%	91.13%	96.45%
LDA	87.50%	88.84%	87.50%	87.65%	96.22%
KNN	91.59%	92.32%	91.59%	91.67%	98.31%
DTC	94.47%	94.62%	94.47%	94.50%	94.57%
SVM	93.99%	94.05%	93.99%	94.01%	98.71%
RFC	**97.12%**	**97.15%**	**97.12%**	**97.12%**	**99.77%**
ABC	91.35%	91.54%	91.35%	91.39%	96.87%
XGB	96.39%	96.48%	96.39%	96.41%	98.99%
NB	87.74%	87.79%	87.74%	87.76%	94.63%
MLP	95.19%	95.27%	95.19%	95.21%	99.44%

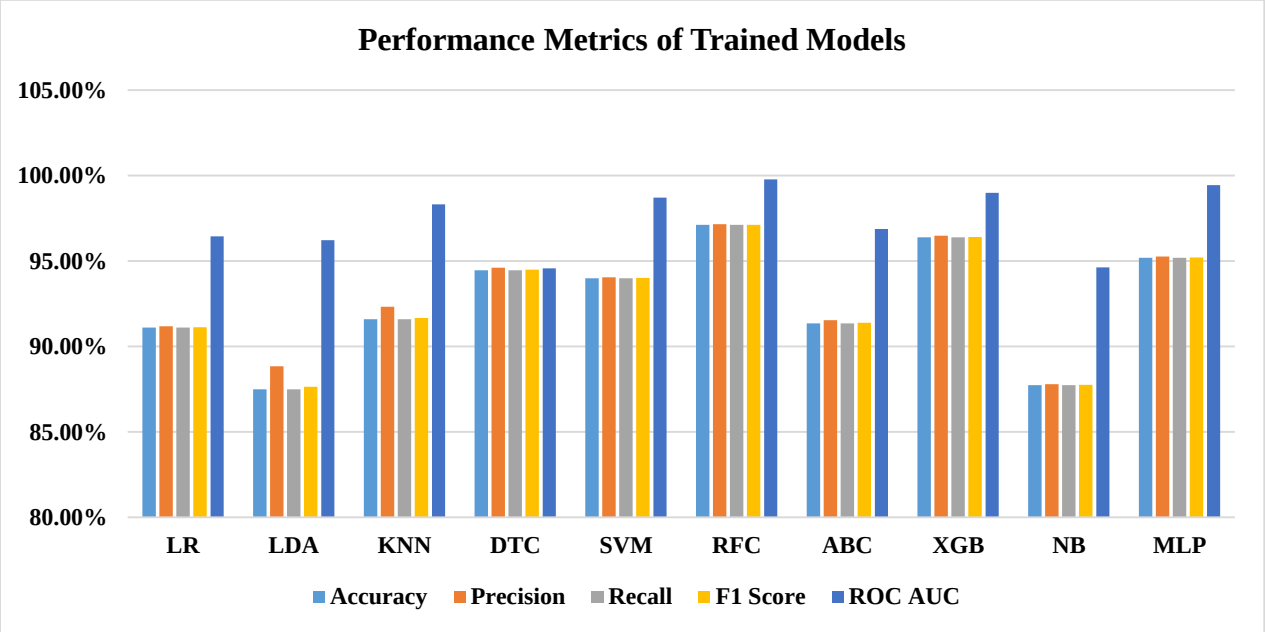

Figure 7: Performance Metrics of Trained Models

To visualize comparative strengths and weaknesses, we plotted confusion matrices highlighting true/false positives and negatives, and overlaid all models' ROC curves in a single chart.

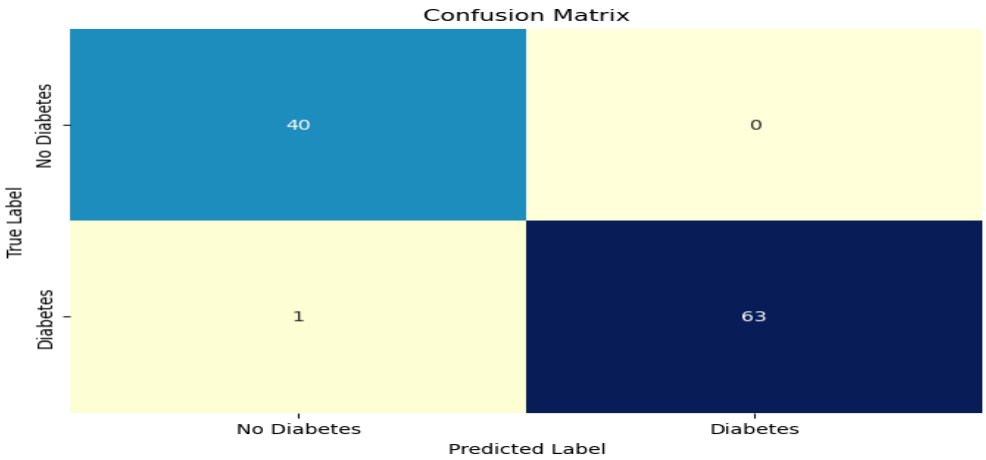

Figure 8: Confusion Matrix for Tuned Random Forest Classifier

The confusion matrix above demonstrates the high accuracy of our model, showing very few misclassifications.

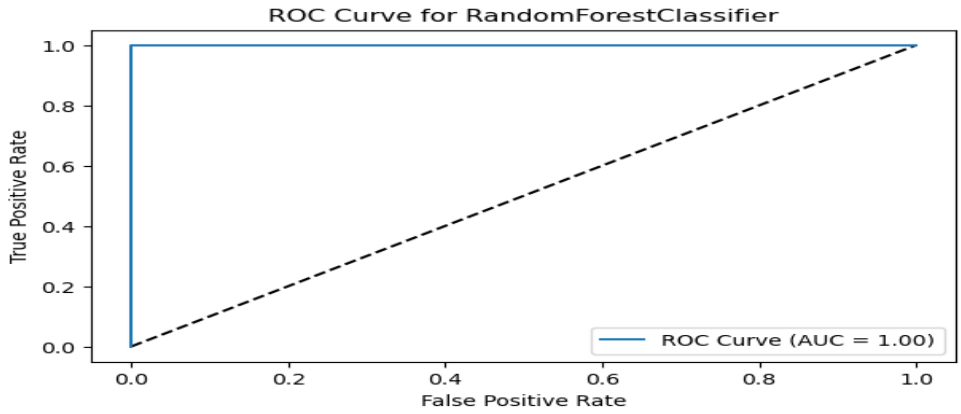

Figure 9: ROC Curve for Tuned Random Forest Classifier

The ROC curve, with an AUC of 100%, indicates that the tuned Random Forest Classifier is an almost perfect discriminator between positive and negative diabetes risk cases on the test set.

3. Model Deployment: Global Accessibility Through Technology

The final step in bringing this project to life was deploying the web application. Bringing our high-performing, multilingual model to users required a robust deployment strategy that prioritizes accessibility and ease of use. We developed a user-friendly web application to facilitate this, enabling individuals from diverse linguistic backgrounds to access early diabetes risk prediction.

Tech Stack:

- ✓ **Frontend:** Built with **Flask**, **HTML**, **JavaScript**, and **CSS** to create an interactive web application, prominently featuring a crucial language toggle that allows users to instantly switch the interface.
- ✓ **Backend:** Powered by **Python** and also utilizing the **Flask framework**, this layer handles data processing, model inference, and dynamic content delivery, ensuring rapid response times for predictions.
- ✓ **Translation:** Our system integrates a **pre-translated dictionary-based approach** for all content translation. This is further complemented by rigorous manual validation for medical terms, guaranteeing not only linguistic accuracy but also vital cultural relevance in each of the 10 supported languages (Afar, Afan Oromo, Amharic, Arabic, Chinese, English, French, Hindi, Somali, and Tigrigna).

User Flow: The application's user experience is designed to be intuitive and straightforward:

1. Users first **select their preferred language** from the available options (e.g., Amharic).
2. They then **input their age, gender, and localized symptom descriptions** via the intuitive graphical user interface.
3. Finally, they **receive an instant risk prediction** accompanied by a clear explanation, all presented in their chosen language.

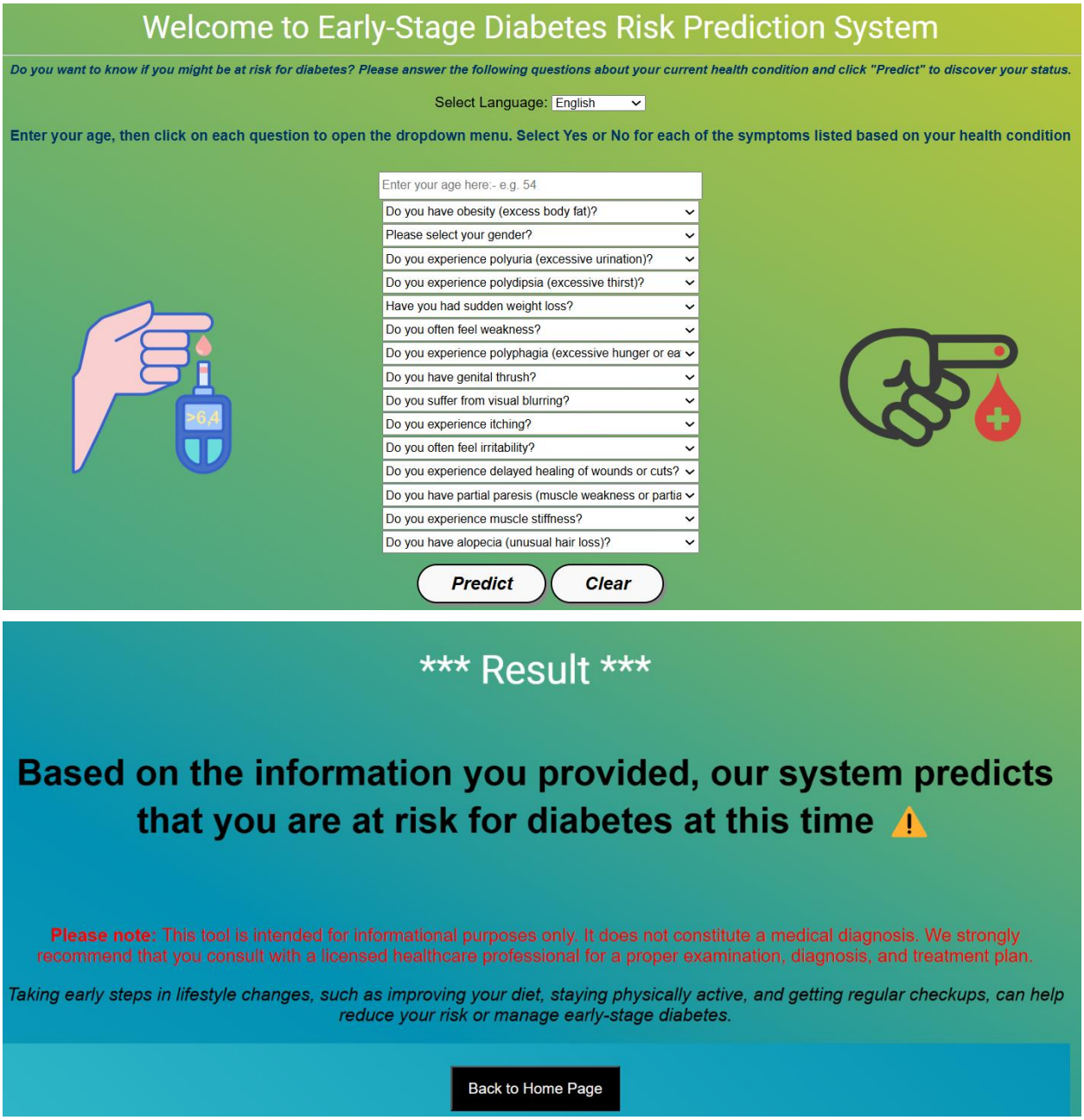

Figure 10: Screenshot of our deployed Multilingual AI Diabetes Risk Prediction System

IV. Discussion

The results of our model training and evaluation are highly promising, demonstrating that machine learning can effectively predict early-stage diabetes risk using a relatively simple, symptom-based dataset. The Random Forest Classifier consistently outperformed other models, achieving high accuracy, precision, and recall, along with an outstanding ROC AUC score. This suggests its

strong capability in correctly identifying both individuals at risk (true positives) and those not at risk (true negatives), while minimizing false positives and false negatives, which are crucial in a health screening context.

A key factor in achieving these results was the meticulous data preprocessing, including handling categorical variables and balancing the dataset using *RandomOverSampler*. This step was vital in preventing the model from being biased towards the more prevalent class. The feature importance analysis (Figure 5) reinforced existing medical understanding, highlighting polydipsia, polyuria, and sudden weight loss as major indicators, thus lending credibility to our model's predictions.

Despite these successes, certain limitations and challenges were encountered. Our Random Forest model achieved an impressive accuracy of 99.04%; however, we faced challenges including managing the small dataset size (520 records) and ensuring dialect-specific translations. While the UCI dataset is well-suited for a proof-of-concept, its size might limit generalizability to extremely diverse, large-scale populations. The initial class imbalance also necessitated oversampling, which, while effective, can sometimes lead to overfitting on the synthetic samples if not carefully managed. We learned that medical translations require cultural nuance (e.g., "blurred vision" in Tigrigna may not have a direct translation), and the user interface design must accommodate specific linguistic requirements, such as right-to-left languages (Arabic and Tigrigna). Furthermore, this project represents an initial technical validation; real-world clinical deployment would require extensive clinical trials and regulatory approvals.

V. Conclusion: Implications and Inspirations

Our "Breaking Language Barriers in Healthcare" project profoundly demonstrates how the integration of machine learning with dedicated multilingual design can democratize early diabetes screening. By prioritizing linguistic diversity, we have created a tool that empowers communities in Ethiopia and globally. Our system not only addresses immediate healthcare needs but also contributes significantly to broader global health initiatives, aligning directly with **Sustainable Development Goal (SDG) 3: Good Health and Well-Being** (facilitating early detection and prevention of diabetes) and **SDG 10: Reduced Inequalities** (ensuring vital health information is accessible to non-English speakers). The robust performance of our Random Forest model,

underpinned by rigorous preprocessing and feature selection, underscores the feasibility and impactful potential of this approach.

Recommendations

For those inspired to continue leveraging digital health and AI for social good, we offer the following recommendations:

- ✓ **Prioritize Accessibility:** Integrate language, digital literacy, and internet access considerations from the project's inception.
- ✓ **Embrace Cultural Adaptation:** Go beyond mere translation to ensure content and concepts resonate with diverse cultural contexts.
- ✓ **Leverage Open-Source Resources:** Utilize platforms like the UCI Machine Learning Repository and libraries like scikit-learn for impactful development.
- ✓ **Foster Collaboration:** Engage medical professionals, linguists, and community leaders to ensure practical relevance and ethical implementation.
- ✓ **Focus on Actionable Insights:** Design systems that empower users to take concrete steps towards improved health outcomes.

VI. Future Work: Reflecting on the Journey

Our journey in developing this multilingual AI system has opened numerous avenues for future exploration and refinement to maximize its real-world impact:

- ✓ **Enhanced Accessibility:** Future work will focus on developing dedicated **mobile applications** for increased reach, and integrating **voice input** functionalities to assist low-literacy users.
- ✓ **Advanced Model Development:** We aim to explore more sophisticated techniques like **deep learning architectures** for potential accuracy enhancements and greater generalizability. Integrating **explainable AI (XAI)** will also be crucial to provide users with deeper insights into their predictions, fostering trust.

- ✓ **Real-World Integration and Validation:** We envision implementing the model in real-world clinical settings, coupled with rigorous validation through **clinical trials**, and exploring secure integration with existing healthcare systems.
- ✓ **Broader Language and Data Scope:** Continuously updating language support to include more dialects and investigating the ethical integration of **external data sources** will further strengthen the model's robustness and personalized predictions.
- ✓ **Feedback Loop Mechanisms:** Designing a system that allows for user feedback or follow-up data to continuously improve the model over time.

By leveraging technology and machine learning, we can take significant strides toward a healthier future for all. Join us in this journey to make diabetes risk prediction accessible and understandable for everyone, regardless of their language.

References

[1] J. B. Echouffo-Tcheugui, L. Perreault, L. Ji, and S. Dagogo-Jack, “Diagnosis and Management of Prediabetes: A Review,” JAMA, vol. 329, no. 14, pp. 1206–1216, Apr. 2023, doi: 10.1001/jama.2023.4063.

[2] U. Hostalek, “Global epidemiology of prediabetes - present and future perspectives,” Clin. Diabetes Endocrinol., vol. 5, no. 1, p. 5, May 2019, doi: 10.1186/s40842-019-0080-0.

[3] “Report on the Global Distribution of Native and Non-Native English Speakers.” Accessed: Jun. 15, 2025. [Online]. Available: https://www.americantesol.com/tesol-report.html

[4] UCI, “Early Stage Diabetes Risk Prediction.” UCI Machine Learning Repository, 2020. doi: 10.24432/C5VG8H.

[5] Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction* (2nd ed.). Springer. (Chapter 4: Linear Methods for Classification)

[6] Fisher, R. A. (1936). The Use of Multiple Measurements in Taxonomic Problems. *Annals of Eugenics, 7*(2), 179-188.

[7] Rish, I. (2001). An empirical study of the naive Bayes classifier. In *IJCAI 2001 Workshop on Empirical Methods in Artificial Intelligence* (pp. 41-46).

[8] Altman, N. S. (1992). An Introduction to Kernel and Nearest-Neighbor Nonparametric Regression. *The American Statistician, 46*(3), 175-185.

[9] Breiman, L., Friedman, J., Olshen, R., & Stone, C. (1984). *Classification and regression trees*. Wadsworth International Group.

[10] Breiman, L. (2001). Random forests. *Machine Learning, 45*(1), 5-32.

[11] Freund, Y., & Schapire, R. E. (1997). A Decision-Theoretic Generalization of On-Line Learning and an Application to Boosting. *Journal of Computer and System Sciences, 55*(1), 119-139.

[12] Cortes, C., & Vapnik, V. (1995). Support-Vector Networks. *Machine Learning, 20*(3), 273-297.

[13] Chen, T., & Guestrin, C. (2016). XGBoost: A Scalable Tree Boosting System. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 785-794).

[14] Goodfellow, I., Bengio, Y., & Courville, A. (2016). *Deep Learning*. MIT Press. (Chapter 6: Deep Feedforward Networks)