Interpretable Machine Learning for Proactive Diabetes Screening: Insights from a Stacking Ensemble on BRFSS Data

***Abstract*—Detection of diabetes is a critical challenge in health care, as detection contributes to the prevention of serious complications and enhances patient outcomes. This study proposes an ensemble model for diabetes and prediabetes detection using the CDC BRFSS 2015 health indicators. After cleaning and balancing the dataset, several machine learning models were compared and the top three performers ExtraTrees, RandomForest, and DecisionTree were selected as base learners. These were combined in a stacking ensemble with CatBoost as the meta learner. The proposed model, a stacking ensemble tuned with RandomizedSearchCV, achieved a test accuracy of 98.87% and a ROC–AUC of 0.9978, outperforming all single models and showing strong, stable results across cross-validation. Explainable AI techniques, including permutation importance, SHAP, and LIME, were used to explain the model's decisions, consistently highlighting BMI, Age, and General Health as the most influential features. This work demonstrates that a carefully tuned ensemble approach, combined with explainable AI, can deliver highly accurate and transparent diabetes prediction for large scale health data.**

***Keywords*—Diabetes prediction, stacking, RandomizedSearchCV, hyperparameter tuning, Explainable AI, SHAP, LIME, Permutation importance, feature selection, Health indicators, classification accuracy, ROC–AUC, Healthcare analytics.**

I. Introduction

Diabetes detection is the process of finding out if someone has diabetes or is moving toward it, using blood glucose measures and key risk signs, so care can start early and prevent serious harm to the eyes, kidneys, nerves, and heart [1]. It matters because diabetes often begins quietly with few symptoms, so people may not notice problems until complications appear [2]. The burden is large worldwide, and timely detection helps lower the chance of long term damage and death related to high blood sugar. To conclude, diabetes detection aims to identify disease or high risk early enough to guide treatment, lifestyle changes, and follow up before severe complications develop [3].

Nowadays, diabetes is diagnosed mainly with standard tests such as HbA1c, fasting plasma glucose, oral glucose tolerance tests, and sometimes random glucose when classic symptoms are present [4]. Because many patients show few early signs, clinicians also consider risk factors such as excess weight, high blood pressure, and abnormal lipids to decide who should be screened sooner. To support faster, scalable, and more consistent detection, machine learning methods are used to learn patterns from health data, often combined with class imbalance handling and careful model tuning [5]. Explainable AI tools are increasingly being applied so that model decisions can be interpreted by clinicians and aligned with medical reasoning.

Prior studies show growing interest in automating diabetes detection, but many focus on small or narrow datasets, rely on a few algorithms, and often report single metrics without model transparency, which limits generalizability and clinical uptake. Building on those gaps, this research uses the large CDC BRFSS 2015 indicators to perform multiclass detection (non diabetes, prediabetes, diabetes) with a balanced, transparent, and tunable pipeline that emphasizes fair evaluation across classes and interpretable outputs. The proposed model is a stacking ensemble from the top three base models while rotating meta learners, explains the final ensemble with SHAP, LIME, and permutation importance, and refines it with RandomizedSearchCV and Optuna to maximize reliable performance. The key contributions of this study include:

- **Evidence Driven Stacking:** Selected the top three models by accuracy as base learners and systematically rotated the remaining models as the meta learner to find the strongest ensemble.
- **Transparent Ensemble Explanations:** Applied SHAP, LIME, and permutation importance to the best stacking configuration for both global and case level interpretability.
- **Separate Hyperparameter Tuning:** Tuned the final stacking ensemble in two independent runs, one with RandomizedSearchCV and another with Optuna using stratified cross validation for reliable selection.

The paper is organized in the sequenced as: SectionII provides an outline of existing works related to this study. SectionIII provides an in-depth description of the methodology have employed in this study. Next, SectionIV analyzes the experimental result. Finally, SectionV gives a brief conclusion and future research directions in this field.

II. Related Research

The related work is a critical element in the structure of research paper to understand the landscape of knowledge in the chosen field. Shows the methods, findings, and limitations of previous studies and identifies gaps where new contributions can be made.

Recent end to end systems pair high performing gradient boosting models with explainability and deployment, exemplified by Tasin et al. [1] who fuse Pima with a private Bangladeshi cohort, address imbalance via SMOTE/ADASYN, explain with SHAP/LIME, and deliver web and Android tools

for instant screening. El Sofany et al. [6] follow a similar public private design with SMOTE, broad model benchmarking led by XGBoost, SHAP based transparency, domain adaptation checks, and a mobile app for point of care use. Dharmarathne et al. [7] center usability by embedding XGBoost with local and global SHAP into a self explainable interface that surfaces per prediction reasons.

Methodological pipelines on tabular clinical data emphasize preprocessing and comparative baselines, with Febriana et al. [2] showing Naive Bayes consistently beating k Nearest Neighbors across multiple splits and 10 fold validation on Pima. Yakut [8] reports Random Forest as best among Extra Trees and Gaussian Process Classifier after min–max scaling and 5 fold validation, while Bhat [9] attains strong Decision Tree scores that reinforce tree based competitiveness on this feature set. Beyond baselines, Patro [4] boost average ML accuracy via correlation based modeling and a CNN, and Sampath [10] show a hybrid AdaBoost+XGBoost ensemble with SMOTE and K fold validation achieving high AUC.

Larger or multiclass clinical contexts and best practice frames broaden scope: Chou [3] train on 15,000 outpatient records and find boosted trees with AUC 0.991 in a single center female cohort, while Abnoosian [5] design an AUC weighted ensemble for a three class Iraqi dataset with near perfect cross validated performance. Jiang [11] mine 252,176 community follow ups to build a Random Forest risk model and an operational scorecard for large scale screening, as Oikonomou and Khera [12] stress pairing discrimination with calibration, external validation, equity, and net clinical benefit.

Many of the studies face same limitations such as heavy reliance on Pima or single site cohorts constrains generalizability; missing or imputed measures add uncertainty; imbalance and domain shift threaten robustness. This study closes those gaps with a lean, evidence driven stacking setup. It ranks models by cross validated accuracy, uses the top three as base learners, and rotates the rest as the meta learner to pick the strongest ensemble, improving stability across datasets. It then explains the final stack with SHAP, LIME, and permutation importance for both global and per case views, keeping results transparent and useful. Finally, it tunes the stack twice—once with RandomizedSearchCV and once with Optuna—under stratified cross validation, alongside principled imbalance handling and calibration checks, so the model is stable, reproducible and ready for practical use.

III. Methodology

The methodology presents a clear, sequential approach to solving the problem, from data collection to diabetes detection which is presented in Fig. 1. It assures productive data handling, effective model training, and reliable, understandable results. The focus is on boosting performance, ensuring transparency, and addressing any challenges to deliver accurate and actionable outcomes.

A. Dataset Description

The dataset used is the CDC Diabetes Health Indicators from BRFSS 2015, comprising 253,680 survey records with 21 predictor variables and a three class outcome Diabetes (0 = no diabetes, 1 = prediabetes, 2 = diabetes). Features span demographics and health behaviors, including high blood pressure, high cholesterol, recent cholesterol check, BMI, Smoking, heart disease, recent physical activity, fruit and vegetable intake, heavy alcohol use, healthcare coverage, cost related care avoidance, self rated general health, difficulty walking, and age. It is provided as an analysis ready CSV with integer/categorical encodings and no reported missing values, facilitating direct use in machine learning workflows. Class imbalance is present (prediabetes and diabetes are less frequent than, non diabetes), and each row represents one adult respondent, enabling multiclass prediction and model explainability on population scale survey data grounded in the 2015 BRFSS methodology and documentation.

B. Data Preprocessing

Missing Value Handling: A two step deletion strategy was used first, columns with more than 50% missing values were removed (kept only features with at least 50% non null entries); second, any remaining records containing at least one missing value were dropped, yielding a complete case dataset without imputation and preserving the native distribution of retained variables.

Dataset Balancing: The target Diabetes_012 was skewed, so a bootstrap oversampling strategy was applied to equalize classes. Specifically, minority classes were expanded with replacement to the size of the majority class while leaving the majority unchanged, then all subsets were concatenated and shuffled for a balanced training table. After balancing, each class (0 = no diabetes, 1 = prediabetes, 2 = diabetes) contains 213,703 records, for a total of 641,109 instances. This simple resampling reduces majority bias, improves minority recall, and stabilizes learning under stratified cross validation, while preserving the original feature distributions of each class.

Correlation Heatmap: The correlation heatmap in Fig. 2 shows how the target Diabetes_012 relates to the top correlated features and how those features relate to each other. General Health has the strongest positive link with the target at about (0.30), followed by HighBP (0.27), BMI and DiffWalk (both 0.22), HighChol (0.21), and Age (0.19), marking them as the most informative predictors. HeartDiseaseorAttack and PhysHlth also show positive associations with the target (0.18), whereas `Income` (-0.17) and Education (0.13) are inversely correlated, suggesting higher socioeconomic status aligns with lower diabetes risk. Among feature–feature relationships, GenHlth aligns strongly with PhysHlth (0.52) and DiffWalk (0.46), and Income correlates with Education (0.45), which are clinically sensible patterns. Most remaining correlations are modest, indicating limited redundancy so each variable contributes complementary information for modeling.

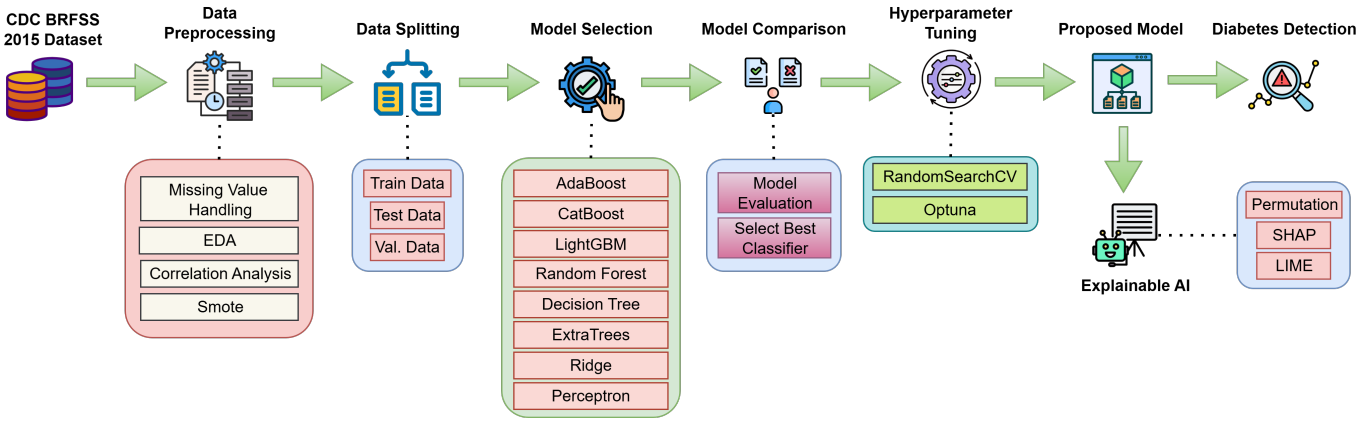

Fig. 1: Methodology of Diabetes Detection.

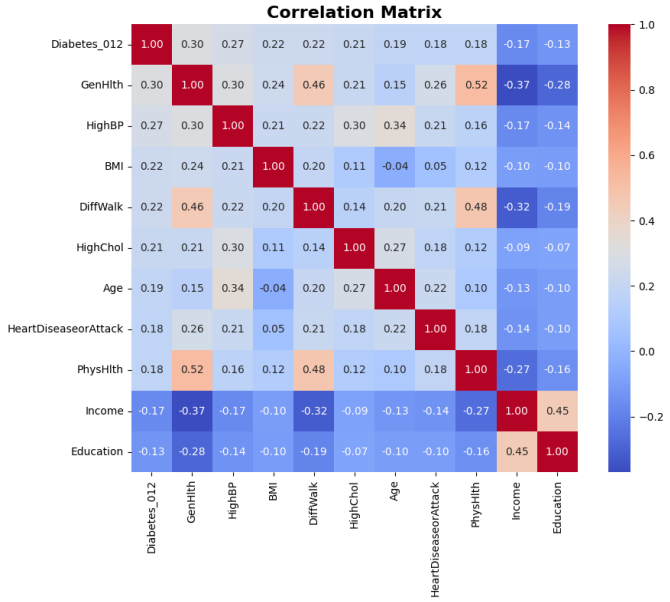

Fig. 2: Correlation Heatmap of Top 10 Features with Label Column.

C. Data Splitting

The dataset was partitioned with a stratified two step split to preserve class ratios of Diabetes_012 and to avoid leakage. First, 70% of data was allocated to the training and the remaining 30% was then split evenly into validation and test (15%/15%). The training set is used for fitting and internal cross validation during model selection; the validation set supports hyperparameter tuning, thresholding, and early stopping; and the test set is held out once for an unbiased estimate of generalization.

D. Model Selection

The classification models used in this work employ different techniques to predict outcomes. Multiple machine learning models AdaBoost, CatBoost, LightGBM, Random Forest, Decision Tree, Extra Trees, RidgeClassifier, and Perceptron were applied stratified train/validation/test split using accuracy and precision as the primary criteria. The top performer by accuracy was then examined in the full training set and audited with a classification report, confusion matrix, ROC curves, AUC, and 5-fold cross validation accuracy to verify stability. At the same time, the top three pretrained models by validation accuracy were explicitly recorded for downstream ensembling.

For the ensemble, a stacking approach was adopted where those top three pretrained models served as fixed base learners, and several candidates were rotated as the meta learner to find the best combiner. The winning stack was selected by test accuracy and then fully evaluated with stratified 5 fold cross validation, ROC–AUC (one vs rest), and confusion matrix to confirm that gains held across classes, providing a robust foundation for later explanation and calibration.

E. Proposed Model

The proposed model is a stacking ensemble that uses the top three performers from initial machine learning models as base learners. These models each make their own predictions, which are then fed into meta learner that was selected after outperforming other candidates in this role. The entire ensemble was then fine-tuned for optimal performance using RandomsearchCV and Optuna. The proposed model architecture is shown in Table 3.

Ultimately, explainable AI techniques like Permutation Importance, SHAP, and LIME are used to interpret the model's decisions, ensuring that the key health factors driving its predictions are identified, making the model both accurate and transparent. The proposed model configuration is depicted in Table I.

TABLE I: Configuration of the Proposed Model.

Component	Configuration
Base learners	ExtraTrees, RandomForest, DecisionTree
Meta learner	CatBoost
Tuned depth	9–10
Learning rate	0.05–0.064
Iterations	300–313

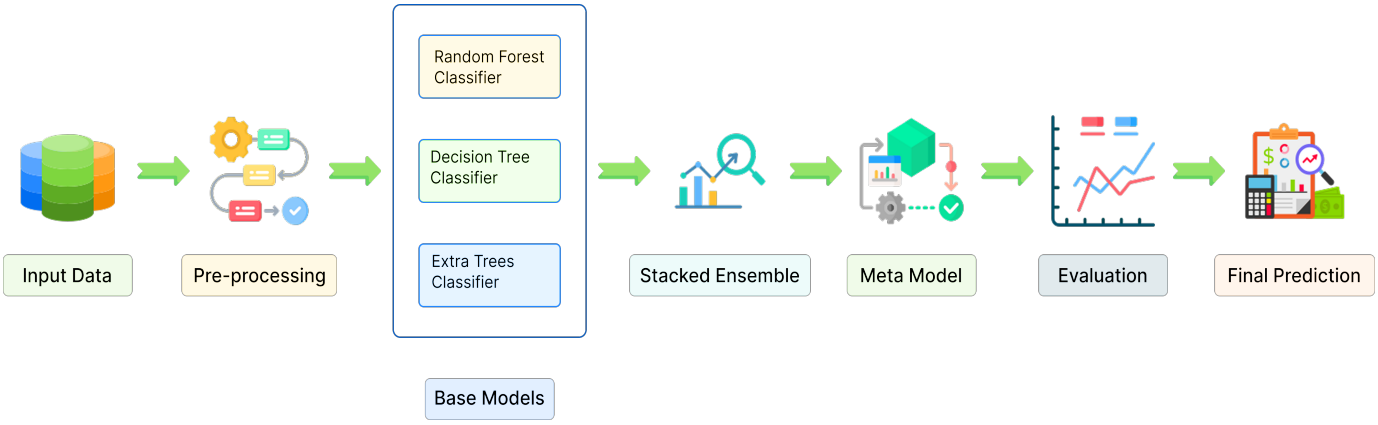

Fig. 3: Proposed Model Architecture

F. Hyperparameter Tuning

The stacking model uses three base learners (RandomForest, DecisionTree, ExtraTrees) and CatBoost as the final estimator. A compact search space targets the meta learner's key knobs depth, learning rate [0.01, 0.05, 0.1], and iterations. RandomizedSearchCV samples 20 combinations with 3 fold stratified cross validation, refits the best configuration, and then evaluates it on the test set with accuracy, a full classification report, a confusion matrix, external stratified CV, and multiclass ROC AUC with one vs rest curves. This gives a fast, reliable way to locate a strong CatBoost stack without exhaustively trying every setting.

The same stack is optimized by letting Optuna propose CatBoost depth (3–10), learning rate (0.01 – 0.2, log scale), and iterations (200–500). Each trial builds the stack and scores it via 3 fold stratified cross validation; after 20 trials, the best parameters are fixed and the model is retrained on the full train data. Final evaluation mirrors the first approach test accuracy, classification report, confusion matrix, and multiclass ROC AUC with per class curves—yielding a second, independent tuning pass that confirms and often refines the best settings found by random search.

IV. Result Analysis and Discussion

The result analysis for the diabetes detection, several models were applied. The result of these models is evaluated using several metrics such as accuracy, precision, recall, F1 score, and AUC. The model with the highest performance is chosen according to these metrics.

The baseline comparison covered eight diverse classifiers and showed that ExtraTrees delivered the strongest single model performance with accuracy 0.9671 in Table II and macro precision 0.9682 on the test split, while RandomForest and DecisionTree followed closely at 0.9570 and 0.9352 accuracy, respectively. Classwise scores for ExtraTrees were solid, with ROC–AUC 0.9968, indicating both high discrimination and stable internal validation. Based on these results, the top three models selected as base learners for stacking were ExtraTrees, RandomForest, and DecisionTree, providing complementary strengths from deep ensembles and a simpler tree.

TABLE II: Performance Comparison of Different ML Model

Model	Accuracy	Precision	Recall	F1-Score
Perceptron	0.4123	0.5395	0.4123	0.3156
Ridge	0.5137	0.4989	0.5137	0.4895
AdaBoost	0.5236	0.5147	0.5236	0.5149
Light GBM	0.5963	0.5968	0.5963	0.5951
CatBoost	0.6944	0.6951	0.6944	0.6944
Decision Tree	0.9352	0.9404	0.9352	0.9342
Random Forest	0.9578	0.9570	0.9578	0.9568
Extra Trees	0.9671	0.9682	0.9671	0.9670
Proposed	**0.9887**	**0.9800**	**0.9800**	**0.9800**

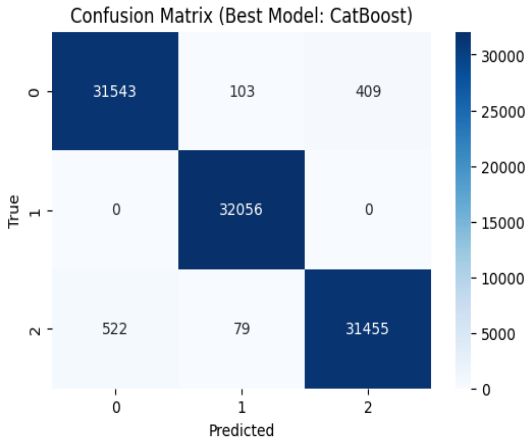

Fig. 4: Confusion Matrix of Proposed Model

The stacking ensemble fixed those three models as the base layer and rotated several candidates as the meta learner, with CatBoost emerging best at 0.9884 test accuracy, surpassing the single model baseline by over two percentage points. The ensemble achieved macro one vs rest ROC–AUC 0.9977. 5 fold CV accuracy 0.9801 ± 0.0004, and near symmetric per class performance (class 0: 0.98/0.98/0.98; class 1: 0.99/1.00/1.00; class 2: 0.99/0.98/0.98 for precision/recall/F1), showing improved

recall in the harder classes without sacrificing overall balance in Fig. 4. These gains suggest the meta learner effectively reconciles different error patterns from the base models and generalizes better on the held out set.

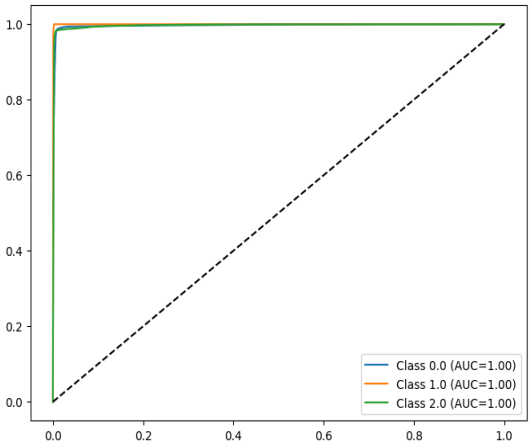

Fig. 5: ROC curve of Proposed Model.

RandomSearchCV is the best tuner with 0.9887 test accuracy, it edged out Optuna with a slightly higher test accuracy and external cross validation score while matching macro ROC–AUC in Fig.5. Optuna converged to a very similar region and produced virtually the same results (0.9885 test accuracy; 0.9978 macro ROC–AUC), confirming robustness to the tuning method. Given the small but consistent edge in test accuracy and CV, RandomizedSearchCV is selected as the best hyperparameter tuning approach for the final model.

The proposed model is a stacking ensemble that brings together top three models as base learners, with one acting as the meta learner. After building this ensemble, RandomizedSearchCV was used to tune the CatBoost meta learner's key hyperparameters depth, learning rate, and number of iterations by randomly sampling combinations and evaluating each with cross validation to find the best settings. This approach ensures the final model is not only robust and accurate, but also finely optimized for performance as shown in the TableI. By combining the strengths of multiple algorithms and carefully tuning their combination, this model delivers highly reliable and practical results for diabetes prediction.

A. Explainable AI

Explainable AI was applied using permutation importance, SHAP, and LIME to show which factors drive the stacked model's predictions and to link global patterns with case-level reasons.

Permutation importance highlights BMI (0.61) as the strongest contributor, followed by Age (0.57) and Income (0.54), with GenHlth, Education, PhysHlth, MentHlth, Smoker, Sex, and HighChol completing the top ten, indicating that body composition, demographics, socioeconomic status, and self reported health jointly shape risk predictions shown in Fig. 6. The tight error bars suggest stable rankings across

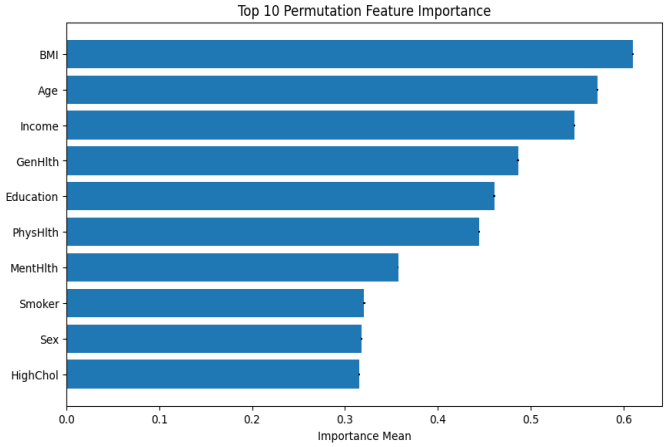

Fig. 6: Explainable AI Using Permutation Importance.

repeats, reinforcing that these variables consistently matter for the model on unseen data.

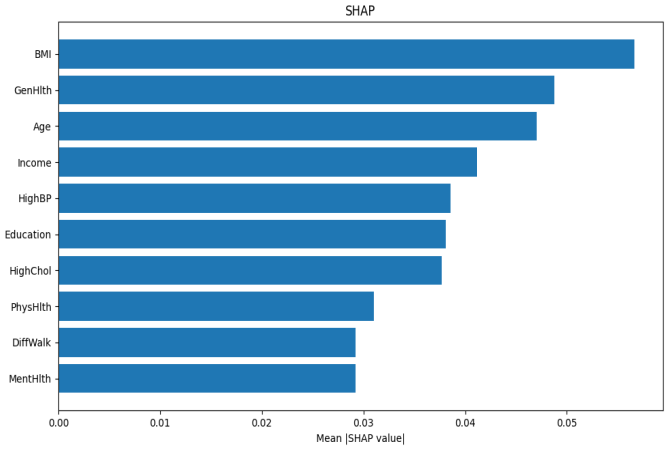

Fig. 7: Explainable AI Using SHAP.

SHAP global importance echoes this picture, placing BMI (0.056), GenHlth (0.048), Age (0.047), and Income (0.041) among the most influential features, with HighBP, Education, HighChol, PhysHlth, DiffWalk, and MentHlth also contributing meaningfully to the final probabilities shown in Fig. 7. Because SHAP aggregates marginal effects across classes, these bars capture how much each feature moves predictions on average, aligning well with the permutation ordering and supporting the overall feature hierarchy.

LIME provides a local explanation for a specific test instance, where Age (0.001), BMI (0.001), and Income (0.0008) dominate that individual decision alongside PhysActivity, GenHlth, Education, HighChol, and PhysHlth, offering a concise reason code at the point of prediction shown in Fig. 8. This local view complements the global summaries by showing which factors actually tipped the balance for a single person, improving transparency for user facing interpretation and auditability in practice.

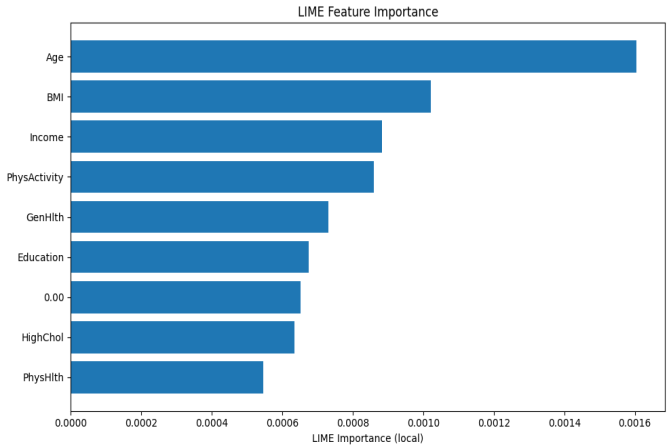

Fig. 8: Explainable AI Using LIME.

V. Conclusion and Future Works

This work proposes a comprehensive workflow for diabetes and prediabetes detection. It compares a broad set of conventional machine learning models and then builds an evidence driven stacking ensemble that uses ExtraTrees, RandomForest, and DecisionTree as base learners with CatBoost as the meta learner. After class balancing and stratified data splits, the ensemble is tuned with RandomizedSearchCV and delivers the best overall performance, with strong accuracy, ROC–AUC, and stable cross-validation. Explainable AI is used throughout: permutation importance, SHAP, and LIME consistently show BMI, Age, Income, and General Health among the most influential features, while HighBP, HighChol, PhysHlth, DiffWalk, and MentHlth add complementary insights. The proposed tuned ensemble generalizes better than single models and remains interpretable at both the global and case levels, making it well suited for population-scale screening and decision support.

Several directions can further improve on proving that the model works beyond this dataset by testing it on newer dataset and on real clinical data from different settings, while also tightening decision quality with better calibration, and cost aware thresholds that fit clinical use. To stay reliable as populations change, the roadmap includes monitoring data drift, updating the model with continual or federated learning, and using privacy preserving training to protect sensitive health information. On the modeling side, benchmarking modern tabular transformers, building patient similarity graphs, and adding time aware methods for long-term risk could capture patterns that static features miss today. Reliability and equity will be strengthened through subgroup fairness audits, bias mitigation strategies, and prevalence aware recalibration so performance remains consistent across demographic groups. Finally, the goal is to ship a lightweight API or dashboard with built in explanations, confidence intervals, and actionable risk thresholds, then pilot it prospectively in real clinics to confirm impact and guide everyday decision making.

References

[1] I. Tasin, T. U. Nabil, S. Islam, and R. Khan, "Diabetes prediction using machine learning and explainable ai techniques," *Healthcare technology letters*, vol. 10, no. 1-2, pp. 1–10, 2023.

[2] M. E. Febrian, F. X. Ferdinan, G. P. Sendani, K. M. Suryanigrum, and R. Yunanda, "Diabetes prediction using supervised machine learning," *Procedia Computer Science*, vol. 216, pp. 21–30, 2023.

[3] C.-Y. Chou, D.-Y. Hsu, and C.-H. Chou, "Predicting the onset of diabetes with machine learning methods," *Journal of Personalized Medicine*, vol. 13, no. 3, p. 406, 2023.

[4] K. K. Patro, J. P. Allam, U. Sanapala, C. K. Marpu, N. A. Samee, M. Alabdulhafith, and P. Plawiak, "An effective correlation-based data modeling framework for automatic diabetes prediction using machine and deep learning techniques," *BMC bioinformatics*, vol. 24, no. 1, p. 372, 2023.

[5] 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. 1, p. 337, 2023.

[6] H. El-Sofany, S. A. El-Seoud, O. H. Karam, Y. M. Abd El-Latif, and I. A. Taj-Eddin, "A proposed technique using machine learning for the prediction of diabetes disease through a mobile app," *International Journal of Intelligent Systems*, vol. 2024, no. 1, p. 6688934, 2024.

[7] G. Dharmarathne, T. N. Jayasinghe, M. Bogahawaththa, D. Meddage, and U. Rathnayake, "A novel machine learning approach for diagnosing diabetes with a self-explainable interface," *Healthcare analytics*, vol. 5, p. 100301, 2024.

[8] Ö. Yakut, "Diabetes prediction using colab notebook based machine learning methods," *International Journal of Computational and Experimental Science and Engineering*, vol. 9, no. 1, pp. 36–41, 2023.

[9] S. S. Bhat, M. Banu, G. A. Ansari, and V. Selvam, "A risk assessment and prediction framework for diabetes mellitus using machine learning algorithms," *Healthcare Analytics*, vol. 4, p. 100273, 2023.

[10] P. Sampath, G. Elangovan, K. Ravichandran, V. Shanmuganathan, S. Pasupathi, T. Chakrabarti, P. Chakrabarti, and M. Margala, "Robust diabetic prediction using ensemble machine learning models with synthetic minority over-sampling technique," *Scientific Reports*, vol. 14, no. 1, p. 28984, 2024.

[11] L. Jiang, Z. Xia, R. Zhu, H. Gong, J. Wang, J. Li, and L. Wang, "Diabetes risk prediction model based on community follow-up data using machine learning," *Preventive Medicine Reports*, vol. 35, p. 102358, 2023.

[12] E. K. Oikonomou and R. Khera, "Machine learning in precision diabetes care and cardiovascular risk prediction," *Cardiovascular Diabetology*, vol. 22, no. 1, p. 259, 2023.