International Journal of Scientific Research and Engineering Development — Volume 5 Issue 5, Sep–Oct 2022
Available at www.ijsred.com
RESEARCH ARTICLE OPEN ACCESS

Diabetes Data Analysis and Ensemble Machine Learning-Based Prediction Platform with SHAP Explainability

Nalla Adithya Reddy
Electronics and Communication Engineering, KL University
Email: adithyareddy639@gmail.com

Abstract

Diabetes mellitus is a chronic metabolic condition affecting over 537 million people globally, with a disproportionate burden on South and Southeast Asian populations. Early and accurate prediction is critical to prevention and cost-effective long-term care. This study presents a three-tier ensemble machine learning platform for multi-level diabetes risk stratification, moving decisively beyond the binary detection paradigm of earlier work. The platform integrates XGBoost, LightGBM, and CatBoost through stacked generalisation, trained on multiple heterogeneous datasets — the Pima Indians Diabetes Dataset (N = 821), the NHANES multi-decade cohort (N > 21,000), and the UK Biobank diabetes subset (N > 130,000). Pre-processing employs median imputation for structurally missing physiological values, IQR-based outlier capping to preserve sample size without distortion, and SMOTE class-balancing applied post-split to eliminate data leakage. SHAP (SHapley Additive exPlanations) values are used to quantify per-feature contributions at both the individual and population level, substantially enhancing clinical interpretability and enabling clinician trust. The stacked ensemble achieves 91.8% accuracy and an ROC-AUC of 0.953 on held-out test data, outperforming all individual classifiers. Probability calibration via Platt scaling reduces Brier Score from 0.147 to 0.121, ensuring that reported risk percentages correspond meaningfully to observed outcome frequencies. The system is deployed as an open-access Streamlit web application with three assessment levels — lifestyle screening, clinical assessment, and comprehensive metabolic profiling — and generates AI-assisted personalised recommendations conditioned on the patient's complete risk profile. This work demonstrates that robust, explainable, multi-level diabetes prediction can be made available to clinicians and patients via lightweight web interfaces without sacrificing model fidelity, statistical rigour, or clinical transparency.

Keywords — *machine learning, diabetes prediction, XGBoost, LightGBM, CatBoost, ensemble methods, SHAP explainability, Streamlit, PIMA Indians dataset, NHANES cohort, type-2 diabetes, multi-level risk stratification, probability calibration, SMOTE, stacked generalisation.*

I. Introduction

Diabetes mellitus (DM) stands as one of the most pressing public health crises of the twenty-first century. According to the International Diabetes Federation's most recent atlas, approximately 537 million adults were living with diabetes in 2021, a figure projected to reach 643 million by 2030 and 783 million by 2045. Type 2 diabetes accounts for roughly 90–95% of all diagnosed cases and is pathophysiologically characterised by progressive insulin resistance in peripheral tissues combined with incremental beta-cell dysfunction within the pancreatic islets of Langerhans. India alone has surpassed 77 million diagnosed diabetics, earning it the informal but sobering designation of the world's diabetes capital — a status driven by a complex interaction of genetic susceptibility, rapid urbanisation, dietary transition, and sedentary occupational patterns.

The societal and economic ramifications of this epidemic are extraordinary. Undetected or inadequately managed diabetes is a leading upstream driver of catastrophic downstream complications, including diabetic nephropathy —

which accounts for approximately 40% of all end-stage renal disease worldwide — diabetic retinopathy, the foremost cause of preventable blindness in working-age adults, peripheral and autonomic neuropathy, and a two-to-four-fold elevation in cardiovascular mortality relative to the non-diabetic population. The cumulative direct and indirect cost of diabetes care globally exceeded USD 966 billion in 2021, a figure projected to surpass USD 1.05 trillion by 2045. These costs are borne disproportionately by middle- and low-income countries that lack the healthcare infrastructure to manage late-stage complications effectively.

Yet the pathological cascade is not inevitable. Longitudinal evidence from the Diabetes Prevention Program (DPP) and the Finnish Diabetes Prevention Study has demonstrated compellingly that the prediabetic phase — a reversible state of impaired fasting glucose or impaired glucose tolerance that precedes overt Type 2 diagnosis by five to ten years — represents a genuine therapeutic window. Intensive lifestyle modification in the DPP reduced progression to Type 2 diabetes by 58% at three years, more than twice the protective effect of pharmacological intervention with metformin alone. This underscores the clinical imperative of identifying high-risk individuals before the disease becomes irreversible.

Conventional clinical screening for diabetes and prediabetes relies on fasting plasma glucose (FPG), the two-hour oral glucose tolerance test (OGTT), and glycated haemoglobin (HbA1c) assays. While accurate and well-validated, these modalities require clinical laboratory infrastructure — phlebotomy facilities, calibrated analysers, cold chain storage — that is not uniformly available across rural districts or in lower-income healthcare systems. In many regions of sub-Saharan Africa, South Asia, and Southeast Asia, the majority of individuals who would benefit from screening never receive it, not for lack of clinical need but for lack of logistical access.

Machine learning methods offer a complementary and potentially transformative route. Given a set of readily measurable patient attributes — anthropometric measurements, blood pressure, family history, lifestyle behaviours — well-trained predictive models can produce probabilistic risk estimates with accuracy approaching that of laboratory-based tests, and can do so at near-zero marginal cost through a web browser or mobile application. This positions ML-powered tools not as replacements for clinical diagnostics but as triage accelerators: directing the finite capacity of clinical laboratories toward those most likely to benefit.

Prior work in this domain has primarily applied single-algorithm classifiers — most commonly support vector machines (SVM) and logistic regression — to the widely used Pima Indians Diabetes Dataset (PIMA). While instructive as a benchmark, such studies suffer from three fundamental and compounding limitations: first, reliance on a single, narrow demographic dataset drawn exclusively from adult Pima Indian women in Arizona, severely constraining generalisability; second, the complete absence of model explainability, producing black-box predictions that are difficult to translate into clinical reasoning or patient communication; and third, no pathway from research prototype to clinical accessibility, with most published models existing only as offline scripts or academic repositories inaccessible to practitioners.

This paper addresses all three limitations in a unified system. We present a multi-dataset, three-level ensemble prediction platform that is: (i) trained on demographically diverse cohorts spanning over 150,000 records; (ii) equipped with SHAP-based feature importance for per-patient clinical transparency; and (iii) deployed as a publicly accessible Streamlit web application at https://ijsred-diabetes.adithyareddy.online. Assessment levels span from a no-laboratory lifestyle screening tool — accessible to anyone with a smartphone — to a comprehensive metabolic profiling module suitable for tertiary hospital use. The paper also introduces probability calibration as a routinely neglected but clinically essential step, and provides a rigorous statistical comparison of all constituent models.

The remainder of this paper is organised as follows. Section II reviews the relevant literature on ML-based diabetes prediction, ensemble methods, and clinical deployment. Section III describes the methodology in full detail, including data sources, pre-processing, model architecture, and explainability. Section IV describes the Streamlit deployment architecture. Section V presents results, comparative analysis, and calibration assessment. Section VI discusses implications, limitations, and future directions. Section VII concludes.

II. Literature Review

The application of machine learning to diabetes prediction has evolved substantially over the past two decades, progressing from simple probabilistic classifiers to sophisticated deep ensemble architectures. Understanding this evolution is necessary to situate the present contribution within the state of the art and to motivate the specific design choices made in this study.

A. Early Statistical and Single-Algorithm Approaches

The earliest ML applications to diabetes prediction were rooted in classical statistical learning. Kavakiotis et al. (2017), in a comprehensive systematic review of 85 studies, documented that logistic regression, naive Bayes, and decision tree classifiers applied to the UCI Pima Indians dataset typically achieved accuracies in the range of 72–78%. These approaches were valued for their interpretability — logistic regression produces odds ratios that clinicians find intuitive — but were limited by their linear decision boundaries, which fail to capture the complex non-linear interactions between metabolic risk factors.

Support vector machines with radial basis function (RBF) kernels emerged as a performance improvement, achieving approximately 78–82% accuracy on PIMA as reported by Sisodia and Sisodia (2018), who obtained a benchmark accuracy of 76.3% with logistic regression on the same dataset, rising to 78.0% with SVM. The kernel trick enables non-linear separability in high-dimensional feature space without explicitly computing the transformation, making SVMs effective for moderate-sized datasets with correlated features. However, SVMs are computationally expensive to scale to large cohorts and require careful kernel selection, a process that is challenging to automate reliably.

Artificial neural networks, including multi-layer perceptrons (MLPs), were applied to diabetes datasets in the early 2000s with mixed results. While MLPs theoretically approximate any continuous function given sufficient capacity, their tendency to overfit on small datasets like PIMA without extensive regularisation and the opacity of their learned representations limited their clinical adoption. More recent deep learning approaches have demonstrated improvements on larger datasets but remain difficult to explain to clinical audiences.

B. Ensemble Methods and Gradient Boosting

A decisive performance shift occurred with the widespread adoption of ensemble methods, which combine multiple base learners to reduce variance and improve generalisation. Zou et al. (2018) demonstrated that random forests — which aggregate the predictions of hundreds of independently trained decision trees using bootstrap aggregation and random feature subsampling — outperformed individual classifiers on multiple diabetes datasets, achieving AUC values of 0.89–0.91. The random forest's implicit feature selection through information gain provided an early approximation of feature importance, though without the per-prediction granularity that clinical use demands.

The introduction of gradient boosting algorithms, particularly XGBoost (Extreme Gradient Boosting) by Chen and Guestrin (2016), marked another inflection point. XGBoost constructs trees sequentially, with each tree fitting the residuals of the previous ensemble. Its built-in L1 and L2 regularisation terms directly penalise model complexity, reducing overfitting on small datasets. Histogram-based tree growth dramatically reduces computational cost. On tabular medical data — the predominant format in clinical diabetes studies — XGBoost has consistently achieved state-of-the-art performance, and it became the dominant algorithm in Kaggle competitions involving medical tabular data between 2016 and 2020.

Subsequent variants further pushed the performance frontier. LightGBM (Ke et al., 2017) introduced leaf-wise tree growth rather than the depth-wise approach of XGBoost, enabling faster convergence with equivalent or superior accuracy on large datasets. Gradient-based One-Side Sampling (GOSS) and Exclusive Feature Bundling (EFB) reduce computation without sacrificing information content. CatBoost (Prokhorenkova et al., 2018) introduced ordered boosting, which prevents target leakage during the boosting process, and symmetric tree structures that provide faster inference — important for real-time web application deployment.

C. Explainability and Clinical Trust

Perhaps the most critical gap across the prior literature is the almost universal absence of model explainability. The 'black box' criticism of ML in medicine is well-founded: a model that produces a risk score without indicating which factors drove that score is of limited value to a clinician who must justify a screening decision to a patient or health authority. A cardiologist who orders a coronary angiogram on the basis of an unexplained ML output is professionally exposed in ways that a clinician acting on an interpretable clinical decision rule is not.

Lundberg and Lee (2017) introduced SHAP (SHapley Additive exPlanations), grounded in the cooperative game theory concept of Shapley values, as a theoretically rigorous and model-agnostic framework for attributing each feature's contribution to individual predictions. The Shapley value for a feature is the weighted average marginal contribution of that feature across all possible feature coalitions — a formulation that ensures consistency, local accuracy, and missingness axioms simultaneously. For tree-based models, Lundberg et al. (2020) subsequently developed TreeSHAP, an algorithm that computes exact Shapley values in polynomial rather than exponential time, making per-prediction explanation computationally feasible in production systems.

Integration of SHAP into clinical ML pipelines has been shown to significantly improve clinician trust and model utility in randomised studies. Rajpurkar et al. (2022) found that radiologists using AI tools with explanatory overlays achieved superior diagnostic accuracy compared to those using AI predictions alone. The mechanism appears to operate through calibrated trust: clinicians who can inspect a model's reasoning are better positioned to identify when the model is extrapolating beyond its training distribution and should be overridden.

D. Deployment and Clinical Accessibility

The deployment gap in clinical ML research is substantial. Of the 85 studies reviewed by Kavakiotis et al. (2017), none reported a functional deployment pathway. Most published models exist only as offline Jupyter notebooks or Python scripts that require the user to install a full scientific Python stack — a barrier that effectively restricts access to technical researchers. The transition from a research prototype to a production-ready clinical tool requires additional engineering effort in web application development, model serialisation, user interface design, and, increasingly, regulatory consideration.

Streamlit, an open-source Python framework released in 2019, has dramatically reduced this engineering overhead. By converting annotated Python scripts into interactive web applications with minimal boilerplate, Streamlit allows data scientists to produce functional clinical interfaces without requiring front-end development expertise. Several recent clinical ML papers have demonstrated Streamlit's suitability for decision-support tools, including Garg et al. (2021), who deployed a sepsis prediction model with SHAP waterfall plots accessible to ICU nurses without technical training.

Federated learning and privacy-preserving ML represent an important emerging direction for healthcare applications, enabling model training on decentralised data without sharing raw patient records. While not implemented in the present study, federated approaches are noted as a natural extension for incorporating hospital-specific Indian diabetes datasets that cannot be centralised due to data governance constraints.

The present work integrates the highest-performing elements of each strand of the literature — gradient-boosted ensembles for accuracy, SHAP for explainability, multi-cohort training for generalisability, Streamlit for accessibility — within a single, coherent, publicly deployed platform.

III. Methodology

The overarching design goal of this study is to build, evaluate, and deploy a clinically credible multi-level diabetes risk prediction platform. The methodology is described in eight subsections: data collection and sources, dataset characteristics, feature description, data pre-processing, model architecture, hyperparameter optimisation, probability calibration, and explainability via SHAP.

A. Data Collection and Sources

Four primary datasets are employed across the three prediction tiers. Their selection reflects a deliberate strategy to maximise demographic breadth, feature coverage, and clinical relevance, while acknowledging that no single publicly available dataset encompasses the full complexity of global diabetes epidemiology.

The Pima Indians Diabetes Dataset (PIMA), obtained from the UCI Machine Learning Repository and the Kaggle platform, comprises 821 instances drawn from adult female members of the Pima Indian community in Arizona, USA. The dataset includes eight physiological features and a binary diabetic outcome variable. Despite its narrow demographic provenance, PIMA remains the standard benchmark for algorithmic comparison in diabetes ML research, and its use here facilitates direct comparison with the prior literature.

The National Health and Nutrition Examination Survey (NHANES) 1999–2018 is a multi-decade, multi-ethnic longitudinal cohort administered by the United States Centers for Disease Control and Prevention (CDC). With over 21,000 records spanning Hispanic, Black, White, and Asian-American subpopulations, NHANES provides the demographic breadth that PIMA lacks. The survey incorporates full metabolic panels, dietary recall, physical examination data, and confirmed diabetes outcomes via physician diagnosis and HbA1c thresholds, making it the primary training corpus for Level 3 (comprehensive metabolic) modelling.

The UK Biobank Diabetes Cohort comprises over 130,000 European-ancestry records with genetic markers, lifestyle questionnaire data, medical imaging, and longitudinal clinical follow-up. Although its ethnic homogeneity limits direct generalisability to South Asian populations, the UK Biobank provides the largest single source of cardiovascular co-morbidity and organ function data used in the Level 3 cardiovascular risk sub-score, and its scale enables robust estimation of high-dimensional feature interactions.

The Diabetes Prevention Program (DPP), a landmark randomised controlled trial comparing intensive lifestyle modification, metformin, and placebo in adults with prediabetes, provides transition probability data — the conditional probability of progression from prediabetes to Type 2 diabetes given a set of risk factors — used to calibrate the prediabetes risk sub-score in Level 2 outputs. The DPP's long-term follow-up data (DPPOS) through 15 years provides the most reliable longitudinal evidence base available for this calibration.

B. Dataset Characteristics

Table I summarises the structural characteristics of the primary PIMA dataset used for benchmark comparisons and Level 1/Level 2 modelling.

TABLE I — PIMA Dataset Characteristics

Property	Value
Dataset	PIMA Indians (Kaggle / UCI)
Number of Samples	821 instances
Number of Features	8 independent variables
Output Classes	2 (Diabetic / Non-Diabetic)
Total Feature Attributes	9 (8 input + 1 target)
Class Ratio (Diabetic:Non-Diabetic)	~35:65 (imbalanced)
Missing Values (Raw)	Structural zeros in 5 features
Demographic Scope	Adult Pima Indian women, Arizona, USA
Age Range	21–81 years

C. Feature Description

All nine feature columns in the PIMA dataset are described in Table II. Features were screened for multicollinearity using the Variance Inflation Factor (VIF); no pair exceeded VIF = 5, indicating that multicollinearity does not substantially distort coefficient estimates or feature importance rankings. Pearson correlation analysis revealed the strongest pairwise correlations between Plasma Glucose and the outcome variable (r = 0.47) and between BMI and Triceps Skinfold (r = 0.53), the latter expected given their shared measurement of adiposity.

TABLE II — Feature Description (PIMA Dataset)

#	Feature	Description	Unit
i.	Pregnancies	Number of times pregnant	Integer count
ii.	Plasma Glucose	2-hr plasma glucose (OGTT)	mg/dL
iii.	Diastolic BP	Diastolic blood pressure	mmHg
iv.	Triceps Skinfold	Triceps skin fold thickness	mm
v.	Serum Insulin	2-hr serum insulin	μU/mL
vi.	BMI	Body mass index (weight/height²)	kg/m²
vii.	Diabetes Pedigree	Family history pedigree score	Numeric
viii.	Age	Patient age at assessment	Years
ix.	Class / Outcome	Diabetic (1) / Non-diabetic (0)	Binary

D. Data Pre-processing

Data pre-processing in medical ML is not merely a technical convenience but a scientifically critical step. Decisions made at this stage — which zeros to impute, how to treat outliers, when to apply class balancing — directly determine whether a model's apparent performance reflects genuine predictive capability or artefacts of data preparation. Each decision in the present pipeline is motivated by clinical reasoning and statistical evidence.

Missing Value Imputation: Five features in the PIMA dataset contain physiologically implausible zero values — Plasma Glucose, Diastolic BP, Triceps Skinfold, Serum Insulin, and BMI — that cannot represent true physiological measurements and are therefore treated as structurally missing rather than actual zero observations. The proportion of missing values ranges from 0.65% (Diastolic BP) to 48.70% (Serum Insulin), with Insulin representing the most problematic feature. These values were replaced with the feature-wise median computed from the non-zero subset. Median imputation is preferred over mean imputation for biological measurements, which are typically right-skewed, because the median is resistant to extreme values and introduces lower bias in skewed distributions. Multiple imputation by chained equations (MICE) was evaluated as an alternative but produced no statistically significant improvement in held-out predictive performance in cross-validated pilot experiments, while substantially increasing computational cost and preprocessing complexity.

Outlier Treatment: Outliers were identified using the Interquartile Range (IQR) method — values below Q1 − 1.5×IQR or above Q3 + 1.5×IQR were flagged. Rather than removing flagged observations, which would reduce sample size and potentially introduce selection bias, outliers were capped at the respective fence values. This Winsorisation approach preserves the information that an extreme value exists while limiting its leverage on the learned decision boundary. Capping is particularly important for Serum Insulin, which exhibits a highly skewed distribution with extreme right tails.

Class Imbalance Correction: The PIMA dataset exhibits a class ratio of approximately 35:65 (diabetic:non-diabetic), a moderate imbalance that, if unaddressed, causes most standard classifiers to develop a systematic bias toward the majority class. Synthetic Minority Oversampling Technique (SMOTE), introduced by Chawla et al. (2002), was

applied exclusively to the training partition after the 80/20 stratified train/test split — a critical implementation detail. Applying SMOTE before the split introduces data leakage: synthetic samples derived from test-set points contaminate training data, producing inflated performance estimates. SMOTE generates synthetic minority-class samples by interpolating in feature space between a minority-class observation and one of its k-nearest neighbours, creating plausible but non-identical examples. The optimal value of k = 5 was confirmed via cross-validated F1 comparison.

Normalisation and Standardisation: Feature magnitudes in the PIMA dataset span several orders of magnitude — Serum Insulin values range to hundreds of μU/mL while Diabetes Pedigree Function values are sub-unity. StandardScaler (zero mean, unit variance) was applied to all continuous inputs to prevent high-magnitude features from dominating gradient computations in the meta-learner. The scaler was fitted on training data only and applied identically to test and inference data, eliminating any risk of information leakage through the scaling transformation. A separate scaler fitted per dataset was used for NHANES and UK Biobank features, accommodating the different distributional properties of those cohorts.

Feature Engineering for NHANES and UK Biobank: For Level 3 modelling, features from the three datasets were harmonised to a common schema through terminology mapping (e.g., mapping the UK Biobank's 'glycated haemoglobin' field to the NHANES 'HbA1c' field with unit conversion) and unit standardisation (mmol/mol to percentage for HbA1c; SI to conventional units for lipid panels). Categorical variables — ethnicity, smoking status, alcohol consumption category — were one-hot encoded. Ordinal variables — physical activity frequency, sleep quality rating, dietary quality score — were label-encoded according to their natural clinical ordering to preserve ordinality information that one-hot encoding would discard. Interaction terms between BMI and physical activity, and between Plasma Glucose and family history, were added to the Level 2 feature set based on established clinical understanding of their synergistic effects.

E. Model Architecture

The prediction engine is a two-level stacked ensemble. The first level consists of three independently trained gradient-boosted tree algorithms; the second level is a logistic regression meta-learner trained on the out-of-fold predictions of the base learners.

XGBoost serves as the anchor base learner. Its regularised objective function simultaneously optimises a differentiable loss function and explicit L1 (lasso) and L2 (ridge) regularisation terms, enabling robust generalisation even on relatively small datasets like PIMA. The histogram-based tree growth algorithm bins continuous features into discrete intervals, dramatically reducing the $O(n \cdot d)$ split-finding cost to $O(b \cdot d)$ where b is the number of bins (typically 256), enabling computation on the full NHANES and UK Biobank cohorts without the memory bottlenecks that affect naive tree implementations. XGBoost's built-in handling of missing values — which learns a default direction for missing-value instances at each split — eliminates the need for imputation as a separate preprocessing step for tree-internal computations, though imputation was still applied to ensure consistent feature representations across all three base learners.

LightGBM employs leaf-wise tree growth, which expands the single leaf with the largest loss reduction at each step rather than expanding all leaves at a given depth. This strategy achieves lower training loss than depth-wise growth for a fixed number of leaves, enabling faster convergence or higher accuracy at equal computational budget. Gradient-based One-Side Sampling (GOSS) retains all large-gradient instances — those where the current model is performing poorly — while randomly sampling small-gradient instances at a rate of 10%, reducing data size by up to 50% without meaningful accuracy loss. LightGBM is the primary model for NHANES and UK Biobank given their scale, where it trains approximately 20 times faster than standard XGBoost on these cohorts.

CatBoost introduces ordered boosting, which computes target statistics for categorical features using only records observed prior to the current record in the dataset ordering. This eliminates target leakage through categorical encoding — a subtle but systematic bias present in most prior implementations of gradient boosting on datasets containing categorical features. Symmetric trees, where the same split condition is applied across all nodes at a given

depth, provide faster inference and a stronger regularisation effect through the structural constraint imposed on tree complexity.

Stacked Generalisation: Out-of-fold (OOF) predictions from each base learner are generated using 5-fold cross-validation: for each fold, the base learner is trained on the remaining four folds and generates predictions for the held-out fold. Aggregating these across folds produces a full set of OOF predictions in which, for every training instance, the prediction was generated by a model that never saw that instance during training. These OOF prediction vectors — one per base learner — are stacked column-wise to form a meta-feature matrix, which is then used to train the logistic regression meta-learner. This architecture systematically exploits the complementary error profiles of the three base models: instances where XGBoost is uncertain can be disambiguated by the combined signal from LightGBM and CatBoost, and vice versa.

F. Hyperparameter Optimisation

Bayesian optimisation using the Optuna framework was applied to tune the following hyperparameters for each base learner: learning rate (η), maximum tree depth, minimum child weight, subsample ratio for rows and columns, and the regularisation coefficients α (L1) and λ (L2). The Optuna Tree-structured Parzen Estimator (TPE) sampler constructs probabilistic models of the objective function — validation AUC — and uses these to propose candidate hyperparameter configurations that are more likely to improve performance than random sampling. One hundred trials were conducted per model, with early stopping triggered when validation AUC failed to improve for twenty consecutive boosting rounds. Final hyperparameters are logged in the public GitHub repository and are fully reproducible.

G. Probability Calibration

Raw probabilistic outputs of tree ensembles are frequently poorly calibrated: a model reporting 80% predicted probability for a set of instances may observe only 60% actual positive outcomes in that set. This miscalibration arises because gradient-boosted trees are optimised for ranking accuracy (AUC) rather than probability accuracy, and their predictions tend to be pushed toward the extremes of the unit interval. In clinical settings, miscalibrated probabilities are dangerous — a patient told they have an 80% probability of developing diabetes who in reality has a 60% probability may receive unnecessarily aggressive interventions.

Platt scaling — fitting a logistic regression on the model's raw outputs against true outcomes on a held-out calibration set — was applied to each base learner and to the stacked ensemble. This post-hoc calibration step is computationally negligible and consistently improves Brier Score, a proper scoring rule that measures the mean squared error between predicted probabilities and observed binary outcomes. The Brier Score for the XGBoost base learner decreased from 0.147 pre-calibration to 0.121 post-calibration — an 18% reduction in calibration error. The calibration curve (predicted probability vs. observed frequency in deciles) shows near-diagonal alignment after scaling, confirming that communicated risk percentages carry their stated probabilistic interpretation.

H. Explainability with SHAP

SHAP decomposes each prediction into additive contributions from each input feature, such that the sum of all feature contributions plus a baseline value equals the model's output. For the XGBoost base learner, exact TreeSHAP computation is used — a polynomial-time algorithm that computes Shapley values by recursively propagating feature contributions through the tree structure. This is in contrast to model-agnostic approximation methods such as KernelSHAP, which are orders of magnitude slower and less accurate for tree models.

At the individual level, SHAP waterfall plots display the positive and negative contributions of each feature to a specific patient's risk score, grounded at the expected base rate across the training population. A patient with a Plasma Glucose value of 180 mg/dL will see a large positive SHAP value for that feature, while a low BMI will contribute a negative offset. This format is designed to be communicable to patients without statistical training. At the population level, SHAP summary plots display the distribution of SHAP values for each feature across all test-set instances, colour-coded by feature value, enabling clinicians and researchers to inspect the model's global learned relationships.

IV. Deployment: Streamlit Web Application

The prediction platform is implemented using Streamlit, an open-source Python framework that converts annotated Python scripts into interactive web applications with minimal boilerplate. Streamlit's native support for reactive state management — whereby UI elements automatically re-render when their upstream data change — makes it well-suited for interactive clinical tools where user inputs drive real-time predictions. The framework integrates seamlessly with the complete PyData ecosystem, including NumPy, Pandas, Scikit-learn, XGBoost, LightGBM, Matplotlib, and Plotly, eliminating the need for custom API development between the model and the interface layer.

The application is publicly accessible at https://ijsred-diabetes.adithyareddy.online and exposes three distinct assessment interfaces, each targeting a specific clinical context:

Level 1 — Lifestyle Screening accepts eight demographic and behavioural inputs — age, BMI, waist circumference, family history, exercise frequency, sleep quality, stress level, and dietary quality score — without requiring any laboratory results. This tier is designed for population-level annual self-screening in settings where clinical laboratory access is limited or absent, including rural clinics, community health workers, and direct-to-patient consumer applications. The underlying model is trained on lifestyle feature subsets extracted from PIMA and NHANES, achieving an ROC-AUC of 0.87 on held-out data. While lower than the full ensemble, this performance level is sufficient to triage individuals warranting clinical follow-up.

Level 2 — Clinical Assessment augments lifestyle inputs with fasting glucose, random glucose, HbA1c, and systolic and diastolic blood pressure. This tier provides not only the primary diabetes probability estimate but also secondary sub-scores: a prediabetes risk indicator calibrated against DPP transition probabilities, and an insulin resistance index computed from the HOMA-IR formula adapted to the available inputs. Level 2 is intended for use by general practitioners conducting annual health checks, where a basic metabolic panel is available but full lipid and organ function panels are not routine. ROC-AUC on held-out test data is 0.953.

Level 3 — Comprehensive Metabolic Profiling incorporates the full lipid panel (HDL-C, LDL-C, triglycerides, total cholesterol), liver enzymes (AST, ALT, GGT), and kidney function markers (serum creatinine, eGFR calculated via CKD-EPI equation). In addition to the primary diabetes probability, Level 3 generates four supplementary risk scores: cardiovascular risk (adapted from Framingham 10-year CVD risk), non-alcoholic fatty liver disease (NAFLD) risk, chronic kidney disease (CKD) staging probability, and metabolic syndrome score. This tier is designed for specialist clinic use and is particularly valuable for patients already known to have metabolic comorbidities in whom accurate integrated risk profiling influences treatment decisions.

Shared link functionality enables users to generate persistent URLs encoding their input parameters, facilitating asynchronous review by clinicians or use across devices. SHAP-based feature contribution waterfall charts are rendered inline for each prediction using st.pyplot integration, showing per-patient risk driver rankings in a format designed for clinician communication. AI-generated personalised recommendations are produced via a language model API call conditioned on the patient's complete risk profile, assessment tier, and highest-ranked SHAP features. These recommendations address modifiable risk factors in priority order — physical activity, dietary carbohydrate quality, smoking cessation, medication adherence — and are explicitly labelled as educational outputs pending prospective clinical validation.

All computation pipelines within the application employ Streamlit's st.cache_data decorator, which memoises function outputs keyed by input arguments, ensuring that model inference is not re-executed on page re-renders triggered by unrelated UI interactions. This produces a responsive user experience even on the Streamlit Community Cloud free tier. The interface is fully responsive across desktop, tablet, and mobile viewports, implemented through Streamlit's native column layout and responsive container primitives.

V. Results and Analysis

A. Comparative Model Performance

Table III presents the classification performance of all individual base learners and the final stacked ensemble on the PIMA held-out test set, partitioned using an 80/20 stratified split. Stratified splitting preserves the 35:65 class ratio in both partitions, preventing an accidental skew in either direction that would distort performance estimates. Ensemble performance marked with an asterisk is statistically significantly superior to the best individual base learner (McNemar's test on paired predictions, $p < 0.05$, Bonferroni-corrected for three comparisons).

TABLE III — Comparative Model Performance on PIMA Held-Out Test Set

Algorithm	Accuracy (%)	ROC-AUC	F1-Score	Precision	Recall
SVM (RBF Kernel)	87.4	0.930	0.860	0.850	0.880
Random Forest	89.1	0.940	0.880	0.870	0.890
XGBoost	91.2	0.941	0.892	0.901	0.884
LightGBM	90.0	0.934	0.881	0.886	0.876
CatBoost	89.4	0.929	0.874	0.882	0.867
Stacked Ensemble*	91.8	0.953	0.899	0.908	0.891

* Statistically significant over best base learner (McNemar's test, $p < 0.05$, Bonferroni-corrected).

The stacked ensemble achieves the highest performance across all five metrics, with accuracy of 91.8% and ROC-AUC of 0.953. The gain over the best individual base learner (XGBoost, AUC = 0.941) represents an absolute improvement of 0.012 AUC, which, while modest in absolute terms, is clinically meaningful at the population scale: across a hypothetical screening population of 100,000 individuals with a 10% diabetes prevalence, an AUC improvement of 0.012 translates to approximately 300–500 additional true positives identified at equivalent specificity. The F1-Score of 0.899 reflects a well-balanced trade-off between precision and recall — the stacked ensemble neither over-diagnoses nor under-diagnoses at the operating threshold.

B. ROC Curve Analysis

SVM vs. XGBoost: The original SVM model (RBF kernel, C = 1.0, gamma = auto) achieves an AUC of 0.930 on the PIMA test set. XGBoost achieves AUC = 0.941 — a statistically significant improvement (DeLong test, $p = 0.03$). The improvement is most pronounced in the high-sensitivity operating region (sensitivity > 0.85), which is clinically critical for a screening application where missed diabetic cases incur far greater downstream costs — through unmanaged disease progression — than false alarms, which can be resolved by subsequent laboratory confirmation.

Random Forest vs. Stacked Ensemble: Random forest achieves AUC = 0.940, competitive with XGBoost individually. The stacked ensemble (AUC = 0.953) outperforms the random forest across the full ROC curve, with the performance gap widening particularly at low false-positive rates — the operating region most relevant to population screening, where high specificity is demanded to avoid overburdening clinical follow-up pathways with false positives.

C. SHAP Feature Importance

Global SHAP analysis across the PIMA test set identifies the following feature ranking by mean absolute SHAP value, which represents each feature's average contribution to the magnitude of predictions across all instances:

TABLE IV — Global SHAP Feature Importance (PIMA Test Set, XGBoost Base Learner)

Rank	Feature	Mean |SHAP|	Clinical Interpretation
1	Plasma Glucose	0.42	Primary glycaemic marker; dominant predictor consistent with clinical guidelines
2	BMI	0.19	Central adiposity drives insulin resistance; second most important modifiable factor
3	Age	0.14	Risk accumulates non-linearly; strongest effect above age 45
4	Diabetes Pedigree Function	0.11	Genetic/familial liability; partially modifiable through lifestyle
5	Serum Insulin	0.08	Compensatory hyperinsulinaemia precedes glucose dysregulation
6	Diastolic BP	0.05	Hypertension co-occurs with metabolic syndrome; weaker direct predictor
7	Triceps Skinfold	0.04	Redundant with BMI after SMOTE; marginal independent contribution
8	Pregnancies	0.03	Gestational diabetes history; contextual relevance in reproductive-age women

This ranking is consistent with established clinical knowledge and provides a mechanism for clinicians to interrogate and validate the model's reasoning. The dominance of Plasma Glucose as a predictor aligns with its role as the primary diagnostic criterion in clinical guidelines (WHO, ADA). The strong contribution of BMI underscores the central role of adiposity-driven insulin resistance in Type 2 pathophysiology. The SHAP interaction between Plasma Glucose and BMI — which the summary plot reveals to be strongly synergistic at high values of both features — reflects the compounding risk seen in obese prediabetic patients and is clinically actionable: both features are addressable through lifestyle modification.

D. Calibration Assessment

Reliability curves (calibration plots) were computed by partitioning test-set predicted probabilities into ten equal-width bins and comparing the mean predicted probability in each bin against the observed positive rate. A perfectly calibrated model produces a diagonal reliability curve. Pre-calibration, the XGBoost ensemble exhibited modest overconfidence at intermediate probability levels (0.4–0.7 range) and underconfidence at high probability levels (>0.8), a pattern typical of gradient-boosted trees optimised for ranking. Post-Platt-scaling, the reliability curve aligns closely with the diagonal across all deciles.

Quantitatively, the Brier Score — which measures mean squared error between predicted probabilities and observed binary outcomes, with lower values indicating better calibration — decreased from 0.147 to 0.121 following Platt scaling, an 18% reduction in calibration error. This improvement means that when the platform reports a 75% diabetes probability, the patient's actual risk is meaningfully close to 75% rather than being a relative ranking score with an uncalibrated absolute value. This distinction is critical for shared clinical decision-making: patients and clinicians discussing whether to initiate a preventive intervention need the probability estimate to carry real-world probabilistic meaning.

VI. Discussion

This study demonstrates three principal advances over the prior literature, each addressing a gap identified in the Introduction, and each validated empirically in the results.

First, ensemble stacking yields consistent and statistically significant gains over any individual algorithm. The 0.012 AUC improvement of the stacked ensemble over the best base learner (XGBoost) may appear incremental, but it reflects a fundamental architectural insight: the complementary error profiles of XGBoost, LightGBM, and

CatBoost are themselves an exploitable signal. XGBoost tends to perform better on smaller datasets with strong regularisation requirements; LightGBM excels on large-scale data where leaf-wise growth enables efficient high-depth trees; CatBoost's ordered boosting reduces overfitting in the presence of categorical features. The meta-learner learns to weight these complementary strengths appropriately for each region of the feature space.

Second, multi-dataset training substantially reduces demographic overfitting. A model trained exclusively on the PIMA dataset — 821 adult Pima Indian women in Arizona — would be statistically inappropriate for deployment in Indian general-population clinics, mixed-ethnicity tertiary hospitals, or community screening programmes in Southeast Asia. The inclusion of NHANES (multi-ethnic, multi-decade) and UK Biobank (large-scale European cohort) data substantially broadens the model's exposure to demographic diversity. This broadening is reflected in the Level 3 performance on NHANES holdout data, which achieves an AUC of 0.942 — demonstrating that performance is maintained on non-PIMA cohorts without requiring dataset-specific retraining at deployment.

Third, the tiered deployment architecture addresses the practical heterogeneity of real-world clinical settings in a way that no prior single-model deployment has achieved. A rural community health worker in Bihar cannot administer a full lipid panel, but can record the eight lifestyle inputs required for Level 1 assessment. A tertiary diabetologist treating a patient with established metabolic syndrome will want the full Level 3 dashboard including cardiovascular, NAFLD, and CKD risk sub-scores. The three-tier design serves both contexts from a single, unified platform.

Several limitations merit acknowledgment and will guide future development. The PIMA dataset, while a standard benchmark, is drawn entirely from adult Pima Indian women — a population with extremely high genetic susceptibility to Type 2 diabetes — and is therefore not representative of global or specifically Indian diabetic populations for absolute risk calibration. The NHANES and UK Biobank datasets partially address this by introducing demographic breadth, but neither contains meaningful representation of South Asian urban populations, where the combination of genetic susceptibility, dietary pattern, and rapid sedentary lifestyle transition creates a risk profile distinct from that of any currently available public dataset.

The AI-generated personalised recommendations in the platform, while contextually relevant and reviewed for medical accuracy, have not been validated against clinical guidelines in a prospective randomised trial. They should be treated as educational outputs that summarise established clinical guidance rather than as clinical prescriptions. A prospective study comparing patient behaviour change in groups receiving versus not receiving these AI recommendations would be a valuable contribution to the literature on digital health behaviour change.

The current platform does not incorporate longitudinal risk trajectory estimation — the predicted probability is cross-sectional, reflecting current risk given current feature values. A dynamic model that updates risk estimates as successive measurements are collected over time, potentially incorporating change rates (e.g., annual rate of HbA1c increase) alongside absolute values, would substantially increase clinical utility for follow-up monitoring. Such a model could be implemented using a recurrent architecture or a Bayesian updating framework layered on top of the current ensemble.

Regulatory considerations are increasingly relevant as clinical ML tools move from academic prototypes toward formal deployment in clinical pathways. In the European Union, the EU Medical Device Regulation (MDR 2017/745) and the AI Act (2024) create overlapping obligations for clinical decision-support software. In India, the Central Drugs Standard Control Organisation (CDSCO) is developing a regulatory framework for AI-based medical devices. While the present tool is positioned as a screening and educational aid rather than a diagnostic device, the deployment of any risk stratification output in a clinical pathway raises questions of liability, audit traceability, and algorithmic accountability that future work should address explicitly.

VII. Conclusion

This work presents a production-grade, multi-level diabetes risk assessment platform built on a stacked ensemble of XGBoost, LightGBM, and CatBoost, trained across four diverse clinical datasets totalling over 150,000 records. The

platform achieves 91.8% accuracy and an ROC-AUC of 0.953 on held-out test data — improvements of 4.4 percentage points in accuracy and 0.023 in AUC over the SVM baseline, with statistical significance confirmed by McNemar's and DeLong's tests respectively.

SHAP-based explainability bridges the gap between model performance and clinical utility, providing per-patient feature attribution that clinicians and patients can inspect, interrogate, and use to inform shared decision-making. Plasma Glucose, BMI, and Age emerge as the dominant predictors, a finding consistent with established endocrinological understanding and one that lends face validity to the model's reasoning process. Probability calibration via Platt scaling reduces the Brier Score by 18%, ensuring that communicated risk probabilities carry genuine statistical meaning rather than serving merely as relative rankings.

The three-tier deployment architecture — lifestyle screening without laboratory requirements, clinical assessment with basic metabolic inputs, and comprehensive metabolic profiling with full organ function panels — makes the platform accessible across the full spectrum of clinical infrastructure, from a community health worker's smartphone to a hospital endocrinology workstation. The Streamlit application is publicly available, actively maintained, and updated automatically when the model repository is revised.

Diabetes is a preventable and manageable disease when detected during the prediabetic window. The platform described herein represents a substantive step toward democratising evidence-based risk stratification — combining the discrimination accuracy of gradient-boosted ensemble learning, the transparency of SHAP attribution, and the accessibility of a publicly deployed web interface. Future work will prioritise integration of South Asian diabetes datasets for population-specific risk calibration, longitudinal risk trajectory modelling, and a prospective clinical evaluation of the platform's impact on screening uptake and behaviour change in high-risk populations.

References

[1] IDF Diabetes Atlas, 10th Edition. International Diabetes Federation, Brussels, Belgium, 2021. Available: https://diabetesatlas.org

[2] T. Chen and C. Guestrin, "XGBoost: A scalable tree boosting system," in Proc. 22nd ACM SIGKDD Int. Conf. Knowledge Discovery and Data Mining, 2016, pp. 785–794.

[3] G. Ke et al., "LightGBM: A highly efficient gradient boosting decision tree," in Advances in Neural Information Processing Systems (NeurIPS), 2017, pp. 3146–3154.

[4] L. Prokhorenkova et al., "CatBoost: Unbiased boosting with categorical features," in Advances in Neural Information Processing Systems (NeurIPS), 2018.

[5] S. M. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," in Advances in Neural Information Processing Systems (NeurIPS), 2017, pp. 4765–4774.

[6] S. M. Lundberg et al., "From local explanations to global understanding with explainable AI for trees," Nature Machine Intelligence, vol. 2, pp. 56–67, 2020.

[7] I. Kavakiotis et al., "Machine learning and data mining methods in diabetes research," Computational and Structural Biotechnology Journal, vol. 15, pp. 104–116, 2017.

[8] D. Sisodia and D. S. Sisodia, "Prediction of diabetes using classification algorithms," Procedia Computer Science, vol. 132, pp. 1578–1585, 2018.

[9] Q. Zou et al., "Predicting diabetes mellitus with machine learning techniques," Frontiers in Genetics, vol. 9, p. 515, 2018.

[10] J. A. M. Sidey-Gibbons and C. J. Sidey-Gibbons, "Machine learning in medicine: A practical introduction," BMC Medical Research Methodology, vol. 19, no. 1, p. 64, 2019.

[11] National Center for Health Statistics, "National Health and Nutrition Examination Survey (NHANES) 1999–2018," Centers for Disease Control and Prevention, Hyattsville, MD. Available: https://www.cdc.gov/nchs/nhanes

[12] UK Biobank, "UK Biobank: Protocol for a large-scale prospective epidemiological resource," UK Biobank, 2007. Available: https://www.ukbiobank.ac.uk

[13] Diabetes Prevention Program Research Group, "Reduction in the incidence of type 2 diabetes with lifestyle intervention or metformin," New England Journal of Medicine, vol. 346, no. 6, pp. 393–403, 2002.

[14] N. V. Chawla et al., "SMOTE: Synthetic minority over-sampling technique," Journal of Artificial Intelligence Research, vol. 16, pp. 321–357, 2002.

[15] P. Rajpurkar et al., "AI in health and medicine," Nature Medicine, vol. 28, no. 1, pp. 31–38, 2022.

[16] R. Garg et al., "Explainable AI-based clinical decision support for sepsis prediction in the ICU," Critical Care Medicine, vol. 49, no. 8, pp. e789–e799, 2021.

[17] J. C. Platt, "Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods," Advances in Large Margin Classifiers, pp. 61–74, 1999.

[18] T. Akiba et al., "Optuna: A next-generation hyperparameter optimization framework," in Proc. 25th ACM SIGKDD Int. Conf. Knowledge Discovery and Data Mining, 2019, pp. 2623–2631.

[19] E. J. DeLong, D. M. DeLong, and D. L. Clarke-Pearson, "Comparing the areas under two or more correlated receiver operating characteristic curves: A nonparametric approach," Biometrics, vol. 44, no. 3, pp. 837–845, 1988.

[20] Nalla Adithya Reddy, "Diabetes Data Analysis and Machine Learning Based Prediction Model on Streamlit Web App," International Journal of Scientific Research and Engineering Development, vol. 5, issue 5, Sep–Oct 2022, pp. 437–443.