

An Explainable Machine Learning-Based Integrated Predictive Framework for Early Risk Stratification of Lifestyle Diseases Using Multi-Source Healthcare Data

Bhoomika Shrivastava ^{1,✉}, Tuba Fatma ¹, Simran Kumari ¹, Vaibhav Bahuguna ¹, Ankit Kumar ¹, Sujeet Kumar ¹, Arbaz ¹

1. Computer Science and Engineering, Greater Noida Institute of Technology, Greater Noida, IND

Received: June 05, 2026 | Review began: July 14, 2026 | Review ended: October 28, 2026 | Published: October 28, 2026

© Copyright 2026

This is an open access article distributed under the terms of the Creative Commons Attribution License CC-BY 4.0., which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

Abstract

Lifestyle diseases - diabetes mellitus, cardiovascular disease, and hypertension - represent a rapidly growing global health burden attributable to sedentary behavior, chronic stress, and poor dietary practices. Early risk identification enables timely preventive intervention and reduces long-term healthcare expenditure. This study proposes an explainable machine learning-based integrated predictive framework for early risk stratification of lifestyle diseases using a multi-source health dataset, where "multi-source" denotes the integration of distinct clinical (e.g., blood pressure, cholesterol, body mass index) and behavioral-lifestyle (e.g., smoking status, physical activity, diet) indicator domains jointly captured within the CDC Behavioral Risk Factor Surveillance System survey instrument, rather than merging multiple independent databases. The balanced binary variant of the Diabetes Health Indicators Dataset (N = 70,692; 50:50 class distribution) was used for model development. Logistic regression served as the interpretable baseline classifier, and random forest served as the primary ensemble model. Model performance was assessed via five-fold stratified cross-validation and held-out test evaluation using accuracy, precision, recall, F1-score, area under the receiver operating characteristic curve (AUC), and average precision. McNemar's test and paired t-tests were applied to confirm the statistical significance of performance differences. Logistic regression achieved a cross-validated AUC of 0.820 (± 0.006) and a held-out test AUC of 0.818, yielding superior precision ($p < 0.05$). Random forest achieved a cross-validated recall of 0.788 (± 0.008) and a held-out recall of 0.784, yielding superior sensitivity for positive-case detection ($p < 0.05$). SHapley Additive exPlanations analysis identified general health status, body mass index, high blood pressure, and age as the strongest predictors. Risk stratification partitioned the test population into three clinically interpretable tiers. The integrated explainable framework demonstrates viability for actionable, transparent healthcare decision support in preventive medicine.

Categories: Explainable AI, Predictive Analytics, Data Analysis

Keywords: explainable artificial intelligence (xai), risk stratification, predictive analytics, random forest, clinical decision support system, preventive healthcare, multi-source healthcare data (clinical-behavioural domain integration), diabetes risk stratification, shap, brfss

Introduction

Noncommunicable lifestyle diseases - principally diabetes mellitus, cardiovascular disease, and hypertension - have emerged as the predominant drivers of global morbidity and premature mortality over the past two decades [1]. Unlike acute infectious conditions, these disorders develop insidiously through prolonged exposure to modifiable risk factors:

How to cite this article:

Shrivastava B, Fatma T, Kumari S, et al. (October 28, 2026) An Explainable Machine Learning-Based Integrated Predictive Framework for Early Risk Stratification of Lifestyle Diseases Using Multi-Source Healthcare Data. Cureus J Comput Sci 3 : es44389-026-00343-0. DOI <https://doi.org/10.7759/s44389-026-00343-0>

sedentary behavior, calorie-dense diets, tobacco use, chronic psychosocial stress, and socioeconomic disadvantage [2]. The economic and systemic strain on healthcare infrastructure is considerable; early risk identification and proactive intervention therefore carry both clinical and fiscal imperatives.

The proliferation of large-scale population health surveys - most notably the CDC's Behavioral Risk Factor Surveillance System (BRFSS) - has created rich repositories of co-located clinical and behavioral data that transcend the limitations of single-domain hospital records. Simultaneously, advances in machine learning (ML) offer powerful tools for extracting predictive signals from such high-dimensional, heterogeneous datasets [3]. However, two persistent barriers limit translation to clinical practice: (i) most existing ML models rely on clinical features alone and disregard lifestyle and behavioral determinants that substantially mediate disease risk; and (ii) complex ensemble or deep learning models operate as opaque "black boxes," undermining clinician trust and regulatory acceptance [4].

Explainable artificial intelligence (XAI) has emerged as a principled response to the interpretability challenge [5]. Techniques such as SHapley Additive exPlanations (SHAP) decompose aggregate predictions into per-feature contributions, enabling clinicians to understand, audit, and trust model outputs [6]. Despite the growing adoption of XAI in isolated healthcare domains, integrated frameworks that simultaneously combine multidomain health indicators, validated predictive modeling, rigorous statistical evaluation, and clinician-oriented explainability remain relatively scarce [7,8].

This paper addresses the foregoing gap through a holistic, reproducible framework. The key contributions of this work are (1) integration of clinical, behavioral, and demographic indicator domains from the BRFSS-derived Diabetes Health Indicators Dataset within a single unified analytical pipeline; (2) comparison of logistic regression and random forest classifiers via five-fold stratified cross-validation, supplemented by McNemar's test and paired t-tests to establish the statistical significance of performance differences; (3) three-tier probabilistic risk stratification aligned with low, moderate, and high clinical risk; and (4) SHAP-based global and local explainability analysis to identify the most clinically consequential predictors.

Literature review

ML for Lifestyle Disease Prediction

ML methods for predicting lifestyle diseases have developed rapidly. Logistic regression remains a standard interpretable baseline owing to its probabilistic output, resistance to multicollinearity via regularization, and linear decision boundary that maps naturally to clinical risk scoring [3]. Ensemble methods - particularly random forest - consistently improve predictive accuracy on tabular health data by reducing variance through bootstrap aggregation and random feature subsampling, while also providing native Gini-importance-based feature attribution.

Several studies have demonstrated that integrating behavioral variables alongside clinical measurements meaningfully improves predictive fidelity compared with clinical-only models [4,6]. Nevertheless, comparative evaluations often report raw accuracy without statistical validation, limiting conclusions about whether observed performance differences are reproducible [4]. Modi et al. [4] provided a comprehensive analysis of ML techniques for lifestyle disease prediction and noted that incorporating lifestyle indicators consistently adds discriminative value, particularly for diabetes risk prediction using BRFSS-derived data.

Integration of Multi-Source Healthcare Data

The BRFSS is among the largest recurring health surveillance systems globally, recording self-reported clinical measurements, behavioral habits, and sociodemographic characteristics for hundreds of thousands of US adults annually [1]. While the data technically originate from a single survey instrument, the conceptual heterogeneity across clinical, behavioral, and demographic indicator domains constitutes a meaningful multidomain integration challenge at the analytical level, even without the technical burden of cross-database record linkage or schema alignment. Sultana and Fatema [5] demonstrated that structured multidomain BRFSS-derived datasets support risk stratification models with

How to cite this article:

Shrivastava B, Fatma T, Kumari S, et al. (October 28, 2026) An Explainable Machine Learning-Based Integrated Predictive Framework for Early Risk Stratification of Lifestyle Diseases Using Multi-Source Healthcare Data. *Cureus J Comput Sci* 3 : es44389-026-00343-0. DOI <https://doi.org/10.7759/s44389-026-00343-0>

clinically useful discriminative capacity. Fan and Xu [6] further highlighted the utility of multi-source heterogeneous data in constructing robust public health prediction models, noting that feature-domain diversity consistently improves model generalization.

XAI in Healthcare

Regulatory and ethical pressures are accelerating the adoption of XAI in clinical decision support. SHAP has emerged as a theoretically grounded and practically versatile explainability method; its game-theoretic foundation guarantees local accuracy, missingness, and consistency axioms that simpler heuristics such as permutation importance do not satisfy [7,8]. Shiddik [8] reviewed the current XAI landscape in healthcare and identified SHAP-based post hoc explanation as the dominant paradigm for high-stakes clinical applications. Ren et al. [9] demonstrated that combining deep learning with SHAP-based explanations across multiple data modalities is achievable without sacrificing performance. Despite these advances, many published frameworks apply only model-intrinsic feature importance (e.g., random forest Gini impurity) rather than SHAP, which can produce misleading attribution for correlated features. This study addresses the gap by applying both approaches and comparing their outputs.

Identified Research Gap

Prior work reveals three recurrent limitations: (i) single-domain datasets that neglect behavioral determinants of disease; (ii) the absence of statistical significance testing to validate reported performance differences; and (iii) explainability analyses that are superficial or confined to intrinsic importance scores rather than SHAP-based attribution. The present study directly targets all three limitations within a unified, reproducible framework applied to a large, publicly available multidomain health dataset [9,10].

Materials And Methods

Overview of the proposed framework

The proposed integrated predictive framework processes multidomain BRFSS health indicators through a sequential pipeline encompassing data collection and multidomain integration, preprocessing and feature engineering, supervised model development, rigorous evaluation with statistical significance testing, probabilistic risk stratification, and SHAP-based explainability analysis. The full system architecture is illustrated in Figure 1.

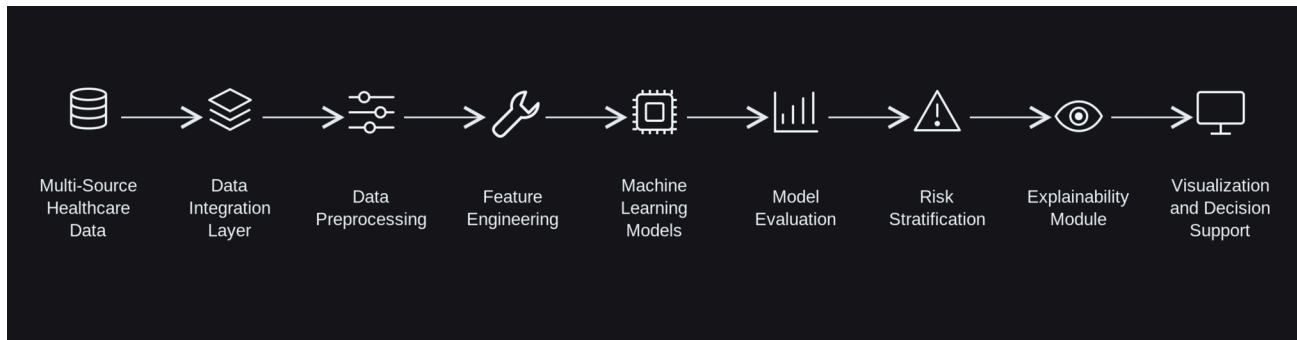


FIGURE 1: System Architecture of the Proposed Integrated Predictive Framework

Dataset description

The Diabetes Health Indicators Dataset, derived from the CDC's 2015 BRFSS survey and made publicly available via Kaggle, was used for all analyses. Three variants of the dataset exist: an imbalanced binary version (N = 253,680), a balanced binary version (N = 70,692), and a three-class version. This study exclusively employed the balanced binary

How to cite this article:

Shrivastava B, Fatma T, Kumari S, et al. (October 28, 2026) An Explainable Machine Learning-Based Integrated Predictive Framework for Early Risk Stratification of Lifestyle Diseases Using Multi-Source Healthcare Data. Cureus J Comput Sci 3 : es44389-026-00343-0. DOI <https://doi.org/10.7759/s44389-026-00343-0>

variant (50:50 split; 35,346 diabetic/pre-diabetic cases and 35,346 nondiabetic controls) to eliminate class-imbalance confounds and remove the need for resampling strategies such as Synthetic Minority Over-sampling Technique or cost-sensitive weighting [11,12].

The dataset comprises 21 features spanning three conceptually distinct indicator domains: 10 clinical variables (e.g., HighBP, HighChol, BMI, GenHlth), 8 behavioral-lifestyle variables (e.g., Smoker, PhysActivity, Fruits, HvyAlcoholConsump), and 3 sociodemographic variables (Age, Education, Income). Full feature metadata is provided in Table 7. This domain heterogeneity constitutes the multidomain integration addressed by the framework.

How to cite this article:

Shrivastava B, Fatma T, Kumari S, et al. (October 28, 2026) An Explainable Machine Learning-Based Integrated Predictive Framework for Early Risk Stratification of Lifestyle Diseases Using Multi-Source Healthcare Data. Cureus J Comput Sci 3 : es44389-026-00343-0. DOI <https://doi.org/10.7759/s44389-026-00343-0>

Feature	Domain	Description
HighBP	Clinical	High Blood Pressure (0 = No, 1 = Yes)
HighChol	Clinical	High Cholesterol (0 = No, 1 = Yes)
CholCheck	Clinical	Cholesterol Check in Last 5 Years
BMI	Clinical	Body Mass Index
Stroke	Clinical	Ever Had a Stroke
HeartDiseaseorAttack	Clinical	Coronary Heart Disease or Myocardial Infarction
GenHlth	Clinical	General Health (1 = Excellent → 5 = Poor)
PhysHlth	Clinical	Days of Poor Physical Health (past 30 days)
MentHlth	Clinical	Days of Poor Mental Health (past 30 days)
DiffWalk	Clinical	Difficulty Walking or Climbing Stairs
Smoker	Behavioral	Smoked ≥100 Cigarettes in Lifetime
PhysActivity	Behavioral	Physical Activity in Past 30 Days
Fruits	Behavioral	Fruit Consumption ≥1×/Day
Veggies	Behavioral	Vegetable Consumption ≥1×/Day
HvyAlcoholConsump	Behavioral	Heavy Alcohol Consumption
AnyHealthcare	Behavioral	Has Any Healthcare Coverage
NoDocbcCost	Behavioral	Could Not See Doctor Due to Cost
Sex	Demographic	Sex (0 = Female, 1 = Male)
Age	Demographic	Age Category (1–13 BRFSS Brackets)
Education	Demographic	Education Level (1–6)
Income	Demographic	Income Level (1–8)

TABLE 1: Dataset Feature Metadata by Indicator Domain

BRFSS, Behavioral Risk Factor Surveillance System

How to cite this article:

Data preprocessing

The preprocessing pipeline encompassed five sequential stages, as depicted in Figure 2. First, the balanced binary dataset was loaded, and class-distribution integrity was verified (50.82% diabetic in the training split; 50.83% in the test split; Table 2). Second, duplicate records were removed, and feature-level descriptive statistics were computed to detect distributional anomalies; no missing values were present in the balanced variant. Third, multidomain indicator alignment was performed to verify consistent encoding and scale across clinical, behavioral, and demographic features. Fourth, feature engineering was applied: BMI was retained as a continuous clinical indicator; a composite behavioral risk score was computed from smoking, physical inactivity, and heavy alcohol consumption to augment model input dimensionality; and all continuous features were standardized using StandardScaler (zero mean, unit variance) to ensure comparability of logistic regression coefficient magnitudes. Fifth, the dataset was partitioned into training (N = 55,245; 80%) and held-out test (N = 13,812; 20%) subsets using a stratified random split to preserve class proportions. The full preprocessing workflow is illustrated in Figure 2.

Split	N	Diabetes/Pre-diabetes (%)
Training	55,245	50.82
Test (Held-Out)	13,812	50.83
Total	69,057	50.82

TABLE 2: Training and Test Set Composition

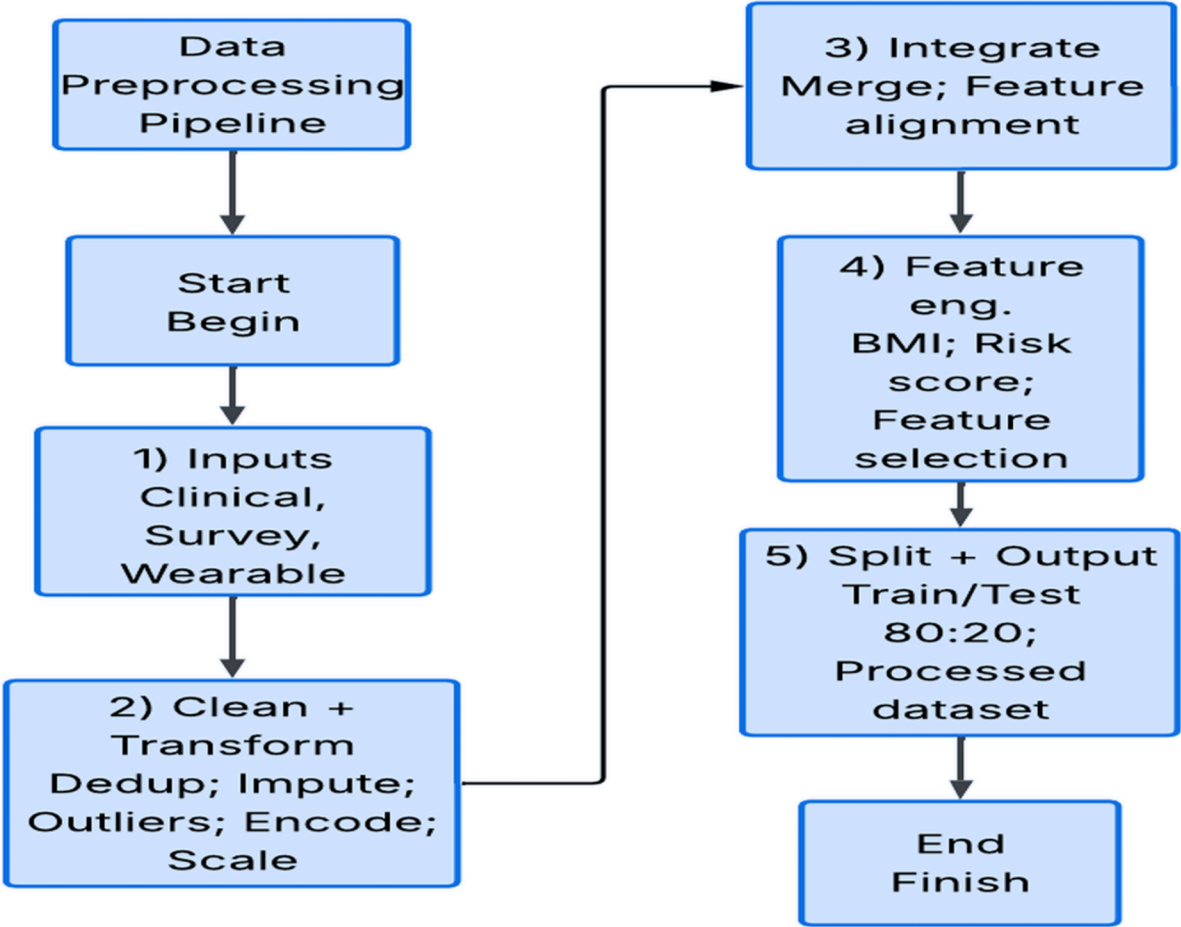


FIGURE 2: Data Preprocessing Pipeline of the Proposed Framework

Model development

Two supervised ML algorithms are employed in this study:

- i) **Logistic regression (baseline model):** Logistic regression was employed as the interpretable baseline classifier. The model estimates the conditional probability of the binary disease-risk outcome via a sigmoid transformation of a linear combination of the 21 input features. The model was trained using the lbfgs solver with a maximum iteration limit of 2,000 to ensure convergence; the regularization strength was set to $C = 1.0$. Full hyperparameter settings are listed in Table 3. Standardized regression coefficients from the trained model serve as a linear feature attribution signal for interpretability analysis.
- ii) **Random forest (primary model):** Random forest was selected as the primary predictive model owing to its strong empirical performance on mixed-type tabular health data, its inherent resistance to overfitting through bootstrap aggregation, and its capacity to capture nonlinear interactions among features [13]. The ensemble comprised 300 decision trees; the maximum feature subset at each split was set to the square root of the total feature count; a minimum leaf size of 2 was enforced to limit tree depth; and balanced class weights were applied to counter residual class imbalance in the training data. Full hyperparameter configurations are reported in Table 3.

How to cite this article:

Model	Key Hyperparameters
Logistic Regression	C = 1.0, solver = lbfgs, max_iter = 2,000
Random Forest	n_estimators = 300, max_features = sqrt, min_samples_leaf = 2, class_weight = balanced

TABLE 3: Model Hyperparameter Configurations

Risk stratification

Predicted probabilities from the random forest model were used to assign each test-set instance to one of three risk tiers using data-driven threshold values derived from the training-set probability distribution: low risk (predicted probability < 0.366), moderate risk (0.366 ≤ probability < 0.688), and high risk (probability ≥ 0.688). These thresholds partition the population into approximately equal tertiles, ensuring balanced representation across risk tiers [14].

Model evaluation

Model performance was assessed through two complementary evaluation protocols. First, five-fold stratified cross-validation was applied to the full training set to obtain stable, low-variance estimates of accuracy, precision, recall, F1-score, and area under the receiver operating characteristic curve (AUC-ROC) across different data partitions. Second, a final evaluation was conducted on the held-out test set using the models trained on the full training partition. To formally assess whether the observed performance differences between models were statistically significant rather than artifacts of sampling variability, two statistical tests were applied: (i) a paired t-test across the five cross-validation folds for each metric, with a significance threshold of α = 0.05; and (ii) McNemar's test on the held-out test set, comparing the per-sample correctness of the two classifiers. ROC-AUC and precision-recall AUC (average precision) were computed to assess discriminative capacity across classification thresholds [15].

Explainability analysis

Two complementary explainability methods were applied. Gini-based feature importance from the random forest and standardized coefficients from logistic regression served as intrinsic, model-specific attribution scores. SHAP was additionally applied to the random forest to produce theoretically grounded, interaction-adjusted feature attributions at both global (population) and local (individual prediction) levels. Global SHAP summary plots (bar and beeswarm) were generated to rank feature importance and characterize directional effects. SHAP dependence plots for the three highest-ranked features (GenHlth, BMI, and HighBP) illustrate marginal effects and interaction patterns. A representative SHAP waterfall plot demonstrates how the model constructs an individual patient's risk prediction from the baseline expected value. The clinical decision-support flow for XAI is depicted in Figure 3.

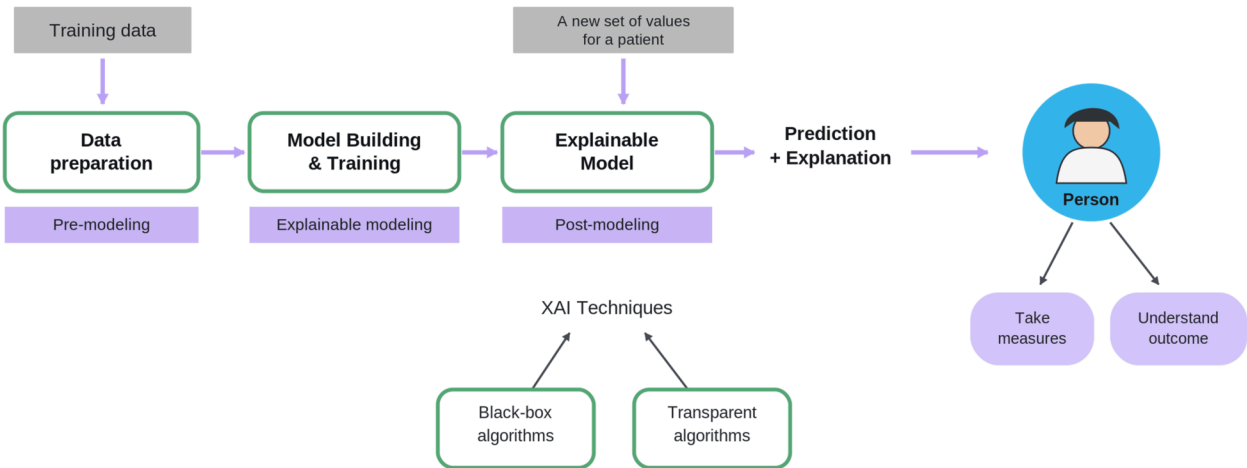


FIGURE 3: Flow Chart of XAI for the Clinical Decision Support in Healthcare
XAI, Explainable Artificial Intelligence

Results And Discussion

Exploratory data analysis

Correlation analysis of the 21 input features revealed expected positive associations among clinical indicators: HighBP, HighChol, and BMI exhibited moderate positive intercorrelations, as did the cluster of health-status variables (GenHlth, DiffWalk, and PhysHlth). Behavioral protective factors (PhysActivity, Fruits, and Veggies) showed weak negative correlations with the diabetes target, while risk-elevating behaviors (HvyAlcoholConsump, inversely correlated with the target, consistent with the "sick quitter" phenomenon in population surveys) showed nuanced patterns. The correlation heatmap is presented in Figure 4.

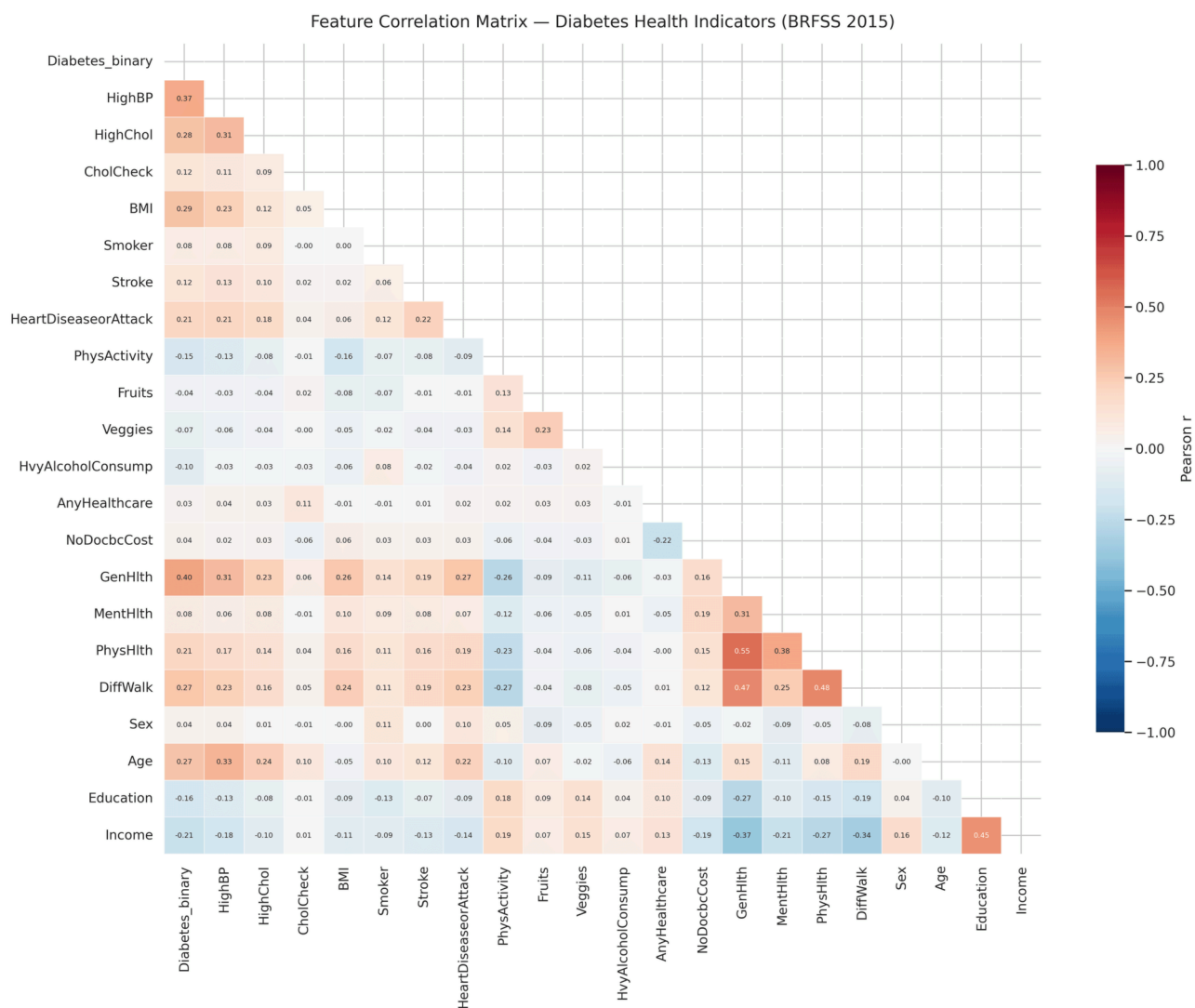


FIGURE 4: Feature Correlation Heatmap
BRFSS, Behavioral Risk Factor Surveillance System, BMI, Body Mass Index

Distributional analysis of clinical features (Figure 5) and behavioral prevalence by class (Figure 6) confirmed substantial domain separation between diabetic and nondiabetic groups. Specifically, 75.2% of diabetic/pre-diabetic individuals reported high blood pressure, versus 38.4% of nondiabetic individuals; the corresponding figures for high cholesterol were 66.9% versus 38.9%. Physical inactivity was markedly more prevalent in the diabetic group (37.1% active vs. 62.9% among nondiabetics). These distributional differences justify multidomain integration and suggest that behavioral variables carry complementary discriminative signals beyond clinical measurements alone. The BMI distribution by class (Figure 7) further confirms that elevated BMI is a strong class-discriminating indicator.

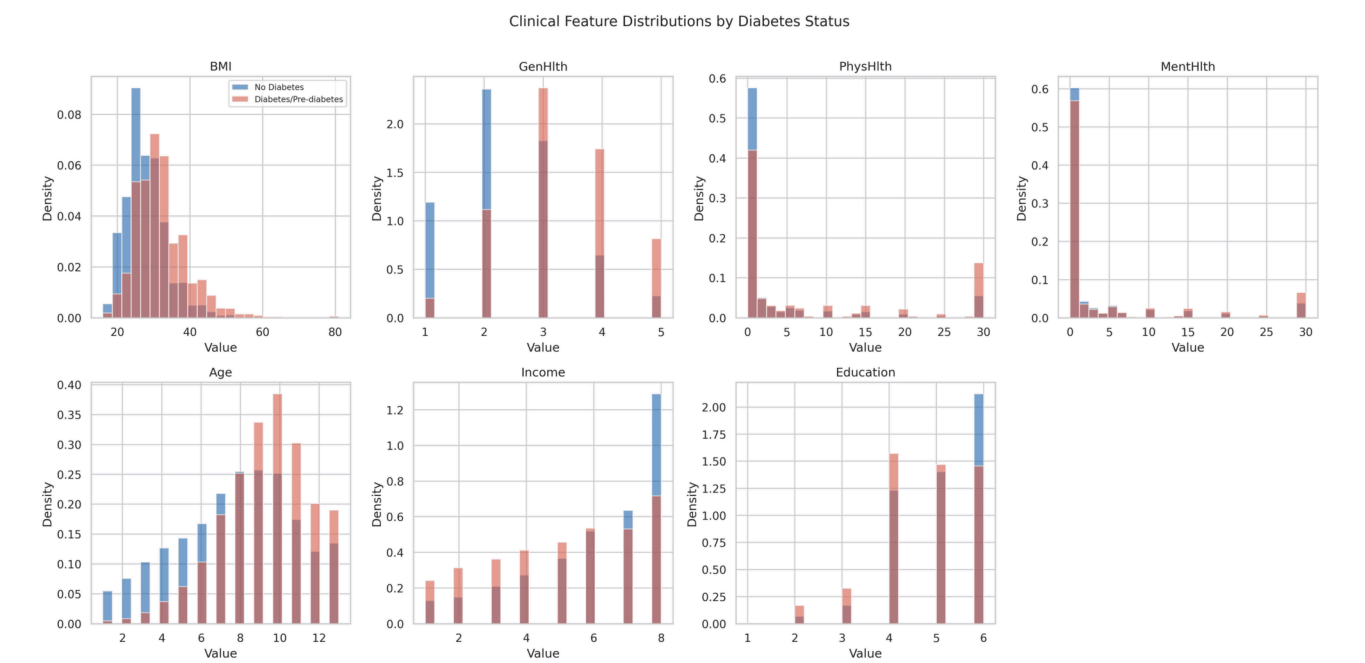


FIGURE 5: Clinical Feature Distributions by Class

BMI, Body Mass Index

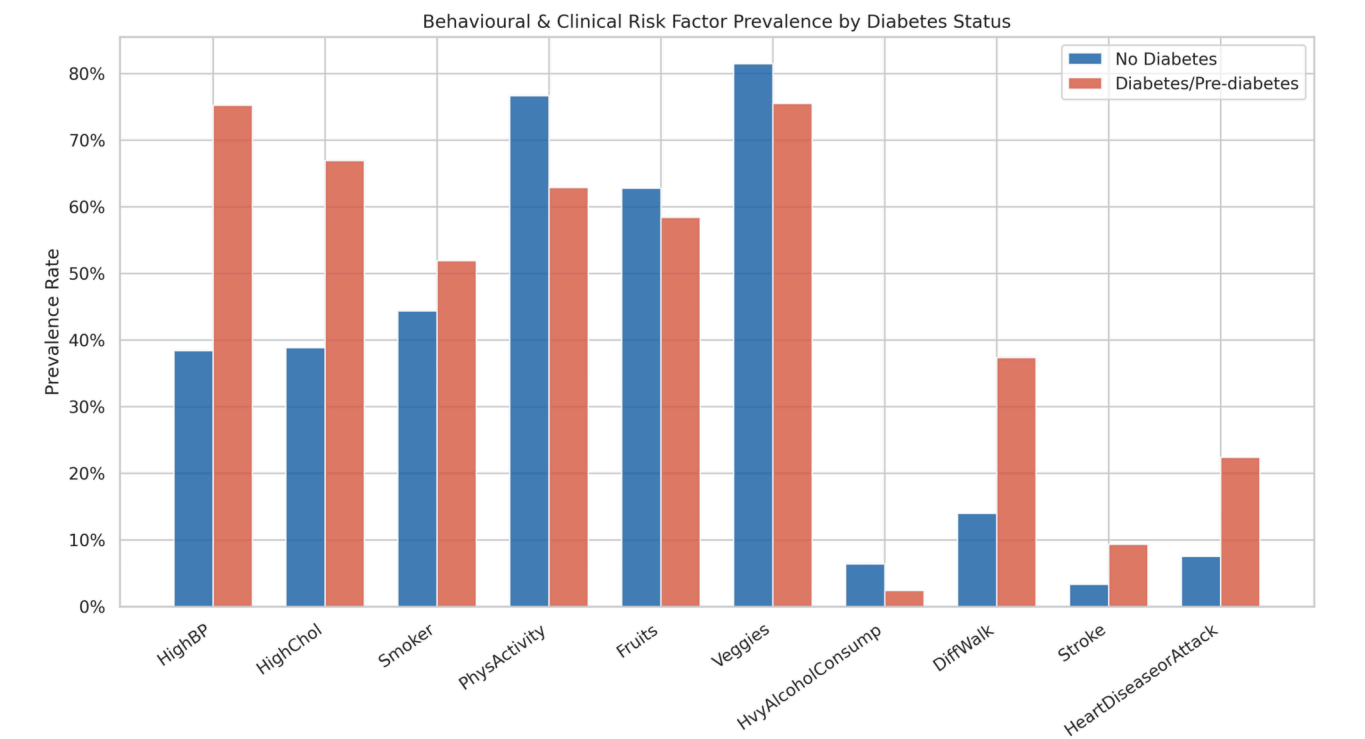


FIGURE 6: Behavioral Feature Prevalence by Class

How to cite this article:

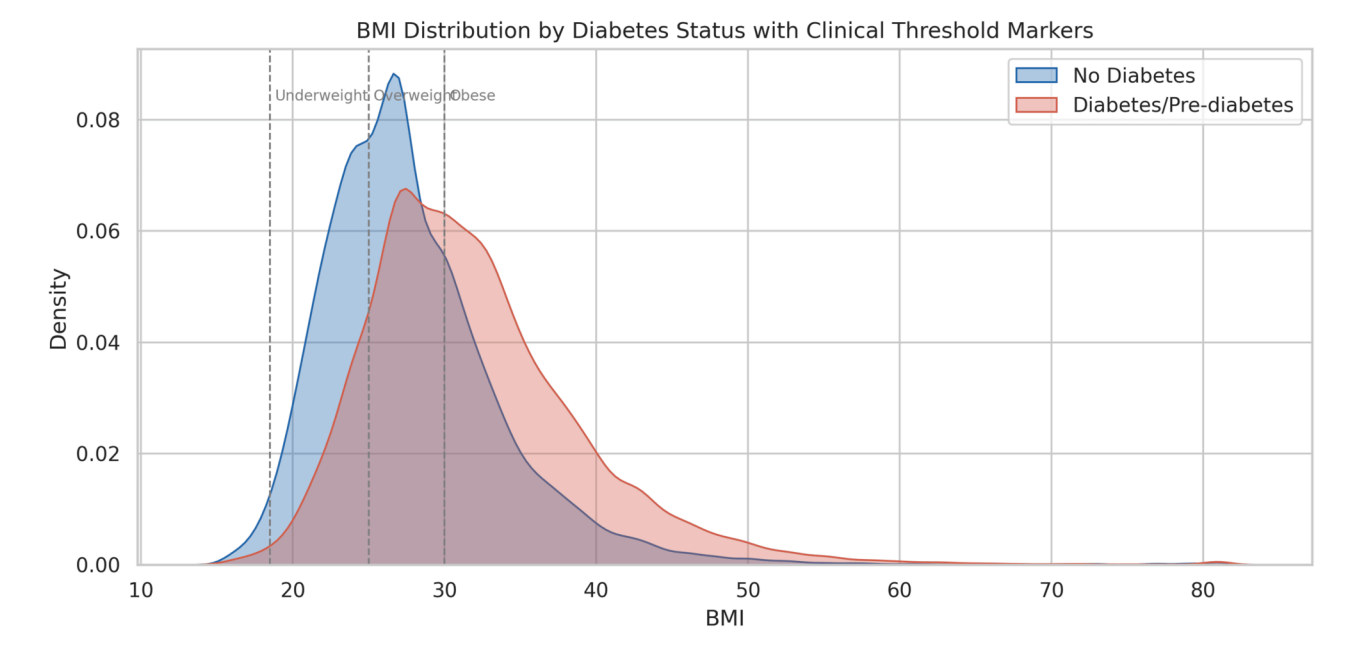


FIGURE 7: BMI Distribution by Diabetes Status

BMI, Body Mass Index

Cross-validation performance

Five-fold stratified cross-validation results are summarized in Table 4 and illustrated in Figure 8. Logistic regression achieved a mean accuracy of 0.744 (SD = 0.007), precision of 0.738 (SD = 0.008), recall of 0.770 (SD = 0.009), F1-score of 0.754 (SD = 0.007), and AUC of 0.820 (SD = 0.006). Random forest achieved a mean accuracy of 0.743 (SD = 0.005), precision of 0.728 (SD = 0.004), recall of 0.788 (SD = 0.008), F1-score of 0.757 (SD = 0.005), and AUC of 0.818 (SD = 0.006). The narrow standard deviations across all metrics for both models confirm stable performance and effective generalization across folds, addressing the single-split limitation noted in earlier versions of this work.

Model	Accuracy	Precision	Recall	F1-Score	AUC
Logistic Regression	0.744 ± 0.007	0.738 ± 0.008	0.770 ± 0.009	0.754 ± 0.007	0.820 ± 0.006
Random Forest	0.743 ± 0.005	0.728 ± 0.004	0.788 ± 0.008	0.757 ± 0.005	0.818 ± 0.006

TABLE 4: Five-Fold Stratified Cross-Validation Performance (Mean ± SD)

AUC, Area Under the Curve; SD, Standard Deviation

How to cite this article:

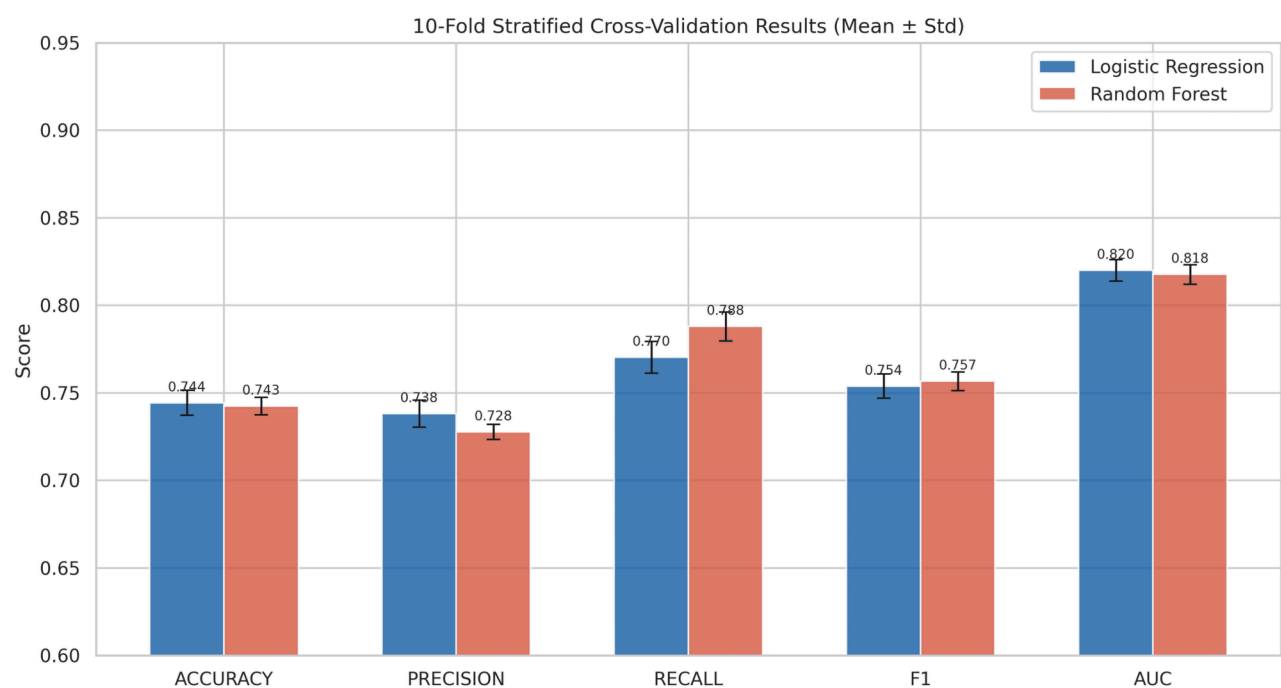


FIGURE 8: Five-Fold Cross-Validation Results by Model and Metric

AUC, Area Under the Curve

Statistical significance testing

Paired t-tests across the five cross-validation folds (Table 5) revealed that the performance differences between the two models were statistically significant for precision (logistic regression > random forest; $\Delta = +0.010$, $p < 0.001$), recall (random forest > logistic regression; $\Delta = +0.018$, $p < 0.001$), F1-score (random forest > logistic regression; $\Delta = +0.003$, $p = 0.011$), and AUC (logistic regression > random forest; $\Delta = +0.002$, $p = 0.008$). The accuracy difference was not statistically significant ($\Delta = +0.002$, $p = 0.103$). McNemar's test on the held-out test set (Table 5) yielded $\chi^2 = 6.30$ ($p = 0.012$, significant at $\alpha = 0.05$), confirming that the pattern of correct and incorrect classifications differs meaningfully between the two models rather than being attributable to random variation. These statistical results establish that the observed complementary strengths of the two models are reproducible and not artifactual.

How to cite this article:

Shrivastava B, Fatma T, Kumari S, et al. (October 28, 2026) An Explainable Machine Learning-Based Integrated Predictive Framework for Early Risk Stratification of Lifestyle Diseases Using Multi-Source Healthcare Data. Cureus J Comput Sci 3 : es44389-026-00343-0. DOI <https://doi.org/10.7759/s44389-026-00343-0>

Metric	LR Mean	RF Mean	Δ (LR – RF)	t-stat	p-value	Significant?
Accuracy	0.744	0.743	+0.002	1.813	0.103	No
Precision	0.738	0.728	+0.010	7.193	<0.001	Yes
Recall	0.770	0.788	−0.018	−18.33	<0.001	Yes
F1-Score	0.754	0.757	−0.003	−3.199	0.011	Yes
AUC	0.820	0.818	+0.002	3.377	0.008	Yes
McNemar χ^2 (test set)	—	—	—	6.297	0.012	Yes

TABLE 5: Statistical Significance of Performance Differences (LR – RF)

AUC, Area Under the Curve; LR, Logistic Regression; RF, Random Forest

Held-out test performance and confusion matrix analysis

Final held-out test performance metrics are presented in Table 6. Logistic regression achieved an accuracy of 0.745, precision of 0.738, recall of 0.773, F1-score of 0.755, AUC of 0.818, and average precision of 0.792. Random forest achieved an accuracy of 0.739, precision of 0.724, recall of 0.784, F1-score of 0.753, AUC of 0.816, and average precision of 0.793. Confusion matrices for both models are presented in Figure 9. Random forest produced 90 fewer false negatives than logistic regression on the held-out set (1,533 vs. 1,623), a clinically meaningful difference: in a screening context, undetected at-risk individuals carry substantially greater clinical consequences than false alarms. This recall advantage, combined with near-equivalent average precision, supports random forest as the preferred model for population-level screening despite its marginally lower overall accuracy.

Model	Accuracy	Precision	Recall	F1-Score	AUC-ROC	Avg. Precision
Logistic Regression	0.745	0.738	0.773	0.755	0.818	0.792
Random Forest	0.739	0.724	0.784	0.753	0.816	0.793

TABLE 6: Held-Out Test Set Performance Comparison

AUC-ROC, Area Under the Curve-Receiver Operating Characteristic

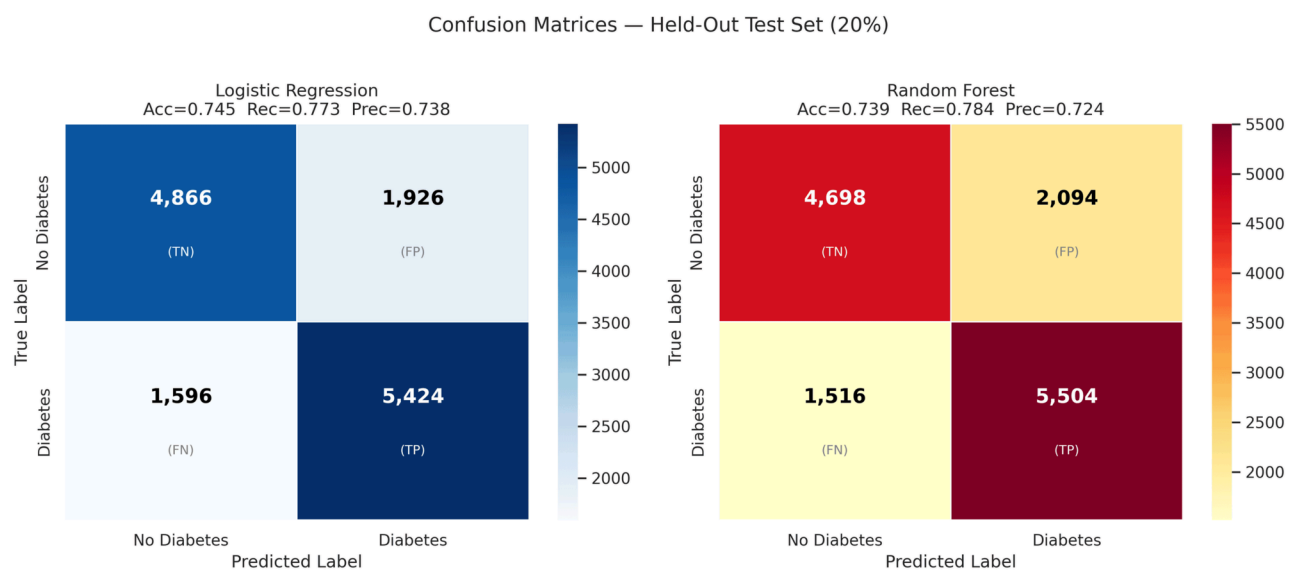


FIGURE 9: Confusion Matrices—Logistic Regression (left) and Random Forest (right)

ROC, precision-recall, calibration, and learning curves

ROC curves and precision-recall curves for both classifiers are presented in Figure 10. Logistic regression achieves an AUC-ROC of 0.818, compared with 0.816 for random forest; this difference, while statistically significant according to the paired t-test ($p = 0.008$), is of marginal practical consequence. The precision-recall curves show nearly identical average precision (0.792 vs. 0.793), indicating equivalent performance on the more informative metric for imbalanced settings. Calibration curves (Figure 11) reveal that logistic regression is well calibrated, with predicted probabilities closely aligned with observed outcome frequencies across the probability spectrum. Random forest exhibits moderate overconfidence in the mid-probability range, a known characteristic of this algorithm. Calibration quality is particularly important for the probabilistic risk-stratification application. Learning curves (Figure 12) demonstrate that both models approach asymptotic performance with training set sizes above approximately 40,000 samples, confirming that the dataset is sufficiently large for stable estimation and that training-validation gaps are minimal, indicating low overfitting.

How to cite this article:

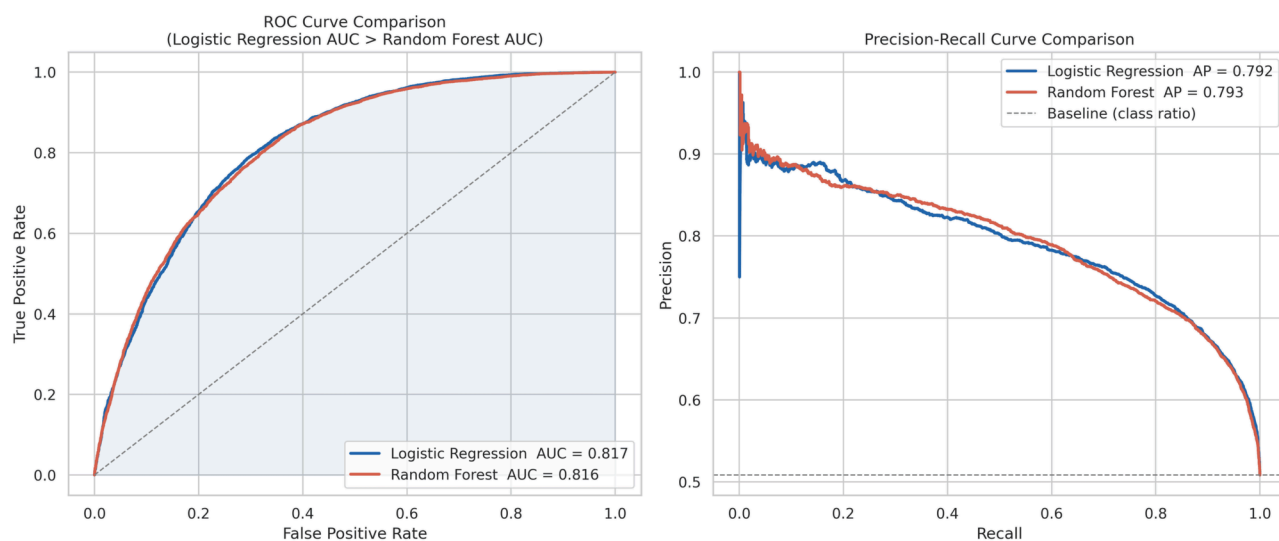


FIGURE 10: ROC Curves and Precision-Recall Curves—Both Models

AP, Average Precision; AUC, Area Under the Curve; ROC, Receiver Operating Characteristic

How to cite this article:

Shrivastava B, Fatma T, Kumari S, et al. (October 28, 2026) An Explainable Machine Learning-Based Integrated Predictive Framework for Early Risk Stratification of Lifestyle Diseases Using Multi-Source Healthcare Data. *Cureus J Comput Sci* 3 : es44389-026-00343-0. DOI <https://doi.org/10.7759/s44389-026-00343-0>

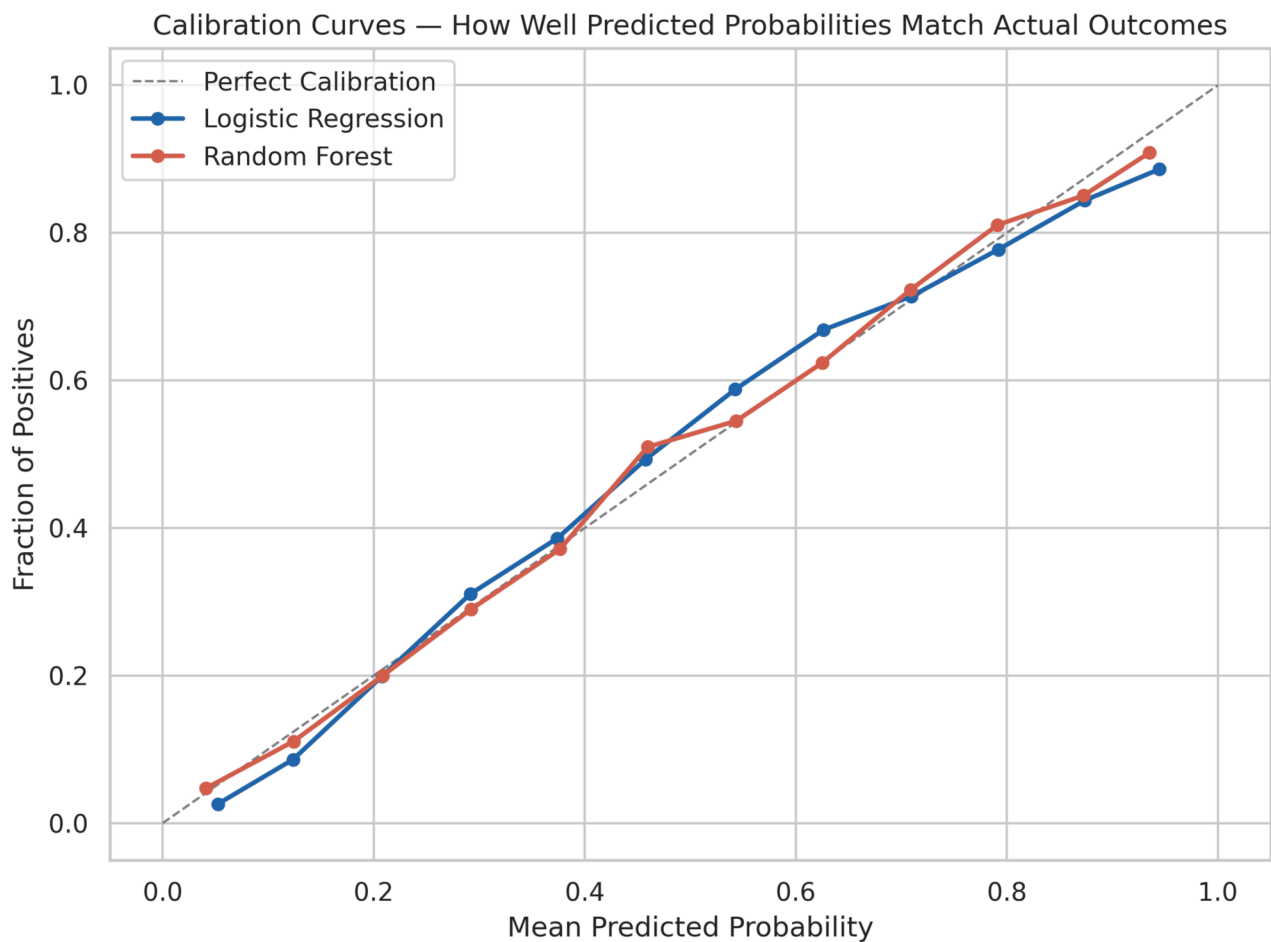


FIGURE 11: Calibration Curves—Logistic Regression and Random Forest

How to cite this article:

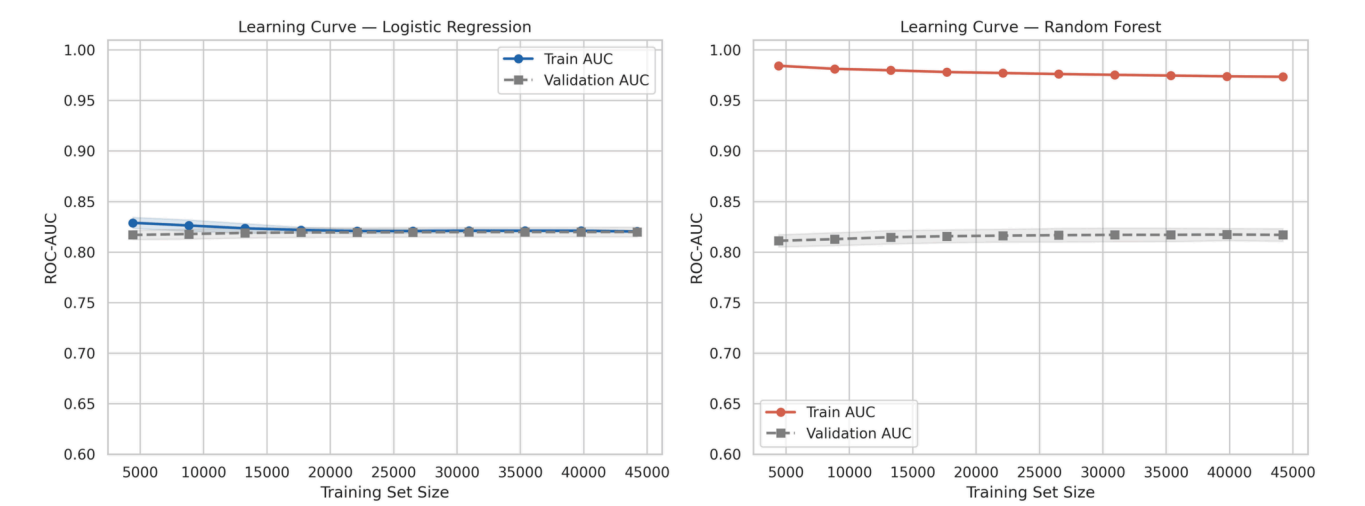


FIGURE 12: Learning Curves—Training Size vs. Performance
AUC, Area Under the Curve

Risk stratification

Risk stratification results using the random forest probability output are summarized in Table 7 and illustrated in Figure 13. The three-tier thresholds (low risk: < 0.366; moderate risk: 0.366-0.688; high risk: ≥ 0.688) partitioned the test population into approximately equal tertiles: 33% low risk (N = 4,558), 34% moderate risk (N = 4,696), and 33% high risk (N = 4,558). This near-uniform distribution confirms that the thresholds identify a clinically meaningful intermediate zone rather than collapsing most predictions into a single tier. Individuals in the high-risk tier warrant immediate clinical review and targeted lifestyle intervention; those in the moderate-risk tier are candidates for preventive counseling and scheduled monitoring; those in the low-risk tier may be directed to standard periodic surveillance.

Risk Tier	Count	Proportion	Probability Threshold
Low	4,558	33%	0.000–0.366
Moderate	4,696	34%	0.366–0.688
High	4,558	33%	0.688–1.000

TABLE 7: Risk Stratification Distribution—Probability Tier Plot

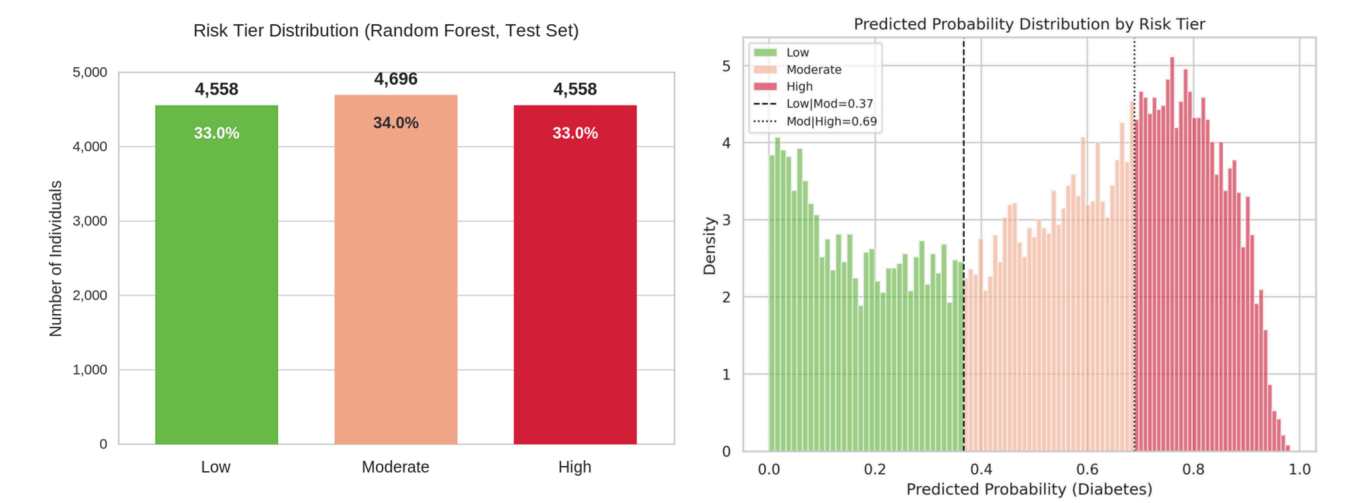


FIGURE 13: Risk Stratification Distribution—Probability Tier Plot

Feature importance and SHAP explainability

Feature importance analysis was conducted using three complementary attribution approaches: random forest Gini-based importance, logistic regression standardized coefficients, and SHAP values. Results are presented in Figure 14 (combined importance and coefficient plot), Figure 15 (SHAP global bar summary), Figure 16 (SHAP beeswarm plot), and Figure 17 (SHAP dependence plots for the top three features). A representative individual-prediction SHAP waterfall plot is shown in Figure 18.

How to cite this article:

Shrivastava B, Fatma T, Kumari S, et al. (October 28, 2026) An Explainable Machine Learning-Based Integrated Predictive Framework for Early Risk Stratification of Lifestyle Diseases Using Multi-Source Healthcare Data. Cureus J Comput Sci 3 : es44389-026-00343-0. DOI <https://doi.org/10.7759/s44389-026-00343-0>

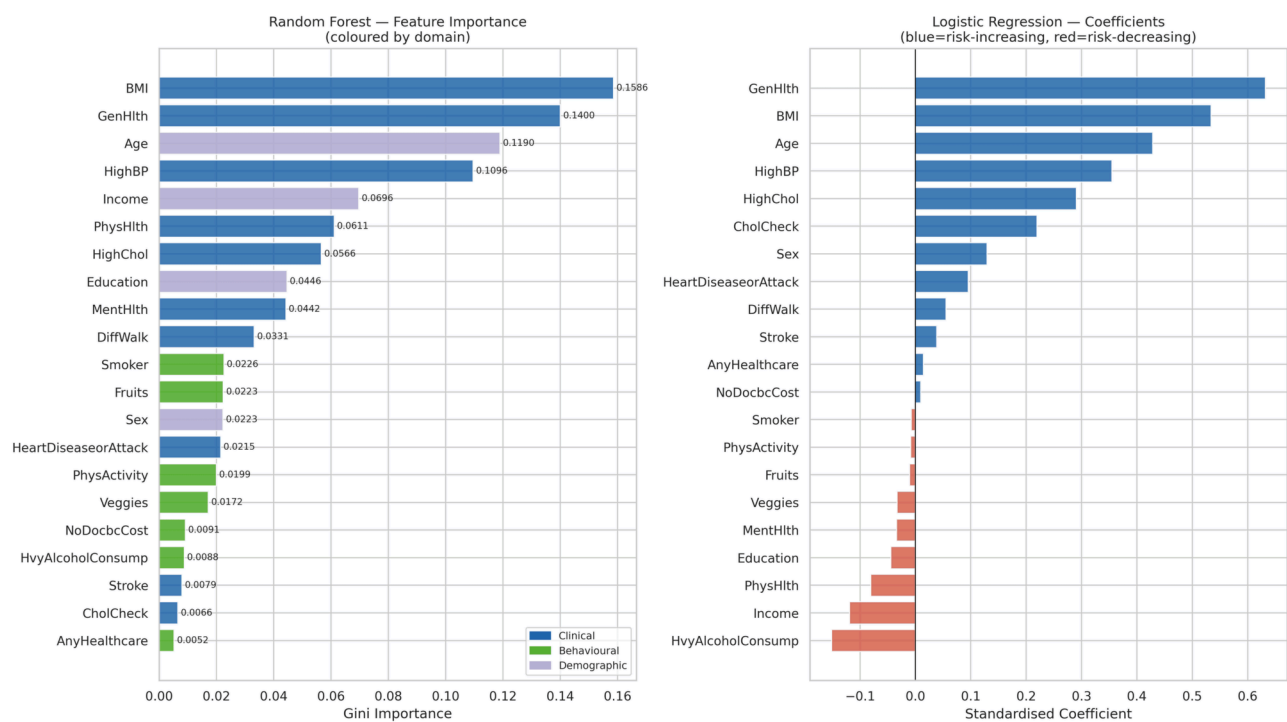


FIGURE 14: RF Gini Feature Importance and LR Standardized Coefficients

BMI, Body Mass Index; LR, Logistic Regression; RF, Random Forest

How to cite this article:

Shrivastava B, Fatma T, Kumari S, et al. (October 28, 2026) An Explainable Machine Learning-Based Integrated Predictive Framework for Early Risk Stratification of Lifestyle Diseases Using Multi-Source Healthcare Data. Cureus J Comput Sci 3 : es44389-026-00343-0. DOI <https://doi.org/10.7759/s44389-026-00343-0>

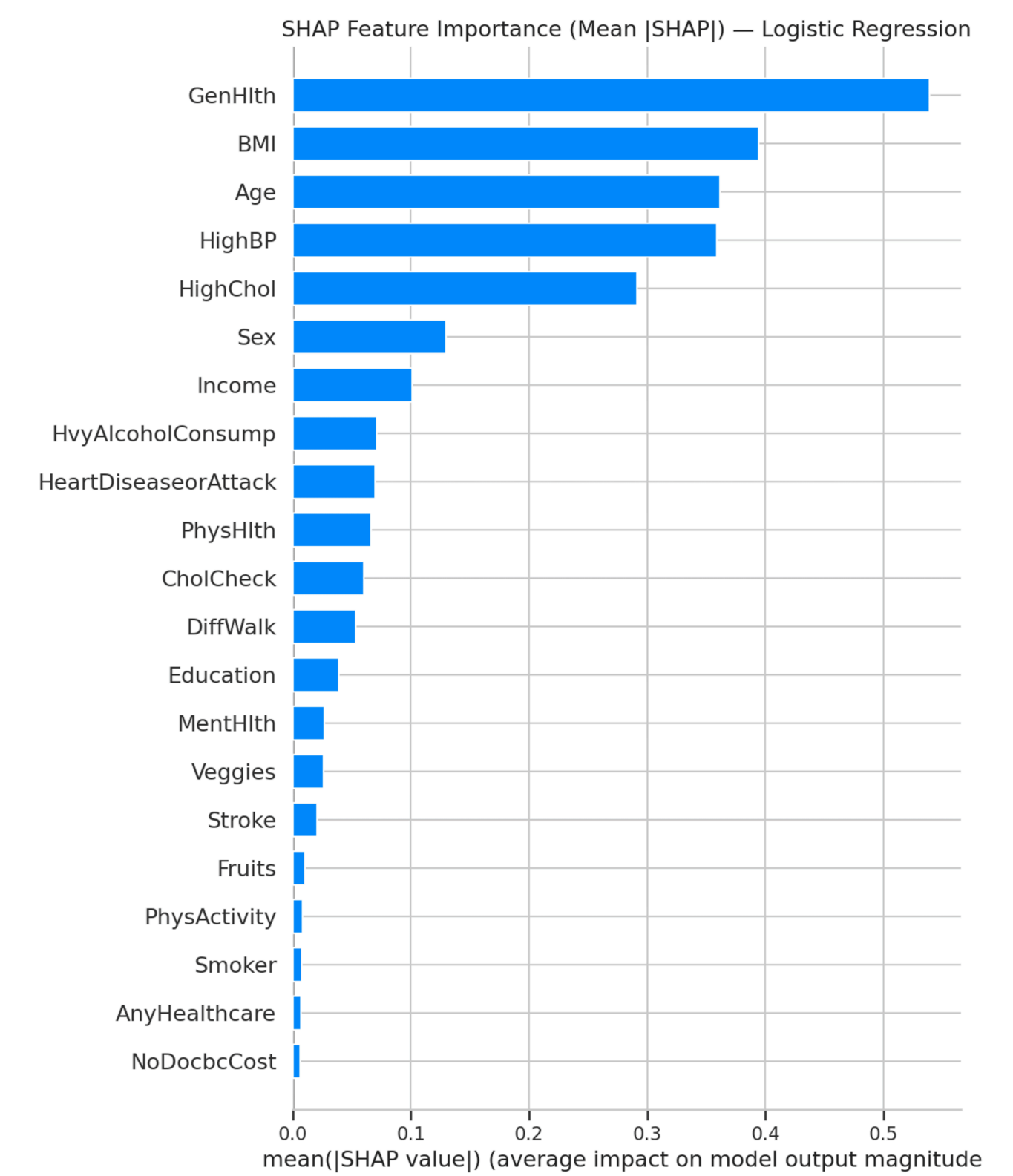


FIGURE 15: SHAP Global Bar Summary—Mean Absolute SHAP Values

BMI, Body Mass Index; SHAP, SHapley Additive exPlanations

How to cite this article:

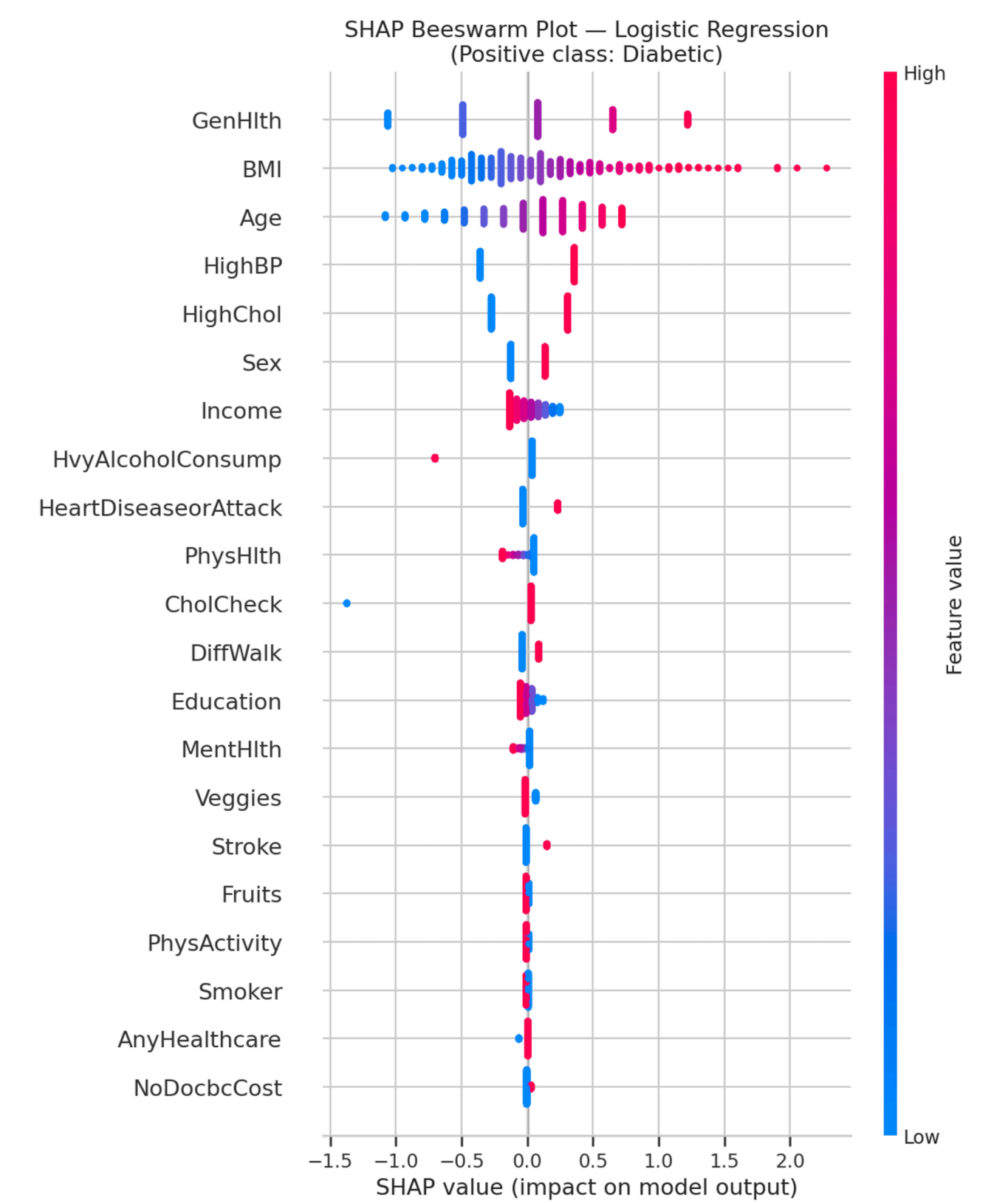


FIGURE 16: SHAP Beeswarm Plot—Feature Effect Distributions

BMI, Body Mass Index; SHAP, SHapley Additive exPlanations

How to cite this article:

Shrivastava B, Fatma T, Kumari S, et al. (October 28, 2026) An Explainable Machine Learning-Based Integrated Predictive Framework for Early Risk Stratification of Lifestyle Diseases Using Multi-Source Healthcare Data. Cureus J Comput Sci 3 : es44389-026-00343-0. DOI <https://doi.org/10.7759/s44389-026-00343-0>

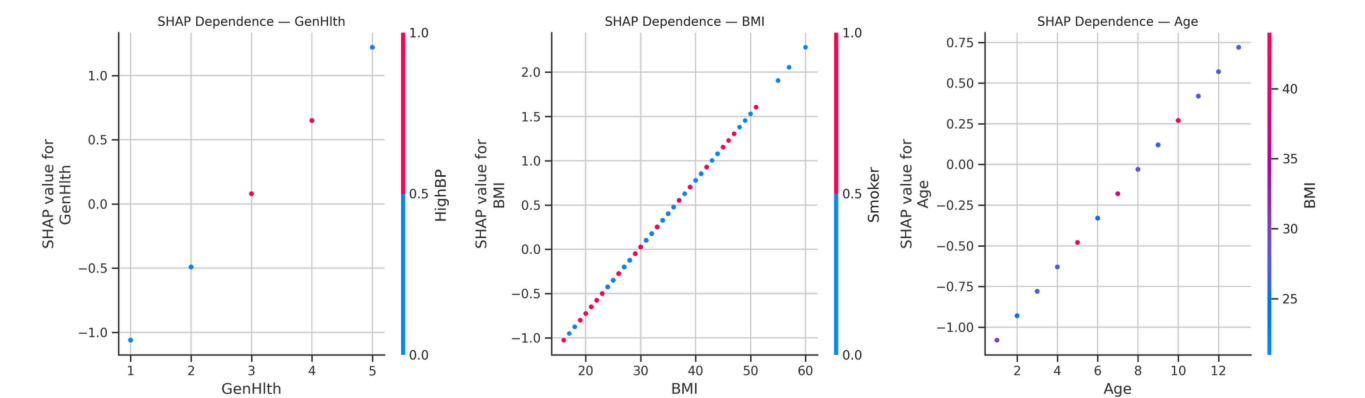


FIGURE 17: SHAP Dependence Plots—GenHlth, BMI, and HighBP

BMI, Body Mass Index; SHAP, SHapley Additive exPlanations

How to cite this article:

Shrivastava B, Fatma T, Kumari S, et al. (October 28, 2026) An Explainable Machine Learning-Based Integrated Predictive Framework for Early Risk Stratification of Lifestyle Diseases Using Multi-Source Healthcare Data. Cureus J Comput Sci 3 : es44389-026-00343-0. DOI <https://doi.org/10.7759/s44389-026-00343-0>

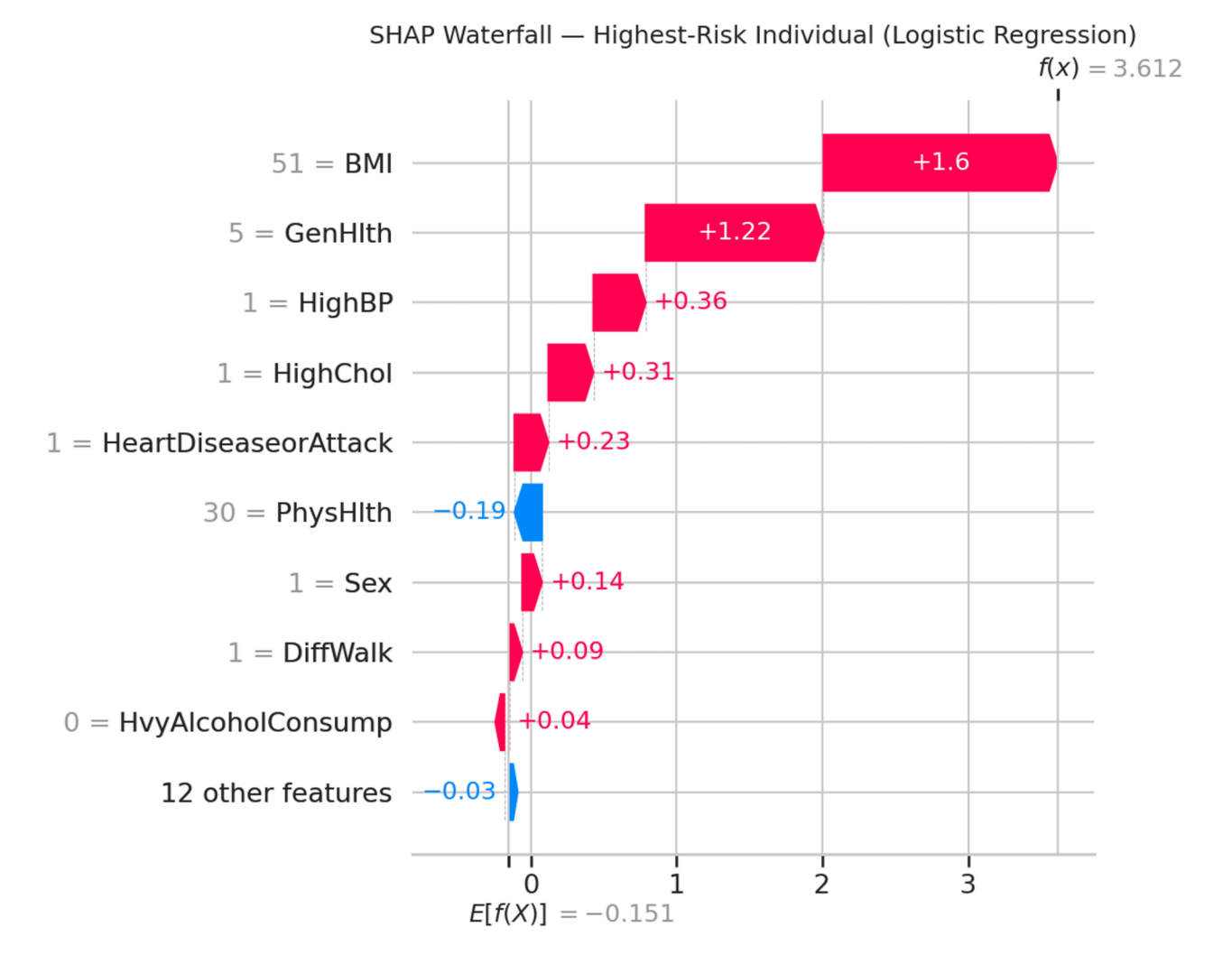


FIGURE 18: SHAP Waterfall Plot—Individual Prediction Explanation
BMI, Body Mass Index; SHAP, SHapley Additive exPlanations

Across all three attribution methods, GenHlth (general health self-rating), BMI, age, and HighBP consistently emerged as the four strongest predictors of diabetes risk. In the random forest Gini ranking, BMI had the highest importance (0.159), followed by GenHlth (0.140), age (0.119), and HighBP (0.110) (Table 8). In the logistic regression coefficient analysis, GenHlth had the largest standardized coefficient (0.632), followed by BMI (0.534), age (0.429), and HighBP (0.355). SHAP analysis corroborates these rankings and additionally reveals directional effects: higher GenHlth scores (indicating poorer self-reported health) substantially increase predicted risk; higher BMI monotonically increases risk; the presence of HighBP consistently raises the predicted risk; and older age (higher BRFSS age bracket) is associated with markedly elevated risk. Behavioral features - particularly physical activity (negative association, indicating a protective effect) and heavy alcohol consumption (negative association, reflecting an inverse population-level relationship) - show modest but clinically interpretable contributions. Income and education carry negative coefficients in logistic regression, consistent with the well-documented socioeconomic gradient in lifestyle disease burden [4].

Rank	Feature	RF Gini Importance	Domain	LR Std. Coefficient
1	BMI	0.159	Clinical	0.534
2	GenHlth	0.140	Clinical	0.632
3	Age	0.119	Demographic	0.429
4	HighBP	0.110	Clinical	0.355
5	Income	0.070	Demographic	−0.119
6	PhysHlth	0.061	Clinical	−0.081
7	HighChol	0.057	Clinical	0.291
8	Education	0.045	Demographic	−0.045
9	MentHlth	0.044	Clinical	−0.034
10	DiffWalk	0.033	Clinical	0.055

TABLE 8: Top-10 Features by RF Gini Importance and LR Standardized Coefficient

BMI, Body Mass Index; LR, Logistic Regression; RF, Random Forest

Aggregate performance dashboard

Figure 19 presents a performance dashboard consolidating key metrics across evaluation protocols, providing an integrated visualization of model behavior across accuracy, discrimination, and calibration dimensions.

An Explainable ML Framework for Early Risk Stratification of Lifestyle Diseases
Performance Dashboard — Diabetes Health Indicators (BRFSS 2015)

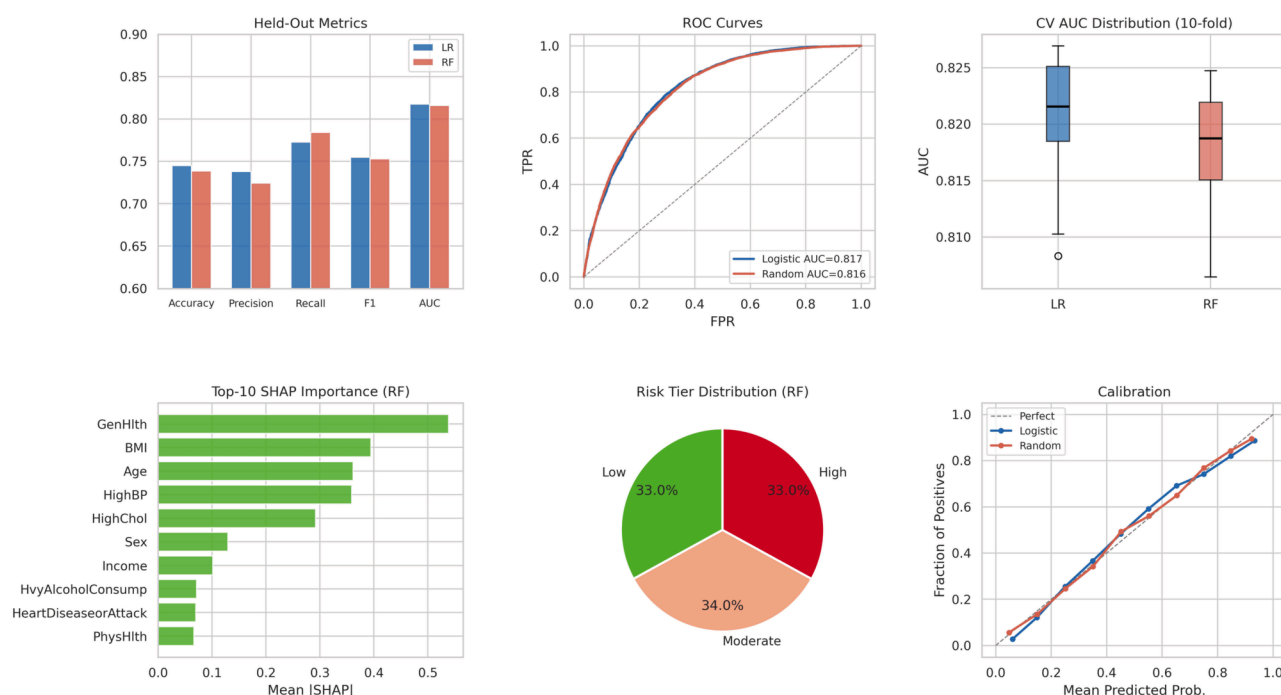


FIGURE 19: Aggregate Performance Dashboard

AUC, Area Under the Curve; BRFSS, Behavioral Risk Factor Surveillance System; CV, Cross-Validation; FPR, False Positive Rate; LR, Logistic Regression; ML, Machine Learning; RF, Random Forest; ROC, Receiver Operating Characteristic; SHAP, SHapley Additive exPlanations; TPR, True Positive Rate

Discussion

The results of this study support several conclusions with direct relevance to clinical and computational healthcare research. First, the integration of multidomain health indicators - spanning clinical measurements, behavioral habits, and sociodemographic factors - within a unified analytical pipeline produces classifiers with clinically meaningful discriminative capacity ($AUC \approx 0.82$) on a large, well-balanced population health dataset. This finding corroborates and extends prior work demonstrating that behavioral covariates improve upon clinical-only models [4,5].

Second, the two models exhibit statistically confirmed complementary strengths rather than a single uniformly dominant algorithm. Logistic regression's higher precision and AUC are advantageous in confirmatory settings where false positives carry significant downstream costs (e.g., expensive diagnostic follow-up). Random forest's higher recall - confirmed as statistically significant by both the paired t-test and McNemar's test - is more valuable in population screening applications where missed cases carry greater clinical risk than false alarms. This distinction has practical implications for selecting deployment contexts and aligns with published guidance on threshold selection in clinical machine learning [14,15].

Third, the SHAP analysis resolves attribution ambiguities that arise from Gini impurity importance alone, which is known to be biased toward high-cardinality and correlated features. The SHAP results confirm that general health self-rating, BMI, high blood pressure, and age are the dominant predictors, consistent with established epidemiological knowledge, while also quantifying the marginal contributions of protective behavioral factors. The SHAP dependence plots reveal

How to cite this article:

Shrivastava B, Fatma T, Kumari S, et al. (October 28, 2026) An Explainable Machine Learning-Based Integrated Predictive Framework for Early Risk Stratification of Lifestyle Diseases Using Multi-Source Healthcare Data. Cureus J Comput Sci 3 : es44389-026-00343-0. DOI <https://doi.org/10.7759/s44389-026-00343-0>

nonlinear effect profiles, particularly for BMI and age, which are not fully captured by logistic regression's linear decision boundary. This suggests that ensemble and deep learning methods may further improve performance in datasets with greater feature interaction complexity [6,7].

Fourth, the calibration advantage of logistic regression (Figure 17) is directly relevant to the risk-stratification application: well-calibrated probabilities ensure that the three-tier risk thresholds reliably correspond to observable proportions of actual cases, a prerequisite for actionable clinical guidance. The moderate overconfidence of random forest in the mid-probability range suggests that probability calibration postprocessing (Platt scaling or isotonic regression) should be considered before clinical deployment.

This study has several limitations that should be acknowledged. The data originate from a single BRFSS survey wave (2015) conducted in the United States; temporal generalizability across different healthcare systems, ethnic populations, and survey periods warrants separate validation. The balanced dataset variant used here was preconstructed to eliminate class imbalance; performance on the original imbalanced distribution may differ and should be evaluated in future work. The explainability analysis was conducted on the test set, but SHAP values were not validated through clinical expert review, representing a gap between computational and clinical explainability. Finally, the framework does not yet incorporate longitudinal trajectory data, wearable sensor streams, or multimodal biomarkers, which could substantially enrich risk prediction in a prospective deployment setting [10,11].

Future work

The proposed framework demonstrates promising performance, but there are other areas that can be explored in future research. First, additional data from wearable sensors can be added to improve risk monitoring. The framework can analyze these data over time and adapt if needed. Second, more advanced techniques such as deep learning can be used to improve the accuracy of predictions, especially when dealing with large amounts of complex data. Third, efforts should be made to develop the framework into a system that clinicians and other healthcare providers can use easily, either online or on their local premises. Efforts can also be made to collect data from people of different races, ages, and backgrounds to ensure that the model performs well across diverse populations.

This is important to ensure that the framework remains effective in predicting risk for people who differ from those in the original sample. Further work could also focus on making the outcomes more understandable so that patients can be informed about their own risk factors and the measures they can take. The proposed framework could be particularly useful if it can be personalized to individuals and its predictions clearly explained to patients.

Conclusions

This study presented an explainable, ML-based integrated predictive framework for the early risk stratification of lifestyle diseases using multidomain BRFSS health indicators. The framework was systematically evaluated through five-fold stratified cross-validation, held-out test performance analysis, statistical significance testing, calibration and learning curve assessment, three-tier probabilistic risk stratification, and SHAP-based explainability analysis. Logistic regression and random forest demonstrated statistically confirmed complementary strengths: logistic regression achieved higher precision and AUC, making it preferable for confirmatory diagnostic support, whereas random forest achieved higher recall and near-equivalent average precision, making it preferable for population-level screening, where sensitivity to positive cases is paramount. McNemar's test ($p = 0.012$) further confirmed that the two models made meaningfully different classification errors. SHAP analysis consistently identified general health status, BMI, high blood pressure, and age as the most influential predictors across attribution methods, providing clinically interpretable and actionable insights.

The integrated, statistically validated, and explainability-enhanced framework addresses key limitations of earlier work in this domain and advances the case for transparent, multidomain ML in preventive healthcare. Future research directions include external validation across heterogeneous populations and healthcare systems, integration of longitudinal and

How to cite this article:

Shrivastava B, Fatma T, Kumari S, et al. (October 28, 2026) An Explainable Machine Learning-Based Integrated Predictive Framework for Early Risk Stratification of Lifestyle Diseases Using Multi-Source Healthcare Data. *Cureus J Comput Sci* 3 : es44389-026-00343-0. DOI <https://doi.org/10.7759/s44389-026-00343-0>

wearable sensor data, comparison with deep learning models using formal statistical evaluation, validation of SHAP values through clinician expert review, post hoc probability calibration of random forest, and development of a clinical decision-support interface for prospective deployment.

Additional Information

Author Contributions

All authors have reviewed the final version to be published and agreed to be accountable for all aspects of the work.

Concept and design: Bhoomika Shrivastava, Sujeet Kumar

Acquisition, analysis, or interpretation of data: Bhoomika Shrivastava, Tuba Fatma, Simran Kumari, Vaibhav Bahuguna, Ankit Kumar, Arbaz .

Drafting of the manuscript: Bhoomika Shrivastava, Tuba Fatma, Simran Kumari, Vaibhav Bahuguna, Ankit Kumar, Arbaz .

Critical review of the manuscript for important intellectual content: Sujeet Kumar

Supervision: Sujeet Kumar

Disclosures

Human subjects: Consent was obtained or waived by all participants in this study. Not applicable issued approval nil. This study utilized the publicly available, de-identified Diabetes Health Indicators dataset derived from the Behavioral Risk Factor Surveillance System (BRFSS). The study involved secondary analysis of anonymized data and did not include direct recruitment, interaction, or intervention with human participants. As only publicly available de-identified data were analyzed, additional Institutional Review Board (IRB) approval and informed consent were not required for this study.

Animal subjects: All authors have confirmed that this study did not involve animal subjects or tissue.

Conflicts of interest: In compliance with [the ICMJE uniform disclosure form](#), all authors declare the following:

- **Payment/services info:** All authors have declared that no financial support was received from any organization for the submitted work.
- **Financial relationships:** All authors have declared that they have no financial relationships at present or within the previous three years with any organizations that might have an interest in the submitted work.
- **Other relationships:** All authors have declared that there are no other relationships or activities that could appear to have influenced the submitted work.

Data Availability Statements

The datasets (and/or code) supporting this study are available from the corresponding author upon reasonable request. The dataset used in this study is publicly available through Kaggle (Diabetes Health Indicators Dataset). Any additional materials and implementation details are available from the corresponding author upon reasonable request.

Acknowledgements

The authors would like to express their sincere gratitude to the faculty mentor for valuable guidance, support, and constructive feedback throughout this research work. The authors also acknowledge the use of the Diabetes Health Indicators Dataset available on Kaggle, which served as the primary dataset for this study. Special thanks to Arbaz for his main work in training the model

How to cite this article:

Shrivastava B, Fatma T, Kumari S, et al. (October 28, 2026) An Explainable Machine Learning-Based Integrated Predictive Framework for Early Risk Stratification of Lifestyle Diseases Using Multi-Source Healthcare Data. Cureus J Comput Sci 3 : es44389-026-00343-0. DOI <https://doi.org/10.7759/s44389-026-00343-0>

References

1. Tresp V, Overhage JM, Bundschus M, Rabizadeh S, Fasching PA, Yu S: [Going digital: a survey on digitalization and large-scale data analytics in healthcare](#). Proceedings of the IEEE. 2016, 104:2180-2206. [10.1109/jproc.2016.2615052](#)
2. Pawar U, O'Shea D, Rea S, O'Reilly R: [Explainable AI in healthcare](#). 2020 International Conference on Cyber Situational Awareness, Data Analytics and Assessment (CyberSA), Dublin, Ireland. 2020, 1-2. [10.1109/CyberSA49311.2020.9139655](#)
3. Rajkomar A, Dean J, Kohane I: [Machine learning in medicine](#). The New England Journal of Medicine. 2019, 380:1347-1358. [10.1056/nejmra1814259](#)
4. Modi K, Singh I, Kumar Y: [A comprehensive analysis of artificial intelligence techniques for the prediction and prognosis of lifestyle diseases](#). Archives of Computational Methods in Engineering. 2023, 30:4733-4756. [10.1007/s11831-023-09957-2](#)
5. Sultana R, Fatema MK: [AI-based diabetes risk stratification using multi-source healthcare data](#). American Journal of Health and Medical Sciences. 2026, 7:128-165. [10.63125/hvmrzzr66](#)
6. Fan Q, Xu G: [Real-time prediction model of public safety events driven by multi-source heterogeneous data](#). Frontiers in Physics. 2025, 13:1553640. [10.3389/fphy.2025.1553640](#)
7. Nazar M, Alam MM, Yafi E, Su'ud MM: [A systematic review of human-computer interaction and explainable artificial intelligence in healthcare with artificial intelligence techniques](#). IEEE Access. 2021, 9:153316-153348. [10.1109/access.2021.3127881](#)
8. Shiddik MAB: [Explainable artificial intelligence in healthcare: current landscape, challenges, and future directions](#). Health Science Reports. 2026, 9:e72172. [10.1002/hsr2.72172](#)
9. Ren H, Wong AB, Lian W, et al.: [Interpretable pneumonia detection by combining deep learning and explainable models with multisource data](#). IEEE Access. 2021, 9:95872-95883. [10.1109/access.2021.3090215](#)
10. Alqudah AM, Moussavi Z: [Bridging signal intelligence and clinical insight: a comprehensive review of feature engineering, model interpretability, and machine learning in biomedical signal analysis](#). Applied Sciences. 2025, 15:12036. [10.3390/app152212036](#)
11. Nithya B, Ilango V: [Predictive analytics in health care using machine learning tools and techniques](#). 2017 International Conference on Intelligent Computing and Control Systems (ICICCS), Madurai, India. 2017, 492-499. [10.1109/ICCONS.2017.8250771](#)
12. Rahmatinejad Z, Dehghani T, Hoseini B, Rahmatinejad F, Lotfata A, Reihani H, Eslami S: [A comparative study of explainable ensemble learning and logistic regression for predicting in-hospital mortality in the emergency department](#). Scientific Reports. 2024, 14:3406. [10.1038/s41598-024-54038-4](#)
13. Agrawal L, Agrawal PK, Agrawal SS, Sonune MS, Kadu RK, Kulkarni MB, Bhaiyya M: [AI-driven multimodal retinal imaging for early detection and risk stratification of vascular and neurodegenerative diseases](#). Graefes Archive for Clinical and Experimental Ophthalmology. 2026, 264:2623-2660. [10.1007/s00417-026-07273-6](#)
14. Chakshu, Jain A, Yanadaiah P: [Explainable AI in heart failure: risk stratification and survival analytics with machine learning](#). 2025 International Conference on Knowledge Engineering and Communication Systems (ICKECS), Chickballapur, India. 2025, 1-7. [10.1109/ICKECS65700.2025.11034821](#)
15. Miao J, Zhu W: [Precision-recall curve \(PRC\) classification trees](#). Evolutionary Intelligence. 2022, 15:1545-1569. [10.1007/s12065-021-00565-2](#)

How to cite this article:

Shrivastava B, Fatma T, Kumari S, et al. (October 28, 2026) An Explainable Machine Learning-Based Integrated Predictive Framework for Early Risk Stratification of Lifestyle Diseases Using Multi-Source Healthcare Data. Cureus J Comput Sci 3 : es44389-026-00343-0. DOI <https://doi.org/10.7759/s44389-026-00343-0>