Journal of Future Artificial Intelligence and Technologies

E-ISSN: 3048-3719

Research Article

An Explainable Machine Learning Framework for Smoking-Associated Health Risk Assessment and Counterfactual Analysis

Somashekar V *

Nitte (Deemed to be University), Nitte Meenakshi Institute of Technology (NMIT), Department of Aeronautical Engineering, Bengaluru, Karnataka 560064, India; e-mail: shekar.mtech.ph.d@gmail.com

* Corresponding Author: Somashekar V

Abstract: This study proposes an explainable machine learning framework for smoking-associated health risk assessment and counterfactual analysis. The objective is to estimate the probability of a model-defined Health Risk Status and examine how this estimated probability changes under a hypothetical modification of smoking status. The study uses the publicly available Body Signal of Smoking dataset from Kaggle, comprising 55,692 records with demographic, anthropometric, physiological, laboratory, and behavioral variables. Health Risk Status was constructed as a binary modeling target by classifying individuals as high risk when at least three of ten selected physiological indicators exceeded their corresponding 75th-percentile thresholds calculated from the training data. This target does not represent a clinically diagnosed disease outcome. The target-generating indicators were excluded from the predictor matrix to reduce direct target leakage. XGBoost was used as the primary model, with sensitivity and ablation analyses examining alternative risk definitions and the contribution of smoking status. Comparative evaluation showed that Random Forest achieved the highest discrimination (ROC-AUC = 0.8920), while XGBoost achieved a ROC-AUC of 0.8348 and was retained for its integration with probability-based counterfactual analysis and SHAP interpretation. Counterfactual analysis compared the observed smoking state with a hypothetical non-smoking state while holding other observed characteristics constant. SHAP analysis identified gender, waist circumference, and LDL as influential predictors, while SHAP stability analysis supported the consistency of the principal feature rankings. The framework provides a model-based approach for assessing health-risk sensitivity to hypothetical smoking-status changes; these estimates should not be interpreted as causal effects or observed clinical outcomes of smoking cessation.

Keywords: Counterfactual Analysis; Explainable Artificial Intelligence; Health Risk Assessment; Machine Learning; SHAP; Smoking Cessation; Smoking Status; XGBoost Classifier.

Received: August, 17th 2026
Revised: September, 7th 2026
Accepted: September, 14th 2026
Published: September, 21st 2026
Curr. Ver.: September, 21st 2026

1. Introduction

Tobacco use remains a major global public-health concern, contributing to approximately 8 million deaths annually, including deaths attributable to both direct tobacco use and exposure to secondhand smoke [1]–[3]. Although smoking is traditionally associated with respiratory diseases and lung cancer, increasing evidence indicates that its effects extend across multiple physiological systems and contribute to cardiovascular, metabolic, hepatic, and renal dysfunction [4]. The complex mixture of thousands of chemicals in tobacco smoke can promote chronic inflammation, oxidative stress, endothelial dysfunction, and cellular damage, thereby affecting multiple biological pathways [5], [6]. These effects may develop progressively and remain clinically unrecognized for prolonged periods, emphasizing the importance of approaches that can identify elevated health risk before severe clinical manifestations occur [7].

Smoking cessation substantially reduces the risk of smoking-related morbidity and mortality, with measurable health benefits occurring after cessation [8]–[12]. However, conventional approaches to smoking-related risk assessment often rely on observed clinical measurements or smoking-status classifications and may not provide an individualized assessment of how predicted health risk could change under alternative smoking conditions [13]–[15]. Consequently, computational approaches capable of integrating multiple biological and behavioral characteristics may provide complementary information for identifying individuals with elevated health risk and exploring alternative risk scenarios [4], [16]–[19]. This suggests that smoking-related risk may be better characterized by considering the combined patterns of demographic, physiological, and behavioral characteristics rather than smoking status in isolation. In this context, a model-based assessment of hypothetical changes in smoking status may provide an additional perspective on the sensitivity of estimated health risk while preserving the observed characteristics of each individual.

Despite advances in machine learning, explainable artificial intelligence, and counterfactual analysis, an important methodological gap remains in integrating these approaches for individualized smoking-related health-risk assessment. Existing studies have largely focused on smoking-status classification, prediction of smoking-associated diseases, or interpretation of observed model predictions, while comparatively less attention has been given to evaluating how model-estimated health risk changes under a specified hypothetical modification of smoking status. The present study addresses this gap by developing a framework that combines probability-based health-risk assessment, explainable machine learning, and constrained counterfactual analysis.

The primary objective of this study is to develop an explainable machine-learning framework for smoking-associated health risk assessment and counterfactual analysis. Specifically, the study aims to:

- Develop an XGBoost-based model to estimate the probability of the model-defined Health Risk Status using available demographic, behavioral, and biological characteristics.
- Identify the demographic, physiological, and behavioral features that contribute most strongly to the model-estimated health-risk probability using SHAP-based analysis.
- Evaluate how the model-estimated health-risk probability changes when smoking status is hypothetically modified from the observed smoking state to a non-smoking state while other observed characteristics remain unchanged.

To achieve these objectives, this study develops a Counterfactual Smoking Risk Assessment Framework that integrates XGBoost-based prediction, SHAP-based interpretation, and counterfactual scenario analysis. Rather than directly predicting smoking status, the framework defines a derived Health Risk Status based on multiple physiological indicators. Importantly, the physiological indicators used to construct the target are excluded from the model predictors to reduce direct target leakage. The framework therefore focuses on assessing a model-defined health-risk state and examining the sensitivity of its predicted probability to a hypothetical modification of smoking status.

The remainder of this paper is organized as follows. Section 2 reviews related studies on smoking-related health assessment, explainable artificial intelligence, and counterfactual analysis. Section 3 describes the study framework, dataset, data preprocessing, model development, experimental configuration, and evaluation procedures. Section 4 presents and discusses the experimental results, including comparative model performance, sensitivity and ablation analyses, SHAP-based interpretation, and counterfactual smoking-state analysis. Finally, Section 5 concludes the study and discusses its implications, limitations, and potential directions for future research.

2. Literature Review

2.1. Machine Learning for Smoking-Related Health Assessment

Machine learning (ML) has increasingly been applied to smoking-related health research, particularly to model relationships among demographic, behavioral, and physiological variables [1], [16], [20], [21]. XGBoost has also been applied to structured smoking-related health and disease-risk modeling [4], [22]–[24]. Existing smoking-related ML studies have predominantly focused on smoking-status classification, prediction of smoking-associated diseases, or

estimation of health-risk categories from observed characteristics [4], [20], [25]. For example, XGBoost-based approaches have been used to predict smoking-induced noncommunicable diseases and identify influential predictors, with analyses primarily focused on predictive performance and feature importance rather than on examining how estimated risk changes under alternative smoking-status conditions [4], [22], [26]. These studies demonstrate the utility of ML methods for identifying relationships between smoking-related characteristics and health outcomes, but they generally treat the observed smoking state as part of the input condition rather than explicitly evaluating hypothetical changes in that state. This distinction provides an opportunity to extend conventional prediction toward individualized, scenario-based health-risk assessment.

Several studies have specifically used the Body Signal of Smoking dataset, although their analytical objectives and target definitions differ from those of the present study. Sung Ahyoung et al. [27] treated smoking status as the binary target and evaluated feature-selection and AutoML-based stacking approaches, reporting an AUC of 0.8209 using a selected subset of predictors. Aishwarya et al. [28] similarly predicted smoking status using a balanced sample of 2,000 observations and evaluated Random Forest, Logistic Regression, Decision Tree, KNN, CatBoost, and artificial neural networks, complemented by SHAP, LIME, QLattice, and Anchor explanations. More recently, Somashekar [29] used the full Body Signal of Smoking dataset to predict the binary smoking label using Random Forest, XGBoost, and Logistic Regression, with SHAP-based interpretation of smoking-associated biomarkers.

In contrast, the present study does not use smoking status as the prediction target. Instead, it constructs a model-defined Health Risk Status from ten physiological indicators, excludes these target-generating variables from the predictor matrix, and retains smoking status as a counterfactual variable. This formulation shifts the analytical focus from predicting smoking status to assessing model-estimated health risk and examining how that estimated probability changes under a specified hypothetical smoking-state modification. The main distinctions between the present study and previous studies using the same dataset are summarized in Table 1.

Table 1. Comparison of Studies Using the Body Signal of Smoking Dataset.

Ref	Target definition	Data/features	Models	Evaluation/ interpretation
[27]	Binary smoking status	55,692 records; feature-selection procedure applied	AutoML and multilayer stacking; RF, XGBoost, CatBoost, and other classifiers	Accuracy, precision, recall, F1-score, AUC
[28]	Binary smoking status	Balanced subset of 2,000 observations from the Body Signal of Smoking dataset	RF, Logistic Regression, Decision Tree, KNN, CatBoost, and ANN	Accuracy, precision, recall, F1-score; SHAP, LIME, QLattice, and Anchor
[29]	Binary smoking status	Full 55,692-record dataset; clinical and biological predictors	Random Forest, XGBoost, and Logistic Regression	Accuracy, F1-score, ROC-AUC, 5-fold CV, SHAP
Ours	Model-defined Health Risk Status: ≥3 of 10 indicators above training-derived P75 thresholds	55,692 records; target-generating physiological indicators excluded from predictors	XGBoost as the primary model; RF, Logistic Regression, and LightGBM for comparison	ROC-AUC, PR-AUC, calibration, Brier score, sensitivity, specificity, F1-score; SHAP stability and counterfactual analysis

2.2. Explainable AI for Health-Risk Assessment

Explainable artificial intelligence (XAI) methods such as SHAP have increasingly been used to interpret machine-learning predictions and identify influential variables in health-risk modeling [30]–[36]. In smoking-related research, recent XAI-based studies have further improved the interpretation of smoking-status predictions by identifying influential clinical and biological variables using approaches such as SHAP, LIME, QLattice, and Anchor [28], [37].

These approaches provide valuable information about the contribution of individual features to observed model predictions. However, feature-level explanations do not directly address an individual-level counterfactual question: how would the model-estimated health risk change if smoking status were hypothetically modified while the individual's other observed characteristics remained unchanged? Thus, conventional ML prediction and feature-level XAI can characterize what is predicted and which variables contribute to the prediction, but provide more limited information about how an individual's predicted risk changes under a specified hypothetical modification [29], [38]–[40].

This distinction is particularly relevant for smoking-related health-risk assessment, where smoking status may be considered not only as an observed characteristic but also as a variable through which alternative risk scenarios can be explored. Accordingly, integrating feature-level explanation with constrained counterfactual analysis can provide complementary information: SHAP can characterize the relative contribution of observed features to the model prediction, while counterfactual analysis can examine changes in the model-estimated probability under a predefined alternative smoking state. This complementary perspective provides the basis for the counterfactual analysis developed in the present study.

2.3. Counterfactual Explainability

Counterfactual analysis extends model interpretation by examining how predictions change under hypothetical modifications of input variables. Recent studies have emphasized the importance of feasibility, actionability, and causal constraints when constructing counterfactual explanations. For example, DiFACE2 introduced a causal-DAG-constrained framework for generating diverse, feasible, and actionable counterfactual explanations, highlighting the role of domain knowledge and causal structure in intervention-oriented counterfactual analysis [41].

In smoking-related health assessment, counterfactual analysis can provide a complementary perspective to feature-level explanations by examining how model-estimated health risk changes under an alternative smoking-status condition. In the present study, smoking status is therefore modified from the observed smoking state to a hypothetical non-smoking state while other observed characteristics are held constant. The analysis does not construct a causal DAG or estimate causal intervention effects; rather, it evaluates the sensitivity of the model-estimated health-risk probability to this specified counterfactual change at the individual level.

Together with the studies discussed in the preceding sections, this perspective provides an opportunity to extend smoking-related health assessment beyond smoking-status prediction and conventional feature-level interpretation toward individualized, scenario-based analysis. The resulting counterfactual estimates represent model-based hypothetical scenarios and should not be interpreted as observed longitudinal effects or causal treatment effects of smoking cessation.

3. Materials and Methods

This study develops a Counterfactual Smoking Risk Assessment Framework that integrates health-risk assessment, explainable machine learning, and counterfactual analysis. The framework estimates the probability of a model-defined Health Risk Status from demographic, anthropometric, behavioral, and physiological characteristics and examines how this estimated probability changes when smoking status is hypothetically modified. Because the dataset records smoking status as a binary variable and does not provide quantitative smoking-exposure measures such as cigarettes consumed per day, smoking duration, or pack-years, the counterfactual analysis is limited to the observed smoking state and a hypothetical non-smoking state. The overall methodological workflow is illustrated in Figure 1.

3.1. Study Framework and Dataset

The study uses the Body Signal of Smoking dataset obtained from Kaggle, comprising 55,692 records and 27 variables representing demographic, anthropometric, behavioral, physiological, and laboratory characteristics [42]. The framework is not intended to estimate cumulative smoking exposure or reconstruct individual smoking histories. Instead, the available physiological and biochemical measurements are used to characterize the observed health-risk profile of the study population. Variables such as blood pressure, lipid measures, blood

glucose, waist circumference, liver-related biomarkers, and renal indicators represent different aspects of cardiovascular, metabolic, hepatic, and renal health. These variables are therefore treated as health-related characteristics within the framework rather than as direct measures or surrogates of cumulative smoking exposure.

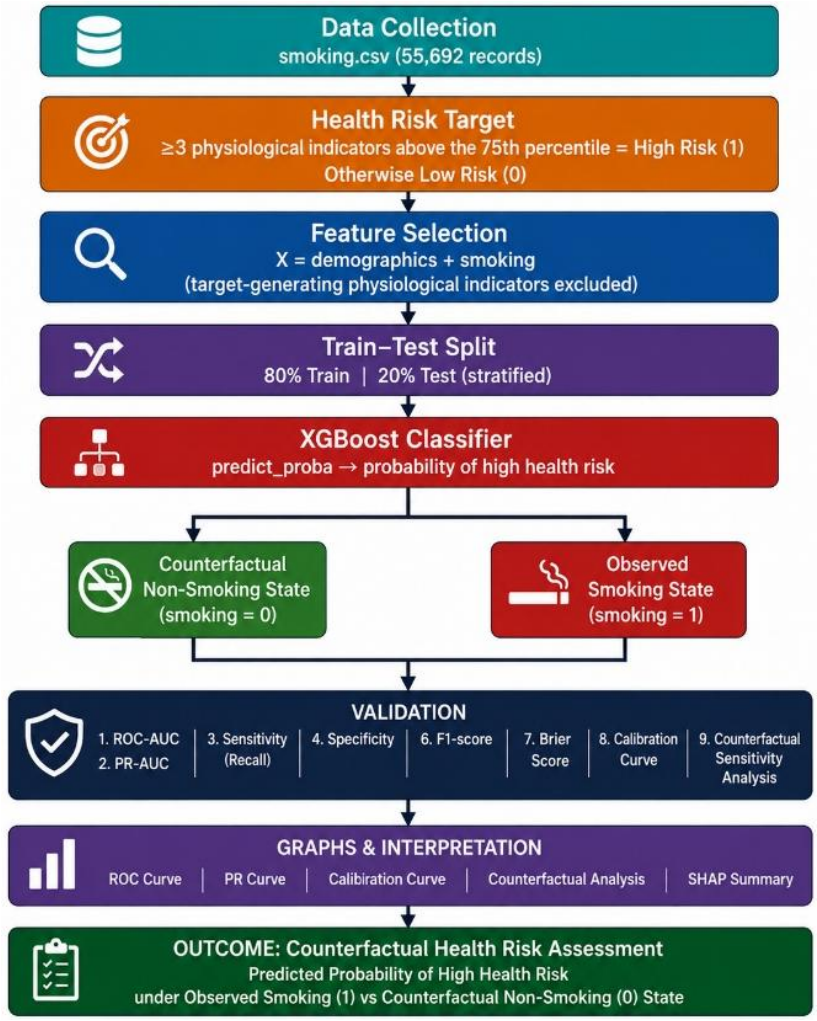

Figure 1. Counterfactual Smoking Risk Assessment Framework based on XGBoost, SHAP, and counterfactual analysis.

3.2 Data Preprocessing and Variable Definition

The dataset contains demographic variables including age, gender, height, and weight; anthropometric measurements such as waist circumference; physiological measurements including systolic and diastolic blood pressure; biochemical variables including fasting blood glucose, total cholesterol, triglycerides, HDL, LDL, hemoglobin, serum creatinine, AST, ALT, and GTP; and behavioral or health-related variables including hearing, dental health, and smoking status. Together, these variables provide a multidimensional representation of the characteristics used for health-risk assessment.

Smoking status is retained as a predictor rather than used as the primary prediction target. Instead, a derived binary target, termed Health Risk Status, is constructed to represent the presence of multiple elevated physiological indicators. Ten indicators are used to construct this target: systolic blood pressure, diastolic blood pressure, fasting blood glucose, total cholesterol, triglycerides, hemoglobin, serum creatinine, AST, ALT, and GTP.

For each indicator, the percentile threshold is calculated using the training data only to prevent information leakage. The primary analysis uses the 75th percentile (P75) as the threshold. An indicator is considered elevated when its value exceeds the corresponding training-set threshold, and the number of elevated indicators is then calculated for each individual. The Health Risk Status is defined as:

$$\text{Health Risk Status}_i = \begin{cases} 1, & \sum_{j=1}^{10} I\left(X_{ij} > T_j\right) \geq 3, \\ 0, & \sum_{j=1}^{10} I\left(X_{ij} > T_j\right) < 3, \end{cases} \quad (1)$$

where X_{ij} represents the value of indicator j for individual i, T_j represents the 75th-percentile threshold of indicator j calculated from the training data, and $I(\cdot)$ is an indicator function that takes a value of 1 when the condition is satisfied and 0 otherwise. Thus, an individual is classified as high risk when at least three of the ten selected indicators exceed their corresponding training-derived thresholds.

The physiological variables used to construct Health Risk Status are excluded from the predictor matrix used for model training. This design prevents direct target leakage and ensures that the model does not reproduce the target using the same measurements from which it was constructed.

Before model development, the resulting Health Risk Status distribution was examined to assess class proportions. In the complete dataset, 21,792 individuals (39.13%) were classified as high risk and 33,900 (60.87%) as low risk. The training and testing subsets contained 17,481 (39.24%) and 4,311 (38.70%) high-risk observations, respectively, indicating that the class proportions were preserved across the two subsets. The resulting distribution is presented in Table 2.

Table 2. Distribution of Health Risk Status

Dataset	Low Risk, n (%)	High Risk, n (%)	Total
Overall	33,900 (60.87%)	21,792 (39.13%)	55,692
Training	27,072 (60.76%)	17,481 (39.24%)	44,553
Testing	6,828 (61.30%)	4,311 (38.70%)	11,139

3.3. Feature Matrix and Data Partitioning

Following the construction of the Health Risk Status target, the predictor matrix was defined using the available demographic, anthropometric, behavioral, and remaining health-related variables. The ten physiological indicators used to construct the target were excluded from the predictor matrix to avoid direct target leakage. Accordingly, the predictor matrix can be represented as

$$\mathbf{X} = [\mathbf{X}_{\text{demo}}, \mathbf{X}_{\text{anthro}}, \mathbf{X}_{\text{behavior}}, \mathbf{X}_{\text{health}}] \quad (2)$$

where $\mathbf{X}_{\text{demo}}$, $\mathbf{X}_{\text{anthro}}$, $\mathbf{X}_{\text{behavior}}$, and $\mathbf{X}_{\text{health}}$ denote the demographic, anthropometric, behavioral, and remaining health-related predictors, respectively. The target-generating physiological indicators are not included in $\mathbf{X}_{\text{health}}$.

The dataset was then divided into training and independent testing subsets using an 80:20 stratified split. Stratification was applied to preserve the proportion of high- and low-risk observations across the two subsets, while a fixed random seed of 42 was used to ensure reproducibility. The training subset was used for model development and hyperparameter selection, whereas the independent test set was reserved for final performance evaluation and counterfactual analysis. The resulting partition can be expressed as

$$\mathcal{D} = \mathcal{D}_{\text{train}} \cup \mathcal{D}_{\text{test}}, \quad \mathcal{D}_{\text{train}} \cap \mathcal{D}_{\text{test}} = \emptyset \quad (3)$$

with

$$|\mathcal{D}_{\text{train}}| = 0.80|\mathcal{D}|, \quad |\mathcal{D}_{\text{test}}| = 0.20|\mathcal{D}| \quad (4)$$

The Health Risk Status thresholds were estimated from the training data and subsequently applied to the independent test set. No test-set information was used for model development, hyperparameter selection, or other model-selection procedures.

3.4. Experimental Configuration and XGBoost Model Settings

All experiments were implemented in Python using the XGBoost and scikit-learn libraries. Categorical variables were converted to numerical representations using label encoding, and observations containing missing values were removed. Model development and hyperparameter selection were performed exclusively using the training subset, while the independent test set was retained for final evaluation.

Hyperparameter selection was performed using GridSearchCV with 5-fold stratified cross-validation, shuffling, and a fixed random seed of 42. ROC-AUC was used as the selection criterion because the study evaluates discrimination between the model-defined high- and low-risk groups across classification thresholds. For XGBoost, the search covered the parameter combinations listed in Table 3. The best estimator identified during cross-validation was subsequently refitted on the complete training subset and evaluated on the independent test set.

Table 3. Model configurations and hyperparameter search spaces

Model	Tuning procedure	Cross-validation	Selection metric	Search space	Final selected configuration
Logistic Regression	GridSearchCV	5-fold stratified CV	ROC-AUC	C = [0.01, 0.1, 1, 10]; class_weight = [None, balanced]; L2 penalty	C = 1; class_weight = balanced; penalty = L2
Random Forest	GridSearchCV	5-fold stratified CV	ROC-AUC	n_estimators = [200, 400]; max_depth = [None, 10, 20]; min_samples_split = [2, 10]; min_samples_leaf = [1, 4]	n_estimators = 400; max_depth = None; min_samples_split = 2; min_samples_leaf = 1
XGBoost	GridSearchCV	5-fold stratified CV	ROC-AUC	n_estimators = [200, 400]; max_depth = [3, 5, 7]; learning_rate = [0.03, 0.1]; subsample = [0.8, 1.0]; colsample_bytree = [0.8, 1.0]; min_child_weight = [1, 5]; gamma = [0, 1]	n_estimators = 400; max_depth = 7; learning_rate = 0.1; subsample = 0.8; colsample_bytree = 1.0; min_child_weight = 1; gamma = 0
LightGBM	GridSearchCV	5-fold stratified CV	ROC-AUC	n_estimators = [200, 400]; max_depth = [3, 5, 7]; learning_rate = [0.03, 0.1]; num_leaves = [31, 63]; subsample = [0.8, 1.0]; colsample_bytree = [0.8, 1.0]; min_child_samples = [5, 20]	n_estimators = 400; max_depth = 7; learning_rate = 0.1; num_leaves = 63; subsample = 0.8; colsample_bytree = 0.8; min_child_samples = 5

The final XGBoost model used the binary logistic objective function and generated probability estimates through its probability output. The selected configuration comprised n_estimators = 400, max_depth = 7, learning_rate = 0.1, subsample = 0.8, colsample_bytree = 1.0, min_child_weight = 1, and gamma = 0, with random_state = 42. This selected XGBoost estimator was used for the final test-set evaluation, SHAP interpretation, and counterfactual analysis.

For the sensitivity and ablation analyses reported in Table 5, an earlier XGBoost configuration was used. These analyses followed the same leakage-controlled target and data-partitioning framework but were conducted using a previous model configuration and are therefore treated as a separate robustness analysis rather than as results from the final selected XGBoost model.

3.5. XGBoost-Based Health Risk Assessment

XGBoost was selected as the primary model for health-risk assessment because of its suitability for structured tabular data and its ability to capture nonlinear relationships and interactions among heterogeneous demographic, anthropometric, behavioral, and health-related variables [43]–[45]. In addition, its probability-based output supports the counterfactual analysis, while its tree-based structure enables SHAP-based interpretation.

The selected XGBoost model was fitted using the training subset and the predictor matrix defined in Section 3.3. Rather than producing only a binary class label, the model generates an estimated probability that an individual belongs to the model-defined high-risk category. For individual i, this probability is denoted by p_i.

The probability output serves as the primary quantity for the subsequent counterfactual analysis. Specifically, the estimated probability under the observed smoking state is compared with the probability obtained after the smoking variable is hypothetically modified, while all other observed predictor values remain unchanged. This allows the analysis to characterize individual-level changes in model-estimated health-risk probability while maintaining the observed characteristics of each individual.

3.6. Sensitivity and Ablation Analysis

To assess the robustness of the Health Risk Status definition and examine the contribution of smoking status to model performance, sensitivity and ablation analyses were conducted using the same target-construction and data-partitioning framework. First, the sensitivity of the Health Risk Status definition to the percentile threshold was evaluated by varying the threshold from the 70th to the 85th percentile while maintaining the requirement that at least three physiological indicators exceed their corresponding thresholds. Specifically, the 70th, 75th, 80th, and 85th percentile configurations were evaluated, with the 75th-percentile definition retained as the primary configuration. All thresholds were calculated exclusively from the training data and subsequently applied to the test data.

Second, sensitivity to the number of elevated physiological indicators required for high-risk classification was examined while maintaining the 75th-percentile threshold. The required number of elevated indicators was varied from two to five to assess whether model performance was sensitive to the selected criterion of at least three elevated indicators. Performance across the alternative definitions was evaluated using ROC-AUC, PR-AUC, sensitivity, specificity, precision, F1-score, and Brier score, providing complementary information on discrimination, classification performance, and probability reliability.

Finally, an ablation analysis was conducted to examine the contribution of smoking status to health-risk assessment. Two otherwise identical XGBoost models were evaluated using the same training and test partitions and predictor framework. The first model included smoking status as a predictor, whereas the second excluded the smoking-status variable. Differences in ROC-AUC, PR-AUC, sensitivity, specificity, precision, F1-score, and Brier score were used to assess the additional predictive information associated with smoking status within the model. Together, these analyses provide complementary checks on the robustness of the model-defined target and the role of smoking status in the predictive framework.

3.7. Comparative Classifier Analysis

Four tabular machine-learning classifiers—Logistic Regression, Random Forest, XGBoost, and LightGBM—were evaluated using the same predictor matrix and stratified training/test partition. All classifiers were tuned using GridSearchCV with 5-fold stratified cross-validation, shuffling, a fixed random seed of 42, and ROC-AUC as the model-selection criterion. Each algorithm was evaluated using a model-specific hyperparameter grid, with hyperparameter selection performed exclusively on the training subset. The selected estimators were then refitted on the complete training subset and evaluated on the independent test set.

The classifiers were compared using ROC-AUC, PR-AUC, sensitivity (recall), specificity, precision, F1-score, accuracy, and Brier score. ROC-AUC and PR-AUC were used to assess discrimination, while sensitivity, specificity, precision, F1-score, and accuracy were calculated at the predefined classification threshold. Brier score was used to assess the reliability of the predicted probabilities. Applying the same data partitioning and evaluation protocol across classifiers enabled a consistent comparison of discrimination, classification performance, and probability estimation.

3.8. Counterfactual Smoking-State Analysis

The counterfactual analysis examines how the model-estimated health-risk probability changes when the smoking-status variable is hypothetically modified. Because smoking status is recorded as a binary variable and the dataset does not provide quantitative smoking-exposure measures such as cigarettes per day, smoking duration, or pack-years, the analysis is restricted to the observed smoking state and a hypothetical non-smoking state. Intermediate values such as 0.25, 0.50, and 0.75 are therefore not considered because they do not correspond to observed or clinically defined levels of smoking exposure in the dataset.

The analysis is restricted to individuals observed as smokers, for whom the smoking-status variable has a value of 1. For each eligible individual, a copy of the observed predictor vector is created, and only the smoking-status variable is changed from 1 to 0. All other observed characteristics are kept unchanged. The final XGBoost model is then applied to both the observed and modified records to obtain the corresponding model-estimated high-risk probabilities. The observed and counterfactual probabilities are defined as:

$$p_i^{(\text{obs})} = f(\mathbf{X}_i, \text{Smoking} = 1), \quad p_i^{(\text{cf})} = f(\mathbf{X}_i, \text{Smoking} = 0) \tag{5}$$

where $\mathbf{X}_i$ represents the other observed predictor characteristics of individual i, $f(\cdot)$ denotes the trained XGBoost model. The individual-level counterfactual probability difference is then calculated as:

$$\Delta p_i = p_i^{(\text{obs})} - p_i^{(\text{cf})} \tag{6}$$

A positive Δp_i indicates that the model estimates a higher health-risk probability under the observed smoking state than under the hypothetical non-smoking state. The distribution of Δp_i across the evaluated individuals is used to characterize the direction and magnitude of the model-estimated changes.

The counterfactual analysis represents model-based sensitivity to a hypothetical change in smoking status rather than an observed longitudinal response or causal treatment effect. Accordingly, changes in predicted probability should not be interpreted as direct evidence of health recovery following smoking cessation. Robustness of the counterfactual findings is further considered through the sensitivity and ablation analyses described in Section 3.6.

3.9. Model Interpretation Using SHAP

SHapley Additive exPlanations (SHAP) was used to interpret the contribution of individual predictors to the XGBoost model output. SHAP values were calculated for the independent test set after final model training. Mean absolute SHAP values were used to quantify the overall contribution of each predictor and establish the global feature-importance ranking. Larger mean absolute SHAP values indicate a greater contribution to the model output, although they do not imply a causal effect.

To assess the stability of the model explanations, SHAP values were additionally evaluated across repeated subsets of the independent test set. For each subset, mean absolute SHAP values were calculated and the corresponding feature rankings were recorded. Explanation stability was assessed using the mean feature rank, rank variability, and the frequency with which a feature appeared among the five highest-ranked predictors. Global SHAP importance and ranking stability were visualized using the resulting feature-importance and ranking measures.

3.10. Model Performance Evaluation

Multiple complementary metrics were used to evaluate classification performance and the reliability of predicted probabilities. ROC-AUC was used to assess discrimination between high- and low-risk individuals across classification thresholds. PR-AUC was also reported to provide complementary information on discrimination, particularly under alternative Health Risk Status definitions with different class proportions. Calibration analysis was used to assess the agreement between predicted probabilities and observed outcome frequencies. Predicted probabilities were grouped into probability ranges and compared with the corresponding observed proportions. The Brier score was additionally calculated to quantify the accuracy of probabilistic predictions, with lower values indicating better probability reliability.

For threshold-dependent evaluation, sensitivity (recall), specificity, precision, F1-score, and accuracy were calculated at the predefined classification threshold. Sensitivity represents the proportion of high-risk individuals correctly identified, whereas specificity represents the proportion of low-risk individuals correctly classified. Precision represents the proportion of individuals predicted as high risk who were actually high risk, while the F1-score provides a balance between precision and recall. All performance metrics were calculated on the independent test set using the same evaluation procedure for the final XGBoost model and the comparative classifiers. The distribution of Health Risk Status was also reported to provide context for interpreting the classification and probability-based performance measures.

4. Results and Discussion

4.1. Predictive Performance of the XGBoost Model

The predictive performance of the XGBoost model was first evaluated on the independent test set using discrimination, classification, and probabilistic metrics. The model achieved a ROC-AUC of 0.8348 and a PR-AUC of 0.7635. At the predefined classification threshold, the model obtained a sensitivity of 0.6698, specificity of 0.8280, precision of 0.7243, F1-score of 0.6960, and accuracy of 0.7643, with a Brier score of 0.1616. These results indicate that the model provides reasonable discrimination and probability estimation for the model-defined Health Risk Status.

The ROC-AUC provides an overall measure of discrimination between the model-defined high- and low-risk categories, while the PR-AUC and threshold-based metrics provide complementary information on classification performance. The Brier score further evaluates the reliability of the predicted probabilities, which is particularly relevant because these probabilities are subsequently used in the counterfactual analysis. Since the target represents a derived physiological-risk category rather than a clinically diagnosed outcome, these performance measures should be interpreted as evidence of predictive performance within the study dataset rather than as evidence of clinical diagnostic performance.

Figure 2 presents the calibration curve of the XGBoost classifier. The x-axis represents the mean predicted probability, while the y-axis represents the observed fraction of positive cases (high-risk individuals). The dashed gray diagonal represents perfect calibration, where predicted probabilities correspond exactly to observed outcome frequencies. A calibration curve that remains close to this reference therefore indicates better agreement between predicted and observed probabilities.

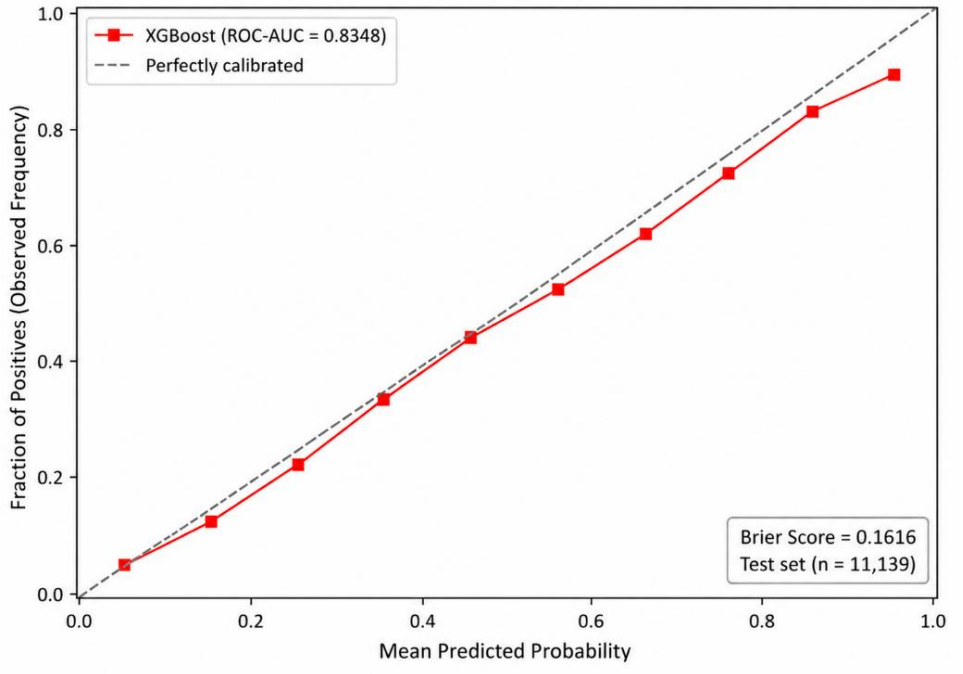

Figure 2. Calibration curve of the XGBoost classifier for health-risk probability prediction.

The red curve remains close to the diagonal across most probability ranges, indicating generally consistent probability estimation. At lower predicted probabilities (0.05–0.35), the curve shows close alignment with the reference line, while the mid-probability range (0.45–0.75) also exhibits relatively stable calibration. At the highest predicted-probability range, a slight deviation from the diagonal is observed, with predicted probabilities marginally lower than the observed fraction, indicating minor under confidence for very high-risk cases. Overall, the deviation is limited relative to the general calibration pattern, supporting the use of the predicted probabilities for the subsequent counterfactual analysis.

4.2. Comparative Classifier Analysis

Table 4 compares the performance of four tabular machine learning classifiers using the leakage-controlled Health Risk Status definition. Random Forest achieved the highest standalone discrimination, with a ROC-AUC of 0.8920 and a PR-AUC of 0.8531, together with a sensitivity of 0.7451, specificity of 0.8587, precision of 0.7806, and F1-score of 0.7624. LightGBM achieved a ROC-AUC of 0.8326 and PR-AUC of 0.7647, whereas XGBoost achieved a ROC-AUC of 0.8348 and PR-AUC of 0.7635. Logistic Regression produced a ROC-AUC of 0.7926 and PR-AUC of 0.7023. The complete comparison, including calibration-related and threshold-based metrics, is presented in Table 4.

Table 4. Comparative Performance of Tabular Machine Learning Classifiers

Model	ROC-AUC	PR-AUC	Sensitivity	Specificity	Precision	F1-score	Brier Score	Accuracy
Logistic Regression	0.7926	0.7023	0.7420	0.6985	0.6242	0.6780	0.1865	0.7160
Random Forest	0.8920	0.8531	0.7451	0.8587	0.7806	0.7624	0.1315	0.8129
XGBoost	0.8348	0.7635	0.6698	0.8280	0.7243	0.6960	0.1616	0.7643
LightGBM	0.8326	0.7647	0.6562	0.8256	0.7174	0.6854	0.1628	0.7573

The substantially higher Random Forest performance warrants consideration of whether the result could arise from target leakage or data-partitioning artifacts. The Health Risk Status target was constructed exclusively from the ten predefined physiological indicators, which were subsequently excluded from the predictor matrix. In addition, the percentile thresholds used for target construction were estimated from the training data only and then applied to the independent test set. All classifiers used the same stratified 80:20 training–testing partition, leakage-controlled predictor matrix, and test-set isolation during model development. These procedures reduce the possibility that the observed Random Forest performance resulted from direct inclusion of the target-generating variables or from inconsistent data partitioning.

Nevertheless, the performance of Random Forest should be interpreted in the context of the derived target and the available cross-sectional feature space. Although the ten target-generating indicators were excluded, some remaining demographic, anthropometric, behavioral, or health-related variables may still be associated with the underlying physiological burden represented by the constructed label. Random Forest may therefore capture complex nonlinear relationships among these proxy characteristics and the model-defined Health Risk Status. This interpretation does not imply equivalent performance for independently measured clinical outcomes.

From a framework perspective, XGBoost was retained as the primary model despite the higher standalone predictive performance of Random Forest. The selection was based on the broader analytical objective of integrating probability-based counterfactual analysis with SHAP-based interpretation and stability analysis. In particular, the XGBoost probability output can be recalculated after the controlled modification of the smoking-status variable, enabling individual-level comparison between observed and hypothetical model outputs. Its tree-based structure also supports consistent SHAP decomposition of model predictions. Thus, the selection of XGBoost represents a framework-level methodological choice rather than a claim of superior standalone predictive performance.

This comparison also highlights an important distinction between predictive performance and framework suitability. Random Forest produced stronger discrimination under the present target definition, whereas XGBoost provided the modeling basis used for the subsequent counterfactual and explainability analyses. The observed differences should therefore be understood as empirical performance characteristics of the evaluated models rather than as evidence that one algorithm is intrinsically preferable across other datasets or clinically defined outcomes.

4.3. Sensitivity and Ablation Analysis

Sensitivity and ablation analyses were conducted to examine the robustness of the Health Risk Status definition and the contribution of smoking status to model performance. Table 5 summarizes the results obtained under alternative percentile thresholds, different numbers of elevated indicators required to define high risk, and the inclusion or exclusion of smoking

status. For clarity, these analyses were conducted using an earlier XGBoost configuration within the same leakage-controlled modeling framework and are therefore not directly comparable with the final XGBoost results reported in Table 4.

Across the percentile-based sensitivity analysis, the ROC-AUC remained relatively stable when the threshold was varied from P70 to P85, ranging from 0.8396 to 0.8430. However, the PR-AUC decreased as the percentile threshold increased, from 0.8281 at P70 to 0.5914 at P85. This pattern indicates that changing the threshold affects not only discrimination but also the composition of the high-risk group captured by the derived target. The Brier score also decreased with higher thresholds, suggesting differences in probability error under alternative target definitions. These results indicate that the model's discrimination is relatively stable across percentile thresholds, whereas the resulting probability and positive-class characteristics are more sensitive to the target definition.

Table 5. Sensitivity and Ablation Analysis of the Health Risk Status Definition and Smoking Status

Analysis	Percentile Threshold	Elevated Indicators	Smoking Status	ROC-AUC	PR-AUC	Brier Score
Threshold sensitivity	P70	≥3	Yes	0.8430	0.8281	0.1620
Threshold sensitivity	P75	≥3	Yes	0.8407	0.7722	0.1578
Threshold sensitivity	P80	≥3	Yes	0.8397	0.7115	0.1471
Threshold sensitivity	P85	≥3	Yes	0.8396	0.5914	0.1168
Indicator-count sensitivity	P75	≥2	Yes	0.8376	0.8710	0.1627
Indicator-count sensitivity	P75	≥3	Yes	0.8407	0.7722	0.1578
Indicator-count sensitivity	P75	≥4	Yes	0.8527	0.6743	0.1288
Indicator-count sensitivity	P75	≥5	Yes	0.8679	0.5755	0.0942
Smoking ablation	P75	≥3	No	0.8383	0.7687	0.1590
Smoking ablation	P75	≥3	Yes	0.8407	0.7722	0.1578

The sensitivity analysis based on the required number of elevated indicators produced a broader range of ROC-AUC values, from 0.8376 for ≥2 indicators to 0.8679 for ≥5 indicators. The corresponding PR-AUC decreased from 0.8710 to 0.5755 as the required number of elevated indicators increased. Thus, a stricter definition of high risk produces a more selective target and changes the balance between discrimination and positive-class representation. The primary P75/≥3 definition provides an intermediate formulation that retains a substantial high-risk group while avoiding an excessively broad or restrictive definition. This balance supports its use as the primary target definition while acknowledging that the choice of threshold and indicator count influences the resulting predictive task.

The smoking ablation analysis showed only a modest difference between models with and without smoking status under the earlier XGBoost configuration. Including smoking status increased the ROC-AUC from 0.8383 to 0.8407 and the PR-AUC from 0.7687 to 0.7722, while the Brier score changed from 0.1590 to 0.1578. The relatively small change suggests that smoking status provides additional predictive information within the modeled risk framework, but that the derived Health Risk Status can also be substantially characterized by the other available predictors. Because these ablation results were obtained using the earlier XGBoost configuration, they should be interpreted as a robustness analysis of the role of smoking status rather than as a direct decomposition of the final model performance reported in Table 4.

Overall, the sensitivity and ablation analyses indicate that the study conclusions are not solely dependent on a single threshold specification or on the inclusion of smoking status. At the same time, the changes in PR-AUC and Brier score across alternative target definitions demonstrate that the operational definition of Health Risk Status materially influences the prediction task. This finding is relevant to the interpretation of the subsequent counterfactual analysis, which is conditioned on the primary P75/≥3 definition.

4.4. SHAP-Based Model Interpretation and Stability Analysis

To further interpret the final XGBoost model, SHAP analysis was used to quantify the contribution of individual predictors to model output. Figure 3 presents the global SHAP

feature importance together with the corresponding ranking stability analysis. The SHAP analysis identified gender, waist circumference, LDL, weight, and age among the most influential predictors of the model-estimated Health Risk Status. Height, smoking status, and HDL also contributed to the model output, although their relative influence was lower than that of the principal features. The distribution of feature importance indicates that the model does not rely predominantly on smoking status alone. Instead, the estimated health-risk probability reflects a broader combination of demographic, anthropometric, and health-related characteristics.

The prominence of waist circumference, LDL, weight, and age is consistent with the multidimensional structure of the derived Health Risk Status, which is constructed from multiple physiological indicators. Since the target represents the simultaneous elevation of several physiological measures, the model may capture relationships between these characteristics and the broader physiological profile represented by the constructed label. The SHAP results should therefore be interpreted as model-attribution patterns rather than as evidence of causal relationships between individual predictors and health risk.

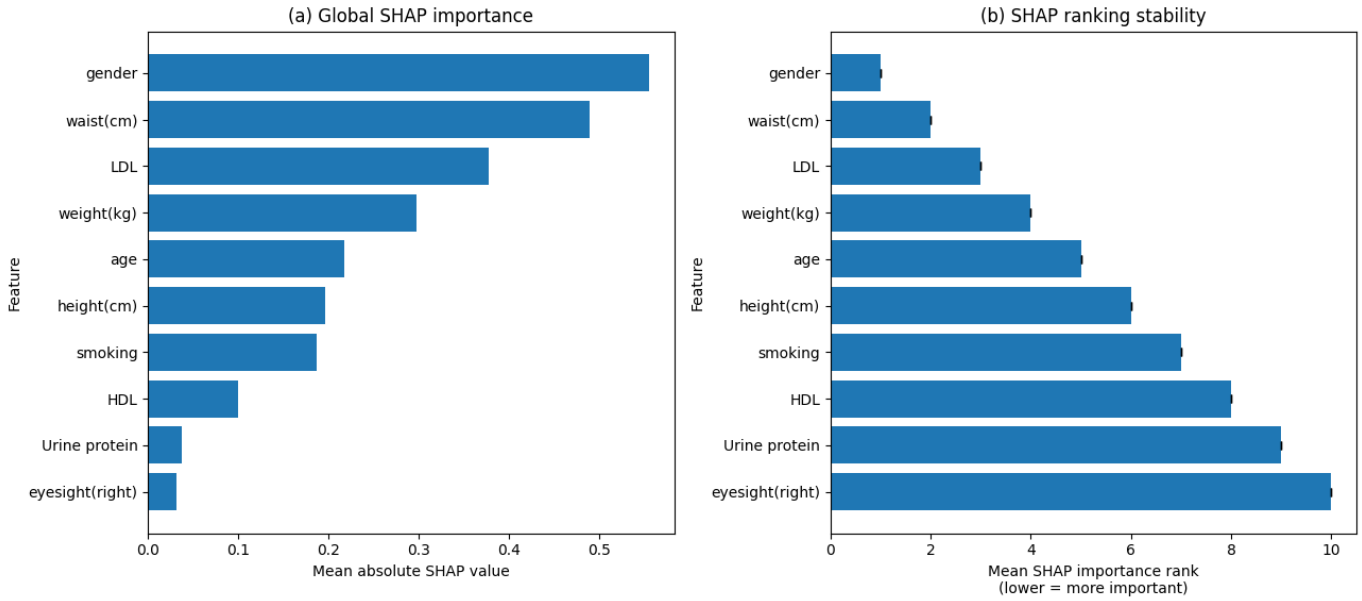

Figure 3. SHAP-based feature importance and ranking stability for the final XGBoost model.

The stability analysis further examined whether the principal feature rankings were consistent across repeated subsets of the independent test set. The leading features remained broadly consistent across these subsets, indicating that the main attribution pattern was not strongly dependent on a single subset of test observations. This provides additional support for interpreting the principal SHAP rankings as a relatively stable characteristic of the fitted model rather than as an artifact of one particular test subset.

The role of smoking status can also be considered in relation to the ablation analysis in Section 4.3. Smoking contributed to the model output but was not among the most dominant predictors identified by the global SHAP analysis. Together, these findings suggest that smoking status provides additional information within the modeled risk framework while the estimated risk remains influenced by a broader set of individual characteristics. The ablation evidence, however, was obtained using an earlier XGBoost configuration and should therefore be interpreted as complementary rather than as a direct quantitative decomposition of the final model.

Taken together, the SHAP importance and stability analyses provide two complementary perspectives on model interpretation: the SHAP values characterize the relative contribution of predictors to the estimated risk, while the stability analysis examines whether the principal feature-ranking pattern is consistent across test subsets. These results provide the interpretive basis for the subsequent counterfactual analysis, in which smoking status is modified while the other observed characteristics are held constant.

4.5. Counterfactual Smoking-State Analysis

Figure 4 illustrates the counterfactual analysis of health-risk probability under the observed and hypothetical non-smoking states. The analysis examines how the model-estimated health-risk probability changes when smoking status is hypothetically modified from smoking to non-smoking while all other observed characteristics are held constant.

The x-axis represents the first 100 smokers selected from the independent test set, and the y-axis represents the health-risk probability predicted by the final XGBoost model. The red line indicates the predicted probability under the observed smoking state, whereas the green line represents the recalculated probability after setting the smoking-status variable to zero. The difference between the two curves therefore represents the change in model-estimated health-risk probability associated with the specified counterfactual input modification.

Figure 4 shows that the hypothetical non-smoking state generally resulted in a lower model-estimated health-risk probability for many individuals, although both the magnitude and direction of the changes varied across individuals. This variation indicates that the model's sensitivity to smoking status is conditioned by the broader predictor profile rather than determined by smoking status alone. The result therefore complements the global SHAP analysis in Section 4.4 by illustrating how the contribution of a single variable can translate into different changes in predicted probability across individuals.

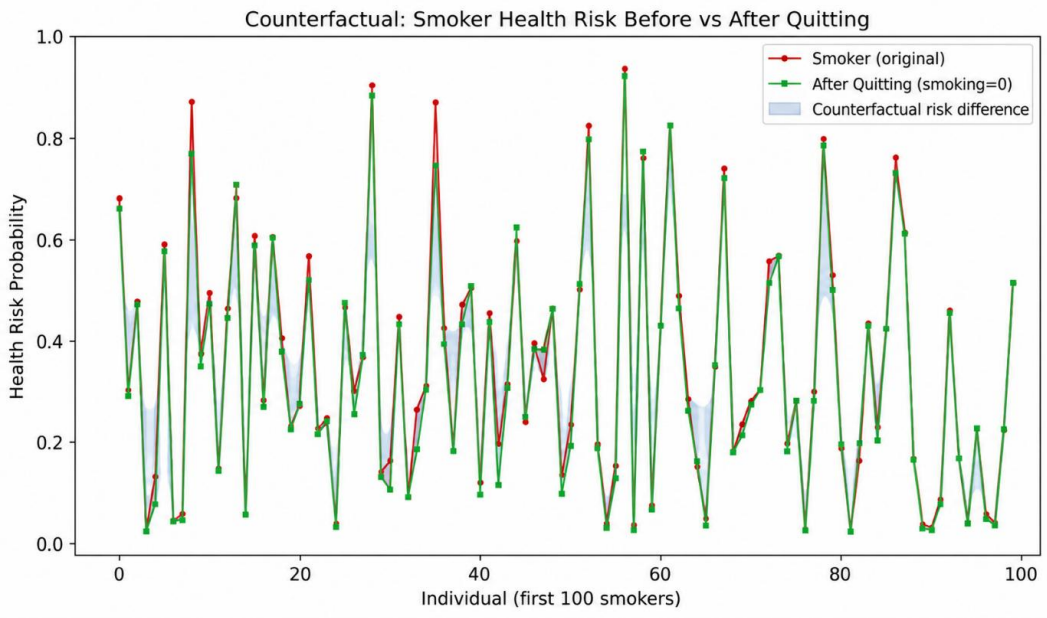

Figure 4. Counterfactual analysis of health-risk probability under observed and hypothetical non-smoking states.

The counterfactual results should be interpreted strictly as model-based hypothetical predictions. Because the underlying data are cross-sectional and the counterfactual modification changes only the smoking-status variable, the analysis does not model the physiological changes that may occur over time following smoking cessation. Consequently, the observed differences cannot establish a causal effect of smoking cessation or quantify an individual's future clinical benefit. Instead, they indicate the sensitivity of the trained model to an alternative smoking-status input while the individual's other observed characteristics remain unchanged.

SHAP and counterfactual analysis therefore provide complementary forms of explainability. SHAP addresses a feature-level "why" question by characterizing the relative contribution of observed predictors to model output, whereas counterfactual analysis addresses a scenario-level "what-if" question by recalculating the predicted probability after a specified modification of smoking status. Their combination allows the framework to relate the broader structure of model predictions to the individual-level response to a predefined smoking-status scenario.

4.6. Individual-Level Counterfactual Risk Changes

Table 6 summarizes the distribution of individual-level counterfactual risk differences for the 4,103 individuals identified as smokers in the independent test set. The counterfactual difference is defined as the predicted probability under the observed smoking state minus the predicted probability under the hypothetical non-smoking state. A positive Δp therefore indicates a lower model-estimated probability under the hypothetical non-smoking condition.

The mean counterfactual difference was 0.074, with a median of 0.070 and an interquartile range of 0.038–0.107. Thus, for the central 50% of smokers, the hypothetical change in smoking status was associated with a decrease of approximately 3.8–10.7 percentage points in model-estimated health-risk probability. Overall, 3,939 individuals (96.0%) had a positive Δp, whereas 164 individuals (4.0%) had a negative Δp.

The range of Δp from −0.080 to 0.279 indicates substantial individual-level heterogeneity in the model response. Although the majority of individuals showed lower predicted risk under the hypothetical non-smoking state, the presence of negative differences demonstrates that this response was not uniform across the test population. This pattern is consistent with the broader feature-level interpretation in Section 4.4, suggesting that the modeled sensitivity to smoking status depends on interactions with an individual's other observed characteristics.

Table 6. Distribution of Individual-Level Counterfactual Changes in Model-Estimated Health-Risk Probability

Statistic	Δp
Number of smokers	4,103
Mean	0.074
Median	0.070
Standard deviation	0.050
Minimum	−0.080
25th percentile	0.038
75th percentile	0.107
Maximum	0.279
Positive Δp, n (%)	3,939 (96.0%)

Δp = predicted probability under the observed smoking state − predicted probability under the hypothetical non-smoking state.

Importantly, Δp represents a change in model-estimated probability rather than a percentage reduction in actual disease risk. The observed differences should therefore not be interpreted as clinically expected benefits of smoking cessation or as estimates of causal treatment effects. Rather, they quantify how the trained XGBoost model responds when the smoking-status input is changed while the remaining observed characteristics are held constant.

4.7. Implications

This study presents an explainable machine learning framework for assessing smoking-associated health risk using a model-defined physiological-risk target and counterfactual analysis. The framework extends conventional smoking-status prediction by examining how the model-estimated health-risk probability changes under a hypothetical modification from smoking to non-smoking. The integration of XGBoost, SHAP, and counterfactual analysis provides complementary information on both the factors contributing to model predictions and the individual-level sensitivity of those predictions to a specified smoking-status scenario.

The results indicate that the estimated risk is influenced by a broader combination of demographic, anthropometric, and health-related characteristics, while smoking status provides additional information within the modeled risk framework. The counterfactual analysis further demonstrates heterogeneity in the model response across individuals, highlighting the value of examining scenario-specific predictions alongside global feature attribution.

Nevertheless, the framework remains exploratory and is based on cross-sectional data and a model-defined health-risk target rather than longitudinal clinical outcomes. The counterfactual estimates should therefore be interpreted as model-based hypothetical scenarios

rather than causal effects or predictions of future health improvement. Further validation using longitudinal data and clinically defined outcomes would be necessary to assess whether this approach can support practical health-risk assessment or smoking-related preventive research.

5. Conclusions

This study developed a Counterfactual Smoking Risk Assessment Framework that integrates XGBoost-based health-risk assessment, SHAP-based interpretation, and individual-level counterfactual analysis. Rather than predicting smoking status itself, the framework constructs a model-defined Health Risk Status from multiple physiological indicators and examines how the estimated probability of high health risk changes when an individual's observed smoking status is hypothetically changed to a non-smoking state while other observed characteristics are held constant. The comparative analysis showed that model performance varied across classifiers, while the sensitivity and ablation analyses demonstrated that the resulting predictive task was influenced by the definition of the derived health-risk target. SHAP analysis further identified the relative contribution of demographic, anthropometric, and health-related variables to the model predictions, while the counterfactual analysis quantified the model's individual-level sensitivity to the specified smoking-status modification.

The findings provide two complementary insights. First, smoking status contributes information to the modeled health-risk assessment but does not determine the prediction independently of the individual's broader observed characteristics. Second, the counterfactual analysis revealed heterogeneous changes in model-estimated risk across individuals, indicating that the modeled response to a smoking-status modification depends on the overall predictor profile. These findings support the use of explainable and scenario-based analysis as a complement to conventional predictive modeling. However, the resulting probability differences describe the behavior of the trained model under a controlled input modification and should not be interpreted as causal effects of smoking cessation or as estimates of future clinical outcomes.

Several limitations constrain the interpretation of these findings. First, the cross-sectional nature of the dataset prevents the assessment of temporal relationships and causal effects between smoking status and health outcomes. Second, the publicly available dataset provides limited information on the original data-collection procedures, participant recruitment, ethics approval, and data-governance framework, which limits assessment of the broader provenance and representativeness of the data. Third, smoking status is represented only as a binary variable, without detailed exposure information such as cigarettes consumed per day, smoking duration, pack-years, or time since cessation. Consequently, the counterfactual analysis represents a discrete smoking-to-non-smoking scenario and cannot capture graded changes in smoking exposure or physiological changes occurring over time following cessation. Finally, Health Risk Status is a derived modeling target based on predefined physiological thresholds rather than a clinically diagnosed disease endpoint. The predictive and counterfactual findings should therefore be interpreted within the scope of this study-specific target rather than generalized directly to clinical disease risk.

Future research should address these limitations through independent and prospective longitudinal cohorts with detailed smoking histories, quantitative exposure measures, time since cessation, repeated physiological measurements, and clinically validated outcomes. Such data would enable the evaluation of dose- and time-dependent counterfactual scenarios and provide a stronger basis for assessing whether model-estimated changes correspond to subsequent health trajectories. External validation across populations and healthcare settings would also be necessary to examine the generalizability of the framework. In addition, incorporating longitudinal measurements could extend the current cross-sectional framework toward individualized risk trajectories, allowing future studies to distinguish changes in model-estimated risk associated with smoking-status scenarios from broader temporal changes in physiological health.

Funding: This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

Data Availability Statement: All data used shall be made available upon request.

Conflicts of Interest: The author declares no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

[1] D. B. Olawade and C. A. Aienobe-Asekharen, "Artificial intelligence in tobacco control: A systematic scoping review of applications, challenges, and ethical implications," *Int. J. Med. Inform.*, vol. 202, p. 105987, Oct. 2025, doi: 10.1016/j.ijmedinf.2025.105987.

[2] M. A. Jonas, J. A. Oates, J. K. Ockene, and C. H. Hennekens, "Statement on smoking and cardiovascular disease for health care professionals. American Heart Association.," *Circulation*, vol. 86, no. 5, pp. 1664–1669, Nov. 1992, doi: 10.1161/01.CIR.86.5.1664.

[3] S. Gupta, V. Kumar, and P. Gupta, "A comprehensive study on the harmful effects of the smoking on human beings," in *Challenges in Information, Communication and Computing Technology*, London: CRC Press, 2024, pp. 577–582. doi: 10.1201/9781003559085-99.

[4] V. Chakma, M. J. H. Nerab, A. Rouf, A. Sayed, H. M. Saim, and M. N. Khan, "Machine Learning Techniques for Predicting SRHD: Smoking-Related Health Decline." Apr. 16, 2025. doi: 10.22541/au.174483083.36026266/v1.

[5] A. B. Alarabi, F. Z. Alshbool, and F. T. Khasawneh, "Impact of smoking on the complement system: a narrative review," *Front. Immunol.*, vol. 16, Jun. 2025, doi: 10.3389/fimmu.2025.1619835.

[6] S. Annareddy, B. Ghewade, U. Jadhav, P. Wagh, and S. Sarkar, "Unveiling the Long-Term Lung Consequences of Smoking and Tobacco Consumption: A Narrative Review," *Cureus*, Aug. 2024, doi: 10.7759/cureus.66415.

[7] E. D. Gometz, "Health Effects of Smoking and the Benefits of Quitting," *AMA J. Ethics*, vol. 13, no. 1, pp. 31–35, Jan. 2011, doi: 10.1001/virtualmentor.2011.13.1.cprl1-1101.

[8] P. Jha *et al.*, "21st-Century Hazards of Smoking and Benefits of Cessation in the United States," *N. Engl. J. Med.*, vol. 368, no. 4, pp. 341–350, Jan. 2013, doi: 10.1056/NEJMsa1211128.

[9] R. Doll, R. Peto, J. Boreham, and I. Sutherland, "Mortality in relation to smoking: 50 years' observations on male British doctors," *BMJ*, vol. 328, no. 7455, p. 1519, Jun. 2004, doi: 10.1136/bmj.38142.554479.AE.

[10] U.S. Department of Health and Human Services, "The Health Consequences of Smoking: 50 Years of Progress: A Report of the Surgeon General," U.S. Department of Health and Human Services, Centers for Disease Control and Prevention, National Center for Chronic Disease Prevention and Health Promotion, Office on Smoking and Health, Atlanta, GA, 2014. [Online]. Available: https://www.ncbi.nlm.nih.gov/books/NBK179276/

[11] E. E. Calle *et al.*, "The American Cancer Society Cancer Prevention Study II Nutrition Cohort," *Cancer*, vol. 94, no. 9, pp. 2490–2501, May 2002, doi: 10.1002/cncr.101970.

[12] T. T. T. Le, D. Mendez, and K. E. Warner, "The Benefits of Quitting Smoking at Different Ages," *Am. J. Prev. Med.*, vol. 67, no. 5, pp. 684–688, Nov. 2024, doi: 10.1016/j.amepre.2024.06.020.

[13] C. M. Chang, S. H. Edwards, A. Arab, A. Y. Del Valle-Pinero, L. Yang, and D. K. Hatsukami, "Biomarkers of Tobacco Exposure: Summary of an FDA-Sponsored Public Workshop," *Cancer Epidemiol. Biomarkers Prev.*, vol. 26, no. 3, pp. 291–302, Mar. 2017, doi: 10.1158/1055-9965.EPI-16-0675.

[14] K. Frost-Pineda *et al.*, "Biomarkers of Potential Harm Among Adult Smokers and Nonsmokers in the Total Exposure Study," *Nicotine Tob. Res.*, vol. 13, no. 3, pp. 182–193, Mar. 2011, doi: 10.1093/ntr/ntq235.

[15] J. N. Khouja, M. R. Munafò, C. L. Relton, A. E. Taylor, S. H. Gage, and R. C. Richmond, "Investigating the added value of biomarkers compared with self-reported smoking in predicting future e-cigarette use: Evidence from a longitudinal UK cohort study," *PLoS One*, vol. 15, no. 7, p. e0235629, Jul. 2020, doi: 10.1371/journal.pone.0235629.

[16] K. Sinha, N. Ghosh, and P. C. Sil, "Harnessing machine learning in contemporary tobacco research," *Toxicol. Reports*, vol. 14, p. 101877, Jun. 2025, doi: 10.1016/j.toxrep.2024.101877.

[17] S. Xiao *et al.*, "Proteomic signatures of smoking and their associations with risk of incident diseases and mortality in diverse populations," *Nat. Commun.*, vol. 17, no. 1, p. 928, Dec. 2025, doi: 10.1038/s41467-025-67656-x.

[18] K. Sugden *et al.*, "Establishing a generalized polyepigenetic biomarker for tobacco smoking," *Transl. Psychiatry*, vol. 9, no. 1, p. 92, Feb. 2019, doi: 10.1038/s41398-019-0430-9.

[19] D. Tang, Z. Guo, J. Chen, Y. Chen, and Y. Zhang, "Different dimensions of smoking behavior and their associations with accelerated composite biomarkers-based biological aging in Chinese older adults," *BMC Geriatr.*, vol. 26, no. 1, p. 14, Dec. 2025, doi: 10.1186/s12877-025-06828-2.

[20] R. Fu *et al.*, "Machine learning applications in tobacco research: a scoping review," *Tob. Control*, vol. 32, no. 1, pp. 99–109, Jan. 2023, doi: 10.1136/tobaccocontrol-2020-056438.

[21] H. Bendotti, S. Lawler, G. C. K. Chan, C. Gartner, D. Ireland, and H. M. Marshall, "Conversational artificial intelligence interventions to support smoking cessation: A systematic review and meta-analysis," *Digit. Heal.*, vol. 9, Jan. 2023, doi: 10.1177/20552076231211634.

[22] K. Davagdorj, V. H. Pham, N. Theera-Umpon, and K. H. Ryu, "XGBoost-Based Framework for Smoking-Induced Noncommunicable Disease Prediction," *Int. J. Environ. Res. Public Health*, vol. 17, no. 18, p. 6513, Sep. 2020, doi: 10.3390/ijerph17186513.

[23] T. Chen and C. Guestrin, "XGBoost: A Scalable Tree Boosting System," in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, Aug. 2016, pp. 785–794. doi: 10.1145/2939672.2939785.

[24] A. Mohamed, M. Abdelrehim, and R. Al-Barazie, "Context matters in machine learning based disease prediction with insights from diverse clinical and symptom data," *Sci. Rep.*, vol. 15, no. 1, p. 42669, Nov. 2025, doi: 10.1038/s41598-025-26855-8.

[25] A. K. Leist *et al.*, "Mapping of machine learning approaches for description, prediction, and causal inference in the social and health sciences," *Sci. Adv.*, vol. 8, no. 42, Oct. 2022, doi: 10.1126/sciadv.abk1942.

[26] Y. Takefuji, "Beyond XGBoost and SHAP: Unveiling true feature importance," *J. Hazard. Mater.*, vol. 488, p. 137382, May 2025, doi: 10.1016/j.jhazmat.2025.137382.
[27] A. Seong, S.-Y. Kang, Y.-G. Song, and G.-W. Kim, "Health-AutoML: An Automatic Adaptive Multi-Layer Stacking Ensemble Learning Framework for Analyzing Healthcare Data," *J. Korean Inst. Inf. Technol.*, vol. 22, no. 1, pp. 23–37, Jan. 2024, doi: 10.14801/jkiit.2024.22.1.23.
[28] S. Aishwarya, P. C. Siddalingaswamy, and K. Chadaga, "Explainable artificial intelligence driven insights into smoking prediction using machine learning and clinical parameters," *Sci. Rep.*, vol. 15, no. 1, p. 24069, Jul. 2025, doi: 10.1038/s41598-025-09409-w.
[29] Somashekar. V, "Multi-Organ Impact of Smoking Using Explainable Artificial Intelligence: Prediction, Validation, and Organ Impact Pattern," *Cureus J. Comput. Sci.*, Jun. 2026, doi: 10.7759/s44389-026-00121-y.
[30] A. Aravindkumar, M. Ramadoss, S. A. Fakhruddin Ahmed, V. Sampath, and K. Lakshminarayanan, "Explainable AI in healthcare: a systematic review of XAI use cases in imaging, diagnostics, and rehabilitation," *Front. Artif. Intell.*, vol. 9, Apr. 2026, doi: 10.3389/frai.2026.1749527.
[31] F. Di Martino and F. Delmastro, "Explainable AI for clinical and remote health applications: a survey on tabular and time series data," *Artif. Intell. Rev.*, vol. 56, no. 6, pp. 5261–5315, Jun. 2023, doi: 10.1007/s10462-022-10304-3.
[32] S. M. Lundberg and S.-I. Lee, "A Unified Approach to Interpreting Model Predictions," in *NIPS'17: Proceedings of the 31st International Conference on Neural Information Processing Systems*, Nov. 2017, pp. 4768–4777. [Online]. Available: https://dl.acm.org/doi/10.5555/3295222.3295230
[33] S. M. Lundberg *et al.*, "From local explanations to global understanding with explainable AI for trees," *Nat. Mach. Intell.*, vol. 2, no. 1, pp. 56–67, Jan. 2020, doi: 10.1038/s42256-019-0138-9.
[34] X. Wang *et al.*, "An explainable artificial intelligence framework for risk prediction of COPD in smokers," *BMC Public Health*, vol. 23, no. 1, p. 2164, Nov. 2023, doi: 10.1186/s12889-023-17011-w.
[35] Y. Nohara, K. Matsumoto, H. Soejima, and N. Nakashima, "Explanation of machine learning models using shapley additive explanation and application for real data in hospital," *Comput. Methods Programs Biomed.*, vol. 214, p. 106584, Feb. 2022, doi: 10.1016/j.cmpb.2021.106584.
[36] A. Prashanthan and J. Prashanthan, "Optimized Explainable Predictive Models for Risk-Based Prioritization in Type 2 Diabetes Prevention among Women with Prior Gestational Diabetes," *J. Futur. Artif. Intell. Technol.*, vol. 2, no. 4, pp. 734–759, Feb. 2026, doi: 10.62411/faith.3048-3719-321.
[37] A. Nayebi, S. Tipirneni, B. Foreman, C. K. Reddy, and V. Subbian, "An Empirical Comparison of Explainable Artificial Intelligence Methods for Clinical Data: A Case Study on Traumatic Brain Injury.," *AMIA ... Annu. Symp. proceedings. AMIA Symp.*, vol. 2022, pp. 815–824, 2022, [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/37128424
[38] M. Prosperi *et al.*, "Causal inference and counterfactual prediction in machine learning for actionable healthcare," *Nat. Mach. Intell.*, vol. 2, no. 7, pp. 369–375, Jul. 2020, doi: 10.1038/s42256-020-0197-y.
[39] A. M. Salih *et al.*, "A Perspective on Explainable Artificial Intelligence Methods: SHAP and LIME," *Adv. Intell. Syst.*, vol. 7, no. 1, Jan. 2025, doi: 10.1002/aisy.202400304.
[40] P. Knab, S. Marton, U. Schlegel, and C. Bartelt, "Which LIME Should I Trust? Concepts, Challenges, and Solutions," in *Communications in Computer and Information Science*, 2026, pp. 28–52. doi: 10.1007/978-3-032-08324-1_2.
[41] G. T. Ayem, A. J. Shiwua, P. T. Atagher, and T. J. Tivkaa, "DiFACE2: Generating Diverse, Feasible, and Actionable Counterfactual Explanations for Post-Conflict Primary Education Interventions Using Causal DAGs," *J. Futur. Artif. Intell. Technol.*, vol. 2, no. 3, pp. 504–520, Nov. 2025, doi: 10.62411/faith.3048-3719-283.
[42] Kukuroo3, "Body Signal of Smoking," *Kaggle*, 2020. https://www.kaggle.com/datasets/kukuroo3/body-signal-of-smoking
[43] C. C. Odiakaose *et al.*, "Hypertension Detection via Tree-Based Stack Ensemble with SMOTE-Tomek Data Balance and XGBoost Meta-Learner," *J. Futur. Artif. Intell. Technol.*, vol. 1, no. 3, pp. 269–283, Dec. 2024, doi: 10.62411/faith.3048-3719-43.
[44] D. R. I. M. Setiadi, K. Nugroho, A. R. Muslikh, S. W. Iriananda, and A. A. Ojugo, "Integrating SMOTE-Tomek and Fusion Learning with XGBoost Meta-Learner for Robust Diabetes Recognition," *J. Futur. Artif. Intell. Technol.*, vol. 1, no. 1, pp. 23–38, May 2024, doi: 10.62411/faith.2024-11.
[45] M. Al-Duais *et al.*, "Comparative Analysis of Machine Learning and Deep learning Techniques for Early Prediction of Breast Cancer," *J. Futur. Artif. Intell. Technol.*, vol. 2, no. 2, pp. 242–254, Jun. 2025, doi: 10.62411/faith.3048-3719-68.