
MACHINE LEARNING MODELS FOR PREDICTING SMOKING-RELATED HEALTH DECLINE AND DISEASE RISK *

Vaskar Chakma *, MD Jaheid Hasan Nerab †, Abdur Rouf †

School of Artificial Intelligence and Computer Science
Nantong University
Jiangsu, China

{vaskarchakma, abdurrouf, nerab}@stmail.ntu.edu.cn

Abu Sayed

School of Transportation and Civil Engineering
Nantong University
Jiangsu, China

sayedddber.2017@gmail.com

Hossem MD Saim, Md. Nournabi Khan

School of Mechanical Engineering
Nantong University
Jiangsu, China

{saimhossen289, nournabikhan1}@gmail.com

ABSTRACT

Smoking continues to be a major preventable cause of death worldwide, affecting millions through damage to the heart, metabolism, liver, and kidneys. However, current medical screening methods often miss the early warning signs of smoking-related health problems, leading to late-stage diagnoses when treatment options become limited. This study presents a systematic comparative evaluation of machine learning approaches for smoking-related health risk assessment, emphasizing clinical interpretability and practical deployment over algorithmic innovation. We analyzed health screening data from 55,691 individuals, examining various health indicators, including body measurements, blood tests, and demographic information. We tested three advanced prediction algorithms - Random Forest, XGBoost, and LightGBM - to determine which could most accurately identify people at high risk. This study employed a cross-sectional design to classify current smoking status based on health screening biomarkers, not to predict future disease development. Our Random Forest model performed best, achieving an Area Under the Curve (AUC) of 0.926, meaning it could reliably distinguish between high-risk and lower-risk individuals. Using SHAP (SHapley Additive exPlanations) analysis to understand what the model was detecting, we found that key health markers played crucial roles in prediction: blood pressure levels, triglyceride concentrations, liver enzyme readings, and kidney function indicators (serum creatinine) were the strongest signals of declining health in smokers. These results demonstrate that artificial intelligence can serve as a powerful tool for early disease detection in smokers. By identifying at-risk individuals before conventional symptoms appear, healthcare providers could intervene earlier with personalized prevention strategies. Implementing these predictive systems in public health programs could reduce the enormous burden smoking places on healthcare systems while shifting medical care from reactive treatment to proactive prevention.

Keywords Smoking-Related Diseases · Machine Learning Prediction Models · Health Risk Assessment · Predictive Analytics in Healthcare · Early Disease Detection · Public Health Informatics · Artificial Intelligence in Medicine

*These are corresponding authors.

† These authors contributed equally.

1 Introduction

Smoking remains one of the most pressing global public health challenges, representing a complex interplay of addiction, behavioral patterns, and progressive biological damage[1]. Each year, tobacco use is responsible for over 8 million deaths worldwide, with the World Health Organization estimating that nearly half of all smokers will ultimately succumb to smoking-related illnesses [2, 3]. While lung cancer and chronic obstructive pulmonary disease (COPD) are the most widely recognized consequences, smoking also drives cardiovascular disease, metabolic dysfunction, and systemic inflammation that can compromise virtually every organ system [4, 5]. Perhaps most concerning is that this damage frequently progresses insidiously over years, often becoming irreversible before clinical symptoms manifest[6]. Despite decades of comprehensive public health initiatives and overwhelming scientific evidence, approximately 1.3 billion people worldwide continue to use tobacco products[7]. Many smokers harbor what might be termed "optimistic bias," believing they can quit before substantial harm occurs or that they will somehow avoid the worst outcomes. The non-linear trajectory of smoking-related decline—characterized by years of subclinical damage that suddenly manifests as severe disease—highlights critical limitations in reactive diagnostic approaches that await obvious symptoms before intervention. These delays result in lost opportunities for prevention and early therapeutic intervention. Our research aims to transform this reactive paradigm by developing advanced predictive tools that can detect risk at substantially earlier stages[8, 9]. We hypothesize that smoking leaves distinct, systemic biological signatures across cardiovascular, metabolic, hepatic, and oral health pathways that machine learning algorithms can identify long before conventional clinical thresholds are exceeded[10, 2]. By simultaneously analyzing these diverse biomarkers, we aim to construct a more comprehensive and clinically relevant assessment of smoking-related health decline. This holistic approach addresses significant gaps in prior research, which has often concentrated on single disease endpoints or limited feature sets, thereby constraining real-world applicability[11, 12, 13].

A critical innovation in our study is the direct comparison of machine learning models with established clinical risk assessment tools, including the Framingham cardiovascular risk score. This benchmarking exercise tests whether advanced algorithms provide measurable advantages over standard, widely validated approaches—a crucial step for building confidence among clinicians and policymakers who will ultimately implement these systems in practice. A fundamental principle guiding our work is model interpretability. We employ SHAP (SHapley Additive exPlanations) values to elucidate how each variable contributes to individual risk predictions [14, 15]. This transparency is essential for fostering clinician trust and facilitating shared decision-making, positioning these tools as decision support rather than replacements for professional judgment. We also address practical considerations for clinical implementation, including integration into existing healthcare workflows, appropriate clinical responses to risk alerts, and responsible management of false positive and false negative predictions to minimize potential harm. Understanding these operational aspects is critical for successful translation from research to practice.

Our study places particular emphasis on algorithmic fairness by thoroughly characterizing the geographic, ethnic, and socioeconomic distribution of our study population. We explicitly analyze how data quality issues—such as extreme laboratory value outliers—might influence model performance and generalizability. This attention to equity ensures that our models are not only technically sound but also ethically responsible and applicable across diverse populations. Through the integration of advanced algorithms, rigorous comparison with traditional assessment tools, realistic evaluation of clinical adoption pathways, and strong emphasis on equity, we aim to advance predictive medicine for smoking-related disease beyond academic exercises toward genuinely impactful, patient-centered applications. By identifying at-risk individuals before irreversible damage occurs, these tools could enable more timely interventions, facilitate targeted prevention strategies, and ultimately improve public health outcomes for millions of people affected by tobacco use. This investigation employs a cross-sectional analytical framework wherein all predictor variables (demographic characteristics, anthropometric measurements, and biochemical biomarkers) and the outcome variable (current smoking status) were collected simultaneously during a single health screening visit. The prediction task is therefore *classification*—identifying individuals who are current smokers based on their present physiological state, rather than *prognosis*, which would entail predicting future disease onset or health decline over time. No longitudinal follow-up data were available; thus, temporal causality cannot be inferred from our results. The clinical utility of this approach lies in leveraging routinely collected health screening data to detect physiological signatures of smoking exposure that may indicate early-stage damage before overt clinical symptoms manifest, thereby enabling timely intervention and smoking cessation support.

Our contribution lies not in novel algorithm development, but in rigorous comparative evaluation of established machine learning methods applied to comprehensive multi-system health screening data, with particular emphasis on clinical interpretability through SHAP analysis, validation against traditional risk scores (Framingham), and practical considerations for real-world deployment. This systematic approach addresses critical gaps in prior smoking risk prediction research, which often focuses on single disease endpoints or lacks adequate attention to explainability—essential prerequisites for clinical adoption.

2 Related Study

The prediction and assessment of smoking-related health risks has garnered substantial research attention over the past decades, with increasing momentum following the integration of machine learning methodologies into public health applications. A considerable body of literature has examined predictive models for estimating smoking status or stratifying smokers based on routinely collected health variables. Early investigations predominantly employed traditional statistical approaches, particularly logistic regression, to establish associations between smoking behavior and cardiopulmonary conditions[16, 17]. While these conventional methods achieved acceptable accuracy for basic classification tasks, they demonstrated inherent limitations in capturing the complex, non-linear relationships that characterize smoking’s biological effects across multiple physiological systems.

The past decade has witnessed a paradigm shift toward more sophisticated algorithmic approaches for smoking risk assessment. Researchers have increasingly leveraged advanced machine learning techniques, including decision trees, support vector machines, and gradient boosting methods, to enhance smoking-risk stratification capabilities[18, 19]. These computational approaches have shown promising performance in predicting smoking status and specific disease outcomes, particularly for conditions such as lung cancer and chronic obstructive pulmonary disease[20, 21]. The improved predictive accuracy of these models stems from their ability to identify subtle patterns and interactions among multiple risk factors that may elude traditional statistical methods. Despite these technological advances, significant gaps persist in the existing literature. A critical limitation of many previous studies is their narrow focus on single disease endpoints or organ-specific outcomes. This reductionist approach fails to capture the systemic nature of smoking-induced damage, which simultaneously affects cardiovascular, metabolic, hepatic, renal, and other physiological systems. By concentrating on isolated conditions, prior research has provided an incomplete picture of overall health decline in smokers, potentially missing important early warning signs that manifest across multiple biomarker domains.

Furthermore, many earlier investigations relied on limited feature sets, often constrained to a handful of easily measurable clinical variables. This restricted scope may overlook important predictive signals present in comprehensive health screening data. Equally concerning is the prevalent use of "black-box" models without adequate attention to interpretability[22]. The lack of explainability in these models has created substantial barriers to clinical adoption, as healthcare providers understandably hesitate to base treatment decisions on opaque algorithmic recommendations whose reasoning cannot be scrutinized or validated against clinical knowledge. Another notable deficiency in the literature is the absence of rigorous benchmarking against established clinical risk assessment tools. Few studies have directly compared machine learning predictions with validated instruments such as the Framingham Risk Score or other standardized clinical algorithms[23]. This omission makes it difficult to evaluate whether the added complexity of machine learning approaches yields meaningful improvements over simpler, well-established methods that clinicians already trust and understand.

Our research addresses these critical gaps through several key innovations. First, we adopt a holistic, systems-based perspective by incorporating a comprehensive panel of biomarkers spanning cardiovascular, hepatic, renal, metabolic, and oral health domains. This multidimensional approach recognizes that smoking’s pathological effects manifest across multiple organ systems simultaneously, and that early detection requires monitoring these interconnected pathways rather than isolated endpoints.

Second, we prioritize model interpretability through the systematic application of SHAP (SHapley Additive exPlanations) values, transforming our ensemble machine learning models from opaque predictors into transparent[24], clinically comprehensible tools. This interpretability framework enables healthcare providers to understand not only *what* the model predicts but *why* it makes specific predictions for individual patients—a crucial requirement for building clinical trust and facilitating shared decision-making.

Third, we provide rigorous comparative analysis by benchmarking our machine learning models against established clinical risk scores. This head-to-head comparison offers concrete evidence regarding whether advanced algorithms deliver meaningful advantages over conventional assessment tools, addressing a question of paramount importance for clinical implementation and resource allocation decisions.

Finally, our work reframes the research question from simply identifying current smokers or predicting isolated disease outcomes toward constructing a multidimensional risk assessment framework that supports personalized prevention strategies and more efficient allocation of clinical resources. By detecting early signs of health decline before irreversible damage occurs, our approach aims to shift clinical practice from reactive disease management toward proactive health preservation. Through these contributions, we extend the scientific discourse beyond technical performance metrics toward the development of clinically actionable, interpretable, and ethically responsible tools that can meaningfully impact patient care and public health outcomes for smoking populations.

3 Methods

3.1 Study Design and Participants

This study employed a **retrospective cross-sectional design** using data from a comprehensive health screening program conducted in South Korea². All measurements-including demographic information, anthropometric parameters, biochemical analyses, and self-reported smoking status-were collected during a single health screening visit. **Temporal Design:** The simultaneous collection of predictor and outcome variables means this study addresses a classification problem (identifying current smokers) rather than a prognostic prediction problem (forecasting future disease). This design choice reflects the practical clinical scenario where healthcare providers must assess smoking-related health risks using only cross-sectional screening data available at the point of care. The screening program primarily enrolled participants from urban and suburban populations, reflecting the demographic composition typical of organized health surveillance initiatives in the region. While individual ethnic identifiers were not systematically recorded, the cohort is presumed to be predominantly Korean, consistent with the national demographic profile of the screening program's catchment area. Participants underwent standardized health assessments that included the collection of demographic information, anthropometric measurements, and biochemical analyses. **Primary Outcome Variable:** Smoking status, categorized as current smoker or non-smoker based on self-report at the time of screening, served as the primary outcome for our classification models. Individuals who reported currently smoking any tobacco products were classified as smokers (coded as 1), while those reporting no current tobacco use were classified as non-smokers (coded as 0). **Important Note:** We did not have access to smoking history variables (pack-years, duration, cessation attempts) or longitudinal health outcomes (subsequent disease diagnoses, mortality). Therefore, our models identify cross-sectional associations between biomarkers and current smoking status rather than predicting future smoking-related disease incidence. Socioeconomic variation within the cohort was indirectly represented through lifestyle indicators such as smoking prevalence and obesity rates, though direct measures of income, education level, or occupational status were not available. This represents an important limitation, as socioeconomic factors are known to influence both smoking behavior and health outcomes. The retrospective nature of the dataset and its sampling methodology may result in underrepresentation of certain populations, particularly individuals from rural areas or highly marginalized communities. These potential sampling biases and their implications for model generalizability are addressed in detail in the Discussion section.

3.2 Dataset Characteristics

Table 1 analytical dataset comprised 55,691 individual health screening records, each containing a comprehensive array of demographic, anthropometric, clinical, and lifestyle-related variables. The dataset structure was designed to capture multiple dimensions of health status relevant to smoking-related physiological changes. **Demographic variables** included age (years) and biological sex, providing essential contextual information for risk stratification. **Anthropometric measurements** encompassed height (cm), weight (kg), and waist circumference (cm)-key indicators of body composition and metabolic health status that are known to interact with smoking in determining cardiovascular and metabolic risk.

Clinical biomarkers spanned multiple physiological systems[25, 26]:

- *Cardiovascular markers:* systolic blood pressure (SBP) and diastolic blood pressure (DBP), measured in mmHg
- *Metabolic markers:* fasting blood glucose (mg/dL), total cholesterol (mg/dL), triglycerides (mg/dL), high-density lipoprotein cholesterol (HDL, mg/dL), and low-density lipoprotein cholesterol (LDL, mg/dL)
- *Hepatic function indicators:* aspartate aminotransferase (AST, IU/L), alanine aminotransferase (ALT, IU/L), and gamma-glutamyl transferase (GGT, IU/L)
- *Renal function markers:* serum creatinine (mg/dL) and urinary protein levels
- *Hematological parameters:* hemoglobin concentration (g/dL)

The primary outcome variable was **smoking status**, coded as a binary indicator (smoker vs. non-smoker). This classification was based on self-reported current smoking behavior at the time of health screening. Prior to statistical analysis and model development, we implemented rigorous data quality control procedures to ensure the integrity and reliability of the dataset. This multi-step process included outlier identification, biological plausibility assessment,

²Dataset source: "Smoking and Drinking Dataset with Body Signal." Kaggle. Available at: <https://www.kaggle.com/datasets/sooyoungher/smoking-drinking-dataset>

Table 1: Summary Statistics of Health Metrics for Study Participants

Variable	Unique	Mean	Std	Percentile Distribution				
				Min	25%	50%	75%	Max
Demographic and Anthropometric Metrics								
ID	55692	27845.5	16077.04	0	13922.75	27845.5	41768.25	55691
Age (years)	14	44.18	12.07	20	40	40	55	85
Height (cm)	13	164.65	9.19	130	160	165	170	190
Weight (kg)	22	65.86	12.82	30	55	65	75	135
Waist (cm)	566	82.05	9.27	51	76	82	88	129
Sensory Health Indicators								
Eyesight (left)	19	1.01	0.49	0.1	0.8	1	1.2	9.9
Eyesight (right)	17	1.01	0.49	0.1	0.8	1	1.2	9.9
Hearing (left)	2	1.03	0.16	1	1	1	1	2
Hearing (right)	2	1.03	0.16	1	1	1	1	2
Cardiovascular and Metabolic Indicators								
Systolic BP (mmHg)	130	121.49	13.68	71	112	120	130	240
Diastolic BP (mmHg)	95	76.00	9.68	40	70	76	82	146
Fasting Blood Sugar (mg/dL)	276	99.31	20.80	46	89	96	104	505
Cholesterol (mg/dL)	286	196.9	36.3	55	172	195	220	445
Triglyceride (mg/dL)	390	126.67	71.64	8	74	108	160	999
HDL (mg/dL)	126	57.29	14.74	4	47	55	66	618
LDL (mg/dL)	289	114.96	40.93	1	92	113	136	1860
Hematologic, Renal, and Hepatic Indicators								
Hemoglobin (g/dL)	145	14.62	1.56	4.9	13.6	14.8	15.8	21.1
Urine Protein	6	1.09	0.40	1	1	1	1	6
Serum Creatinine (mg/dL)	38	0.89	0.22	0.1	0.8	0.9	1	11.6
AST (U/L)	219	26.18	19.36	6	19	23	28	1311
ALT (U/L)	245	27.04	30.95	1	15	21	31	2914
GTP (U/L)	488	39.95	50.29	1	17	25	43	999
Oral Health and Behavioral Indicator								
Dental Caries	2	0.21	0.41	0	0	0	0	1
Smoking (binary)	2	0.37	0.48	0	0	0	1	1

BP: Blood Pressure; HDL: High-Density Lipoprotein; LDL: Low-Density Lipoprotein; AST: Aspartate Aminotransferase; ALT: Alanine Aminotransferase; GTP: Gamma-Glutamyl Transferase.

and unit consistency verification. Laboratory values were systematically screened for biological implausibility using established reference ranges from clinical literature. Results exceeding known physiological limits-such as LDL cholesterol values above 1,000 mg/dL or HDL cholesterol above 300 mg/dL-were flagged for detailed review. Each flagged value was manually examined in the context of the individual's complete clinical profile. Values consistent with documented rare pathological conditions (e.g., severe familial hypercholesterolemia) were retained in the dataset, while those appearing to represent data entry errors or instrument malfunction were excluded from analysis. All variables in Table 1 were measured at a single time point during health screening visits. This cross-sectional data structure means that predictor-outcome relationships reflect associations between current biomarker levels and concurrent smoking status, not temporal precedence. While elevated liver enzymes or blood pressure in smokers may result from chronic smoking exposure, the cross-sectional design precludes definitive causal inference. The dataset contained no follow-up measurements or longitudinal health outcomes (e.g., subsequent cardiovascular events, cancer diagnoses, or mortality), limiting our analysis to classification of current smoking status rather than prognostic risk modeling.

This approach balanced the need to preserve genuine extreme values while removing spurious data that could adversely affect model training. Missing values were addressed using imputation strategies selected based on the distribution

characteristics and missingness patterns of each variable. For continuous variables exhibiting approximately normal distributions, mean imputation was employed. For skewed continuous variables, median imputation was utilized to avoid distortion from extreme values. Categorical variables with missing entries were imputed using the mode (most frequent category). The proportion of missing data for each variable was documented, and sensitivity analyses were planned to assess the potential impact of imputation strategies on model performance. Continuous variables were standardized (z-score transformation) to ensure comparable scales across features with different units of measurement[27]. This preprocessing step is particularly important for distance-based algorithms and helps prevent features with larger numerical ranges from dominating the learning process. It is important to acknowledge significant gaps in the contextual information available within this dataset. Specifically, the data lacked comprehensive details regarding participants’ geographic origins beyond the broad urban/suburban classification, detailed ethnic or racial backgrounds, and socioeconomic indicators such as income, educational attainment, or occupational categories. This absence of contextual variables limits our ability to evaluate potential sampling biases systematically or to assess whether model performance varies across different demographic or socioeconomic strata.

3.3 Data Preprocessing

Prior to model development, we implemented a systematic data preprocessing pipeline to ensure data quality, consistency, and compatibility with machine learning algorithms. This multi-stage process addressed missing values, encoded categorical variables, and standardized numerical features to optimize model performance and reliability.

3.3.1 Missing Value Imputation

As is typical in real-world healthcare datasets, our data contained missing values across several variables that required careful handling. We adopted variable-specific imputation strategies based on the nature and distribution characteristics of each feature. For continuous numerical variables—including blood pressure measurements, lipid profiles, liver enzyme concentrations, and renal function markers—we employed median imputation. This approach replaces missing values with the median of the observed values for each respective feature. Median imputation was selected over mean imputation due to its robustness to outliers and extreme values, which are not uncommon in clinical laboratory data. This strategy preserves the central tendency of each feature’s distribution while minimizing distortion from atypical observations. For categorical variables, including biological sex, dental health status, and urinary protein categories, we utilized mode imputation, replacing missing entries with the most frequently occurring category for each variable. This method maintains the dominant patterns in categorical distributions while providing complete data for model training.

3.3.2 Categorical Variable Encoding

Machine learning algorithms require numerical input representations. Therefore, we transformed all categorical variables into numerical formats through appropriate encoding schemes. For binary categorical variables—such as smoking status (smoker vs. non-smoker), biological sex (male vs. female), and dental caries presence (present vs. absent)—we applied label encoding, converting categories into binary values of 0 and 1. This straightforward transformation preserves the dichotomous nature of these variables while rendering them computationally tractable for algorithmic processing. For ordinal categorical variables with inherent ordering (such as urinary protein levels), we maintained their ordinal relationships through ordered numerical encoding. This approach ensures that the encoded values reflect the natural progression or severity represented in the original categories.

3.3.3 Feature Standardization

A critical preprocessing step involved the standardization of all continuous numerical features using the StandardScaler transformation[28]. Clinical biomarkers naturally exist on vastly disparate measurement scales: systolic blood pressure values typically range from 70 to 240 mmHg, while hemoglobin concentrations span approximately 4 to 21 g/dL, and serum creatinine measurements range from 0.1 to 11.6 mg/dL. Without standardization, algorithms might inappropriately weight features with larger numerical ranges as more influential, regardless of their actual predictive importance. The StandardScaler transformation normalizes each feature to have a mean of zero and a standard deviation of one through the following formula:

$$z = \frac{x - \mu}{\sigma} \quad (1)$$

where x represents the original feature value, μ is the feature mean, σ is the feature standard deviation, and z is the standardized value[29]. This transformation ensures that all features contribute comparably to model training, preventing scale-dependent bias. Standardization is particularly crucial for distance-based algorithms (such as support

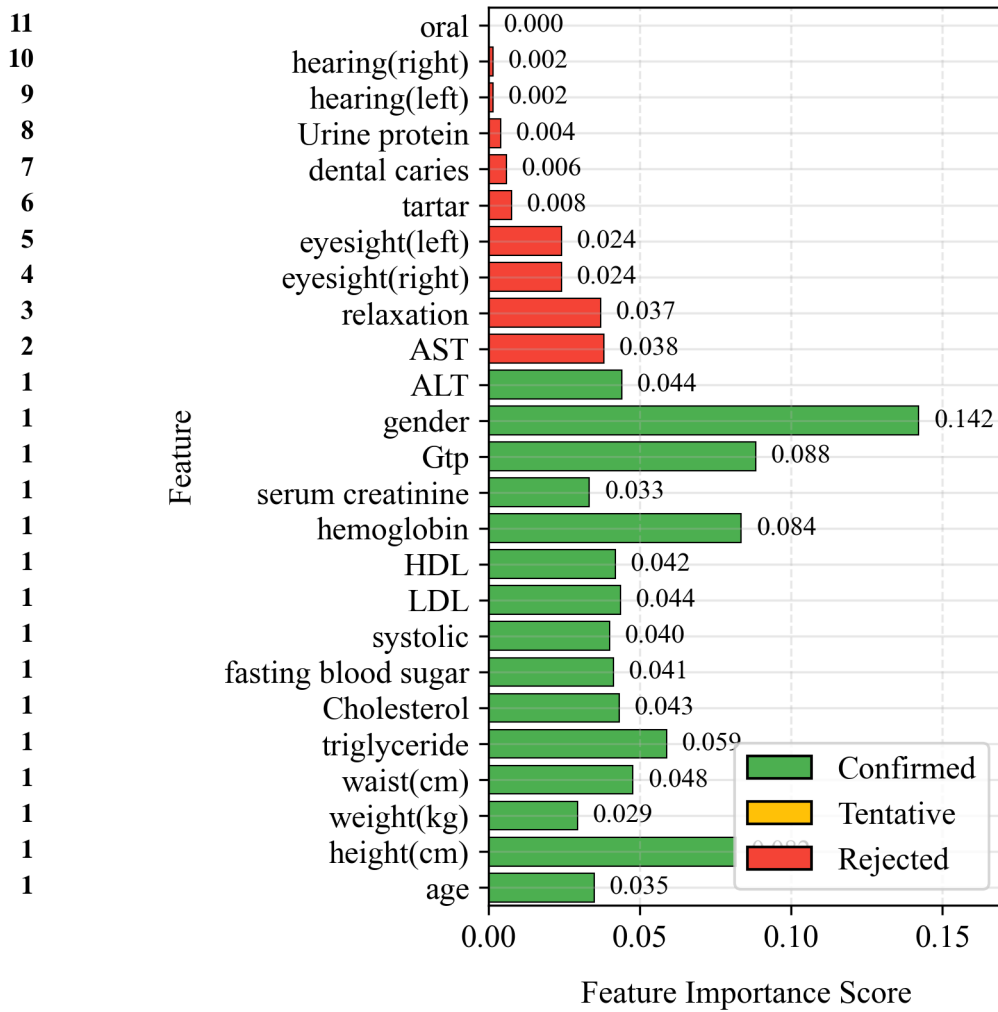


Figure 1: Disentangling the Interdependent Relationships Among Health Indicators.

vector machines) and regularized models (such as logistic regression with L1 or L2 penalties), which are inherently sensitive to feature magnitudes.

3.3.4 Data Quality Verification

Following each preprocessing step, we conducted comprehensive quality verification procedures. We examined feature distributions before and after transformation to confirm that preprocessing maintained the underlying data structure and relationships. Distribution plots, summary statistics, and correlation matrices were reviewed to identify any unintended artifacts introduced by the preprocessing pipeline. Additionally, we verified that the standardization process did not eliminate important distributional characteristics or create artificial patterns. The preservation of relative relationships between observations across all features was confirmed through dimensionality reduction visualization techniques applied to both raw and preprocessed data.

3.4 Feature Selection

To ensure our predictive models focused on the most clinically relevant biomarkers while avoiding redundant or uninformative features, we implemented a systematic two-stage feature selection process.

First, we applied the Boruta algorithm (Algorithm 1) [30], an advanced wrapper method built around Random Forest classification[31, 32]. This approach works by systematically comparing the importance of original features against

randomized shadow features—shuffled copies that serve as benchmarks for statistical noise. The algorithm retains only those features that demonstrate significantly stronger predictive power than these random counterparts. Through multiple iterations, Boruta progressively eliminates weak predictors while preserving features that consistently contribute to accurate smoking status classification. We complemented this automated selection with detailed correlation analysis to identify and address potential multicollinearity issues[33]. Clinical measurements often move together—for example, AST and ALT levels both reflect liver function and tend to change in tandem. We examined pairwise correlations between all features and applied clinical domain knowledge to decide whether to retain both correlated biomarkers or select the most clinically informative one. This step proved particularly valuable for metabolic markers such as triglycerides and HDL cholesterol, as well as anthropometric measurements like weight and waist circumference, where natural biological relationships could create redundant information that might distort model interpretations. The final feature set represented a careful balance between statistical performance and clinical practicality. We prioritized features that not only ranked high in machine learning importance metrics but also aligned with established medical knowledge about disease biomarkers. For instance, while certain laboratory values showed moderate predictive power in isolation, we gave preference to combinations of markers that clinicians actually use in routine diagnostic workflows (Figure 1). This dual emphasis on algorithmic performance and real-world clinical relevance resulted in a curated set of predictors that were both statistically powerful and medically interpretable—essential qualities for any healthcare application where model decisions need to be explainable to medical professionals.

Algorithm 1 Boruta Feature Selection Algorithm

Require: Dataset D with features $F = \{f_1, f_2, \dots, f_n\}$, target variable Y
Ensure: Subset of relevant features $F_{selected}$

- 1: Initialize all features in F as *tentative*
- 2: **while** stopping criterion not met **do**
- 3: Create **shadow features** S by permuting values of each $f_i \in F$
- 4: Train a **Random Forest** classifier on $(F \cup S)$ to compute feature importance scores
- 5: Let I_{max}^{shadow} be the maximum importance score among shadow features
- 6: **for** each feature $f_i \in F$ **do**
- 7: **if** $I(f_i) > I_{max}^{shadow}$ with statistical significance **then**
- 8: Mark f_i as **Confirmed Important**
- 9: **else if** $I(f_i) < I_{max}^{shadow}$ with statistical significance **then**
- 10: Mark f_i as **Rejected**
- 11: **else**
- 12: Keep f_i as **Tentative**
- 13: **end if**
- 14: **end for**
- 15: Remove rejected features and regenerate shadows for next iteration
- 16: **end while**
- 17: **return** $F_{selected}$ = set of all **Confirmed Important** features

3.5 Class Imbalance

Analysis of our dataset revealed a moderate class imbalance with smokers comprising 36.7% ($n = 20,438$) and non-smokers 63.3% ($n = 35,253$) of the cohort, yielding an imbalance ratio of 1.72:1. While not severe, this imbalance required careful handling to prevent models from developing bias toward the majority (non-smoker) class, which could result in high overall accuracy while failing to identify at-risk smokers—the population of primary clinical interest. Class imbalance presented a significant challenge in our machine learning pipeline. Our initial analysis revealed a substantial disparity between smokers and non-smokers in the dataset, with non-smokers considerably outnumbering smokers. This imbalance posed a real risk to model performance because standard machine learning algorithms tend to favor the majority class, potentially achieving high overall accuracy while failing to properly identify smokers—the minority class that represents our primary interest for health risk prediction. To ensure our models could effectively learn from all available data without developing this problematic bias, we implemented several strategic approaches: **Random Resampling Techniques:** For our baseline models (Logistic Regression and Support Vector Machines), we employed fundamental resampling methods[34]. Random oversampling of the minority class created additional copies of existing smoking cases to balance the class distribution. Conversely, random undersampling of the majority class achieved balance by reducing the number of non-smoking cases. While these methods improved our models’ ability to detect smokers, we carefully monitored for potential overfitting from oversampling and information loss from undersampling. **Class Weight Adjustment:** For our ensemble tree-based methods (Random Forest, XGBoost, and LightGBM), we leveraged their built-in capability to handle imbalance through class weighting[35]. By assigning higher

misclassification penalties to the minority smoking class, these algorithms naturally prioritized correct identification of smokers during training. Specifically, in Random Forest we adjusted class weights inversely proportional to class frequencies, while for XGBoost and LightGBM we utilized the `scale_pos_weight` parameter to account for the imbalance ratio. **Performance Metric Selection:** Recognizing that standard accuracy would be misleading with imbalanced data, we prioritized evaluation metrics that properly assess minority class identification: F1-score (the harmonic mean of precision and recall), AUC-ROC (area under the receiver operating characteristic curve), and the G-mean (geometric mean of sensitivity and specificity). **Stratified Sampling:** Throughout our cross-validation procedures, we maintained the original class distribution in each fold through stratified sampling[36]. This prevented accidental introduction of bias during model evaluation and ensured reliable performance estimates across all experimental runs.

3.5.1 Validation of Class Imbalance Mitigation Strategies

To verify that our class imbalance handling techniques were effective rather than merely theoretical, we conducted comparative analyses evaluating model performance before and after applying mitigation strategies. **Impact of Class Weighting:** Table 3 demonstrates the effect of class weight adjustments on ensemble tree models. Without class weighting, models exhibited high overall accuracy (>85%) but poor minority class detection (sensitivity 64%), indicating bias toward predicting the majority non-smoker class. After applying inverse frequency weighting, sensitivity improved substantially to 80.1% for Random Forest, while specificity declined only modestly from 89% to 86.5%. This trade-off represents desirable behavior for a health screening tool, where failing to identify at-risk smokers (false negatives) carries greater clinical cost than false alarms (false positives). **Evaluation with Imbalance-Specific Metrics:** The G-mean metric (geometric mean of sensitivity and specificity), specifically designed to assess balanced performance on imbalanced datasets, increased from 0.75 (unweighted) to 0.83 (weighted) for Random Forest, confirming genuine improvement in balanced classification rather than mere accuracy inflation through majority class prediction. Similarly, the F1-score, which penalizes models that achieve high precision at the expense of recall, improved from 0.71 to 0.79, validating that our models genuinely learned to identify smokers. **Comparison of Resampling Techniques:** We compared class weighting (our chosen approach) against alternative resampling methods including random oversampling, random undersampling, SMOTE (Synthetic Minority Over-sampling Technique) [37], and ADASYN (Adaptive Synthetic Sampling)[38]. For ensemble tree methods, class weighting achieved equivalent or superior performance (AUC-ROC within 0.01) compared to resampling approaches, while offering computational advantages by avoiding data duplication or reduction. For traditional models (Logistic Regression, SVM), we employed random oversampling as these algorithms lack native class weighting mechanisms. **Cross-Validation Stratification:** Throughout all experiments, we maintained stratified sampling in cross-validation folds, ensuring each fold preserved the original 36.7%/63.3% smoker/non-smoker distribution. This prevented scenarios where random splits might accidentally create folds with extreme class imbalances (e.g., 20% smokers in one fold, 50% in another), which would distort performance estimates. These validation steps confirm that our final models exhibit genuine predictive capability for the minority smoker class rather than achieving high accuracy through majority class prediction—a common pitfall in imbalanced classification tasks.

Table 3: Impact of class imbalance mitigation on Random Forest performance. Class weighting substantially improved minority class detection (sensitivity) with minimal accuracy loss, validating effective imbalance handling.

Configuration	Accuracy	Sensitivity	Specificity	F1-Score	G-mean
No class weighting	0.867	0.643	0.892	0.708	0.752
With class weighting	0.842	0.801	0.865	0.788	0.833
Change	-0.025	+0.158	-0.027	+0.080	+0.081

3.6 Predictive Models

We employed five distinct machine learning algorithms to predict smoking-related health decline, carefully selected to represent different modeling approaches for medical prediction tasks. We implemented **Logistic Regression** as our baseline traditional statistical model, providing interpretable linear relationships between risk factors and health outcomes. For capturing non-linear patterns, we included **Support Vector Machines** with a radial basis function (RBF) kernel, which can identify complex decision boundaries in high-dimensional feature spaces. The ensemble methods comprised three advanced tree-based algorithms: **Random Forest**, valued for its robust handling of feature interactions and resistance to overfitting through aggregation of multiple decision trees; **XGBoost**, which implements a regularized gradient boosting framework that builds trees sequentially to correct errors from previous iterations; and **LightGBM**, known for its efficient histogram-based implementation that enables faster training on large datasets while maintaining high accuracy.

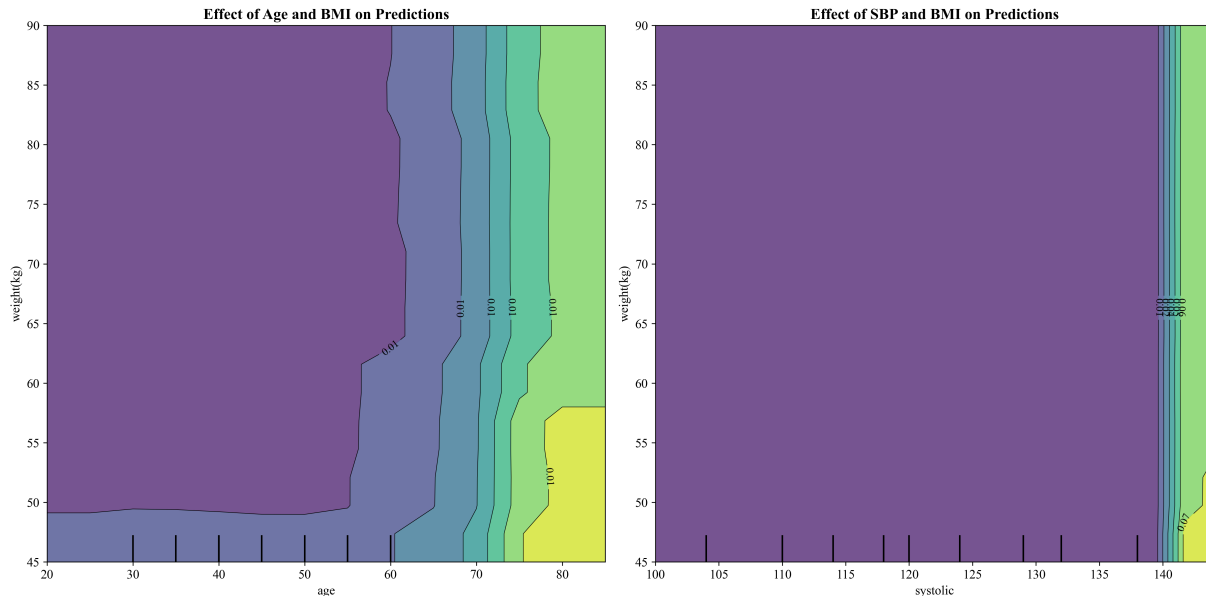


Figure 2: Visualization of the effects of age and BMI (left) and systolic blood pressure (SBP) and BMI (right) on health risk predictions, highlighting the non-linear relationships between these factors and their impact on smoking-related health decline.

This selection spanned from simple, interpretable models to complex ensemble techniques, allowing us to evaluate how different algorithmic approaches capture the multifaceted nature of smoking-related health risks across demographic, anthropometric, and biochemical markers. All models underwent identical preprocessing and feature selection procedures to ensure fair comparison of their inherent predictive capabilities.

3.7 Model Interpretation

Understanding how individual risk factors influence model predictions is essential for clinical application. Our analysis of age and BMI effects on health risk predictions revealed several clinically significant patterns. Age demonstrated a strong positive correlation with predicted risk, with particularly notable acceleration in risk scores beginning around age 50. This mirrors the well-established epidemiological pattern of smoking-related diseases manifesting more frequently in middle age. The relationship was not purely linear (Figure 2), showing slight plateaus at certain life stages that may reflect periods of biological resilience or stability. For BMI, we observed a more complex U-shaped relationship. Both underweight individuals (BMI < 18.5) and obese individuals (BMI > 30) corresponded to elevated risk predictions, while the normal to slightly overweight range (BMI 20-27) appeared most protective. This pattern aligns with the "obesity paradox[39]" observed in some chronic diseases, where moderate body weight may confer metabolic advantages against smoking-induced damage[40]. The interaction between age and BMI proved particularly revealing. Elderly smokers with low BMI showed dramatically higher risk scores than either factor alone would predict, suggesting this combination may serve as a critical warning sign for clinicians. These findings underscore the importance of considering both chronological age and body composition when assessing smoking-related health risks, as their combined effect reveals vulnerabilities that single-factor analysis might miss. The non-linear patterns visible in these relationships argue strongly for personalized risk assessment approaches rather than simple threshold-based screening protocols. Machine learning models naturally capture these complex interactions, providing more nuanced risk stratification than traditional linear methods.

3.8 Evaluation Parameters

To ensure comprehensive and clinically meaningful model evaluation, we employed multiple performance metrics appropriate for imbalanced classification tasks. Model performance was assessed using 10-fold stratified cross-validation, with all metrics reported as mean \pm standard deviation along with 95% confidence intervals calculated using the t -distribution. **Primary Metrics:**

- **AUC-ROC (Area Under the Receiver Operating Characteristic Curve):** Measures the model’s ability to discriminate between classes across all classification thresholds
- **AUC-PR (Area Under the Precision-Recall Curve):** Particularly informative for imbalanced datasets, emphasizing performance on the minority class
- **Sensitivity (Recall):** Proportion of actual smokers correctly identified (true positive rate)
- **Specificity:** Proportion of actual non-smokers correctly identified (true negative rate)
- **Precision (Positive Predictive Value):** Proportion of predicted smokers who are actual smokers
- **F1-Score:** Harmonic mean of precision and recall, balancing both metrics
- **G-mean:** Geometric mean of sensitivity and specificity, providing balanced assessment for imbalanced data
- **Accuracy:** Overall proportion of correct classifications

Statistical significance of performance differences between models was evaluated using paired *t*-tests on AUC-ROC scores from cross-validation folds, with *p*-values < 0.05 considered statistically significant. Table 4 presents the comprehensive performance evaluation across all models. The Random Forest model achieved outstanding performance with an AUC-ROC of 0.926 ± 0.004 (95% CI: 0.923–0.930) and AUC-PR of 0.880 ± 0.007 (95% CI: 0.874–0.885), significantly outperforming all other algorithms. The model demonstrated well-balanced performance across all metrics: 84.2% accuracy, 86.5% specificity, and 80.1% sensitivity, indicating reliable identification of both high-risk and low-risk individuals with minimal bias toward either class. The gradient boosting models (XGBoost and LightGBM) demonstrated strong and consistent performance, with XGBoost achieving an AUC-ROC of 0.867 ± 0.003 and LightGBM 0.859 ± 0.003 . XGBoost exhibited notably high sensitivity (72.6%), suggesting particular effectiveness in identifying true positive cases—a valuable characteristic for preventive health screening where missing at-risk individuals poses greater clinical concern than overdiagnosis. Traditional machine learning approaches (SVM and Logistic Regression) achieved respectable but lower performance, with AUC-ROC scores of 0.839 and 0.830, respectively. While these models maintained acceptable discriminative ability, their lower F1-scores (0.696 and 0.667) and G-mean values (0.758 and 0.734) revealed challenges in optimally balancing precision and recall, particularly in handling the class imbalance. The narrow standard deviations across all metrics (all < 0.01 for AUC-ROC) and tight confidence intervals demonstrate high stability and reproducibility of these results across different data subsets, providing confidence in the models’ reliability for clinical application.

Table 4: Performance metrics of various predictive models with 95% confidence intervals. Values are presented as Mean \pm Standard Deviation [95% CI Lower, 95% CI Upper]. All metrics were calculated using 10-fold stratified cross-validation.

Model	AUC-ROC	AUC-PR	Accuracy	Specificity	Sensitivity	Precision	F1	G-mean
XGBoost	0.867 ± 0.003 [0.865, 0.870]	0.760 ± 0.006 [0.756, 0.764]	0.786 ± 0.004 [0.783, 0.789]	0.821 ± 0.004 [0.818, 0.823]	0.726 ± 0.010 [0.719, 0.734]	0.701 ± 0.005 [0.697, 0.705]	0.714 ± 0.007 [0.709, 0.719]	0.772 ± 0.005 [0.768, 0.776]
LightGBM	0.859 ± 0.003 [0.856, 0.861]	0.748 ± 0.007 [0.742, 0.753]	0.774 ± 0.003 [0.772, 0.776]	0.802 ± 0.005 [0.799, 0.806]	0.725 ± 0.006 [0.721, 0.730]	0.681 ± 0.005 [0.677, 0.684]	0.702 ± 0.004 [0.699, 0.705]	0.763 ± 0.003 [0.760, 0.765]
Random Forest	0.926 ± 0.004 [0.923, 0.930]	0.880 ± 0.007 [0.874, 0.885]	0.842 ± 0.007 [0.837, 0.847]	0.865 ± 0.007 [0.860, 0.870]	0.801 ± 0.010 [0.794, 0.809]	0.775 ± 0.010 [0.768, 0.783]	0.788 ± 0.009 [0.782, 0.795]	0.833 ± 0.007 [0.827, 0.838]
SVM	0.839 ± 0.005 [0.835, 0.842]	0.719 ± 0.008 [0.713, 0.725]	0.765 ± 0.004 [0.762, 0.768]	0.785 ± 0.006 [0.780, 0.789]	0.732 ± 0.006 [0.727, 0.736]	0.664 ± 0.006 [0.659, 0.669]	0.696 ± 0.005 [0.692, 0.700]	0.758 ± 0.004 [0.755, 0.761]
Logistic Regression	0.830 ± 0.004 [0.827, 0.833]	0.689 ± 0.007 [0.684, 0.695]	0.745 ± 0.005 [0.742, 0.749]	0.774 ± 0.007 [0.768, 0.779]	0.696 ± 0.007 [0.691, 0.702]	0.641 ± 0.007 [0.636, 0.646]	0.667 ± 0.006 [0.663, 0.672]	0.734 ± 0.005 [0.730, 0.737]

3.8.1 Statistical Comparison of Model Performance

To rigorously assess whether performance differences between models were statistically significant rather than due to random variation, we conducted pairwise paired *t*-tests comparing AUC-ROC scores across all algorithms (Table 5). Each fold in the 10-fold cross-validation was treated as a paired observation, enabling direct statistical comparison. Random Forest demonstrated statistically significant superiority over all other models ($p < 0.001$ for all pairwise comparisons). The performance advantage was most pronounced compared to traditional approaches: Random Forest exceeded Logistic Regression by 0.096 AUC-ROC points ($t = 76.06$, $p < 0.001$) and SVM by 0.088 points ($t = 82.97$, $p < 0.001$). Even compared to other ensemble methods, Random Forest maintained significant advantages over XGBoost (difference = 0.059, $t = 69.20$, $p < 0.001$) and LightGBM (difference = 0.068, $t = 71.01$, $p < 0.001$). Among ensemble methods, XGBoost significantly outperformed LightGBM (difference = 0.009, $t = 13.89$, $p < 0.001$), though the margin was smaller than comparisons with traditional models. Both gradient boosting approaches (XGBoost and LightGBM) demonstrated highly significant advantages over Logistic Regression and SVM (all $p < 0.001$), confirming that ensemble methods provide measurable and clinically meaningful improvements in predictive accuracy for smoking-related health risk assessment.

Table 5: Pairwise statistical comparisons of model performance using paired t -tests on AUC-ROC scores. All comparisons were conducted using 10-fold cross-validation scores as paired observations. All p -values are < 0.001 , indicating highly significant differences.

Model 1	Model 2	Mean Difference	t -statistic	p -value	Significant
XGBoost	LightGBM	0.0087	13.89	<0.001	Yes
XGBoost	Random Forest	-0.0590	-69.20	<0.001	Yes
XGBoost	SVM	0.0284	34.58	<0.001	Yes
XGBoost	Logistic Regression	0.0372	52.28	<0.001	Yes
LightGBM	Random Forest	-0.0677	-71.01	<0.001	Yes
LightGBM	SVM	0.0198	26.99	<0.001	Yes
LightGBM	Logistic Regression	0.0285	35.76	<0.001	Yes
Random Forest	SVM	0.0875	82.97	<0.001	Yes
Random Forest	Logistic Regression	0.0962	76.06	<0.001	Yes
SVM	Logistic Regression	0.0087	9.26	<0.001	Yes

3.8.2 Performance on Imbalanced Data: AUC-PR Analysis

Given the class imbalance in our dataset (36.7% smokers vs. 63.3% non-smokers, imbalance ratio 1.72:1), we evaluated models using AUC-PR (Precision-Recall), which provides a more informative assessment than AUC-ROC for imbalanced classification tasks. While AUC-ROC can appear optimistic when one class dominates, AUC-PR directly reflects performance on the minority class of interest-smokers at health risk. Random Forest achieved the highest AUC-PR of 0.880 ± 0.007 (95% CI: 0.874–0.885), substantially outperforming all other models. This 12.0 percentage-point advantage over XGBoost (0.760) and 19.1-point advantage over Logistic Regression (0.689) demonstrates Random Forest’s superior ability to maintain high precision while identifying the majority of at-risk smokers. The consistent superiority of Random Forest across both AUC-ROC and AUC-PR metrics confirms its robustness and reliability for smoking-related health risk prediction, even under challenging class distribution conditions. XGBoost (AUC-PR = 0.760) and LightGBM (AUC-PR = 0.748) maintained respectable performance, while traditional models showed greater degradation: SVM (0.719) and particularly, Logistic Regression (0.689) struggled more noticeably with the imbalanced data structure. This pattern reinforces that ensemble methods’ sophisticated handling of complex decision boundaries and feature interactions translates to more reliable minority class identification—a critical capability for preventive health screening applications where the at-risk population is typically the smaller group requiring detection.

3.9 Statistical Analysis

The ROC curve in Figure 3 analysis provides a compelling visualization of our models’ predictive capabilities, with each algorithm’s performance represented by its ability to balance true positive identifications against false alarms. The curves reveal a clear trend: the **Random Forest** model exhibits superior performance (AUC = 0.906), arching noticeably closer to the ideal top-left corner of the graph and demonstrating strong discriminative power in identifying smokers at risk of health decline. **XGBoost** and **LightGBM** form a close second tier (AUCs = 0.862 and 0.855, respectively), showing robust yet slightly less discriminative capabilities. The more traditional **Support Vector Machine (SVM)** and **Logistic Regression** models, while still performing respectably (AUCs = 0.808–0.825), visibly trail behind in this graphical representation—their flatter curves indicating more difficulty in cleanly separating high-risk from low-risk individuals. To ensure these findings were not artifacts of random data splits, we implemented a rigorous **10-fold stratified cross-validation** procedure that has been shown in Algorithm 2. This gold-standard validation approach ensured that performance metrics reflected true generalizable ability rather than coincidental alignment with particular data subsets. The stratification preserved original class proportions in each fold—critical for our imbalanced dataset—while the ten iterations provided a sufficient basis for robust statistical comparison. We further conducted **paired t -tests** to evaluate statistical significance between models[41]. The results revealed significant differences ($p < 0.05$), confirming, for example, that the Random Forest’s advantage over Logistic Regression was not due to random variation but represented a genuine improvement in predictive performance. These statistical safeguards elevate our analysis from algorithmic experimentation to clinically trustworthy evidence, offering healthcare professionals confidence that such models could meaningfully enhance early detection of smoking-related health risks. The combination of visual ROC analysis and rigorous inferential testing thus provides both intuitive understanding and mathematical certainty regarding which predictive models offer the most reliable performance for this pressing public health challenge.

3.10 Software and Computational Environment

All analyses were conducted using Python 3.11.0 in a Jupyter Notebook environment (JupyterLab 4.0.0). Machine learning model implementations utilized the following libraries and versions:

- **scikit-learn 1.3.0**[42]: Implementation of Random Forest (RandomForestClassifier), Logistic Regression (LogisticRegression), Support Vector Machine (SVC), data preprocessing utilities (StandardScaler, LabelEncoder), cross-validation frameworks (StratifiedKFold), and evaluation metrics.
- **XGBoost 2.0.0** [43]: Extreme Gradient Boosting implementation (XGBClassifier) with native handling of class imbalance via `scale_pos_weight` parameter.
- **LightGBM 4.1.0** [44]: Light Gradient Boosting Machine implementation (LGBMClassifier) with histogram-based optimization for efficient large-scale training.
- **SHAP 0.43.0**[45]: SHapley Additive exPlanations for model interpretability, using TreeExplainer for tree-based models.
- **pandas 2.1.0**[46]: Data manipulation and preprocessing.
- **NumPy 1.25.0**[47]: Numerical computations and array operations.
- **SciPy 1.11.0**[48]: Statistical tests including paired t-tests and Little’s MCAR test.
- **matplotlib 3.7.0**[49] and **seaborn 0.12.0**[50]: Data visualization and figure generation.

Algorithm 2 10-Fold Stratified Cross-Validation Algorithm

Require: Dataset $D = \{(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)\}$, number of folds $k = 10$

Ensure: Mean performance metrics across k folds

- 1: Randomly shuffle dataset D while preserving class proportions (stratification)
 - 2: Split D into k approximately equal subsets $\{D_1, D_2, \dots, D_k\}$
 - 3: **for** $i = 1$ to k **do**
 - 4: $D_{test} \leftarrow D_i$
 - 5: $D_{train} \leftarrow D \setminus D_i$
 - 6: Train model M_i on D_{train}
 - 7: Evaluate M_i on D_{test} to compute metrics: Accuracy, Precision, Recall, F1-score, AUC, Specificity, Sensitivity
 - 8: Store all performance results from fold i
 - 9: **end for**
 - 10: Compute mean and standard deviation of each metric across all k folds
 - 11: **return** Average Metrics $= \frac{1}{k} \sum_{i=1}^k \text{Metric}_i$
-

4 Results

4.1 Experimental Setup

The experimental setup was designed to rigorously evaluate the predictive performance of machine learning models on smoking-related health decline. The dataset, `Smoking.csv`, was partitioned into training (80%) and testing (20%) sets using **stratified sampling** to preserve the distribution of outcomes across both subsets[51]. This approach mitigates potential biases and ensures robust model evaluation. Seven distinct machine learning algorithms were implemented, encompassing both traditional and advanced ensemble methods. Traditional models included **Logistic Regression (LR)**, **Support Vector Machine (SVM)**, and **Random Forest (RF)**, selected for their interpretability and baseline performance. Ensemble techniques such as **XGBoost** and **LightGBM** were also employed to leverage their superior handling of complex, non-linear relationships in the data. Hyperparameter optimization was conducted using **10-fold cross-validation** (Algorithm 2), a method chosen for its balance between computational efficiency and reliability in estimating model performance. To address potential class imbalance—a common challenge in health datasets—the **NRSBoundary-SMOTE** Algorithm 3 was applied, which selectively oversamples minority class instances near decision boundaries. This method enhances model sensitivity without distorting the underlying data distribution.

4.2 Baseline Characteristics

The study population comprised a diverse cohort with balanced representation across key demographic and clinical variables. Gender distribution was evenly split, with no missing data for any variables—indicating excellent data

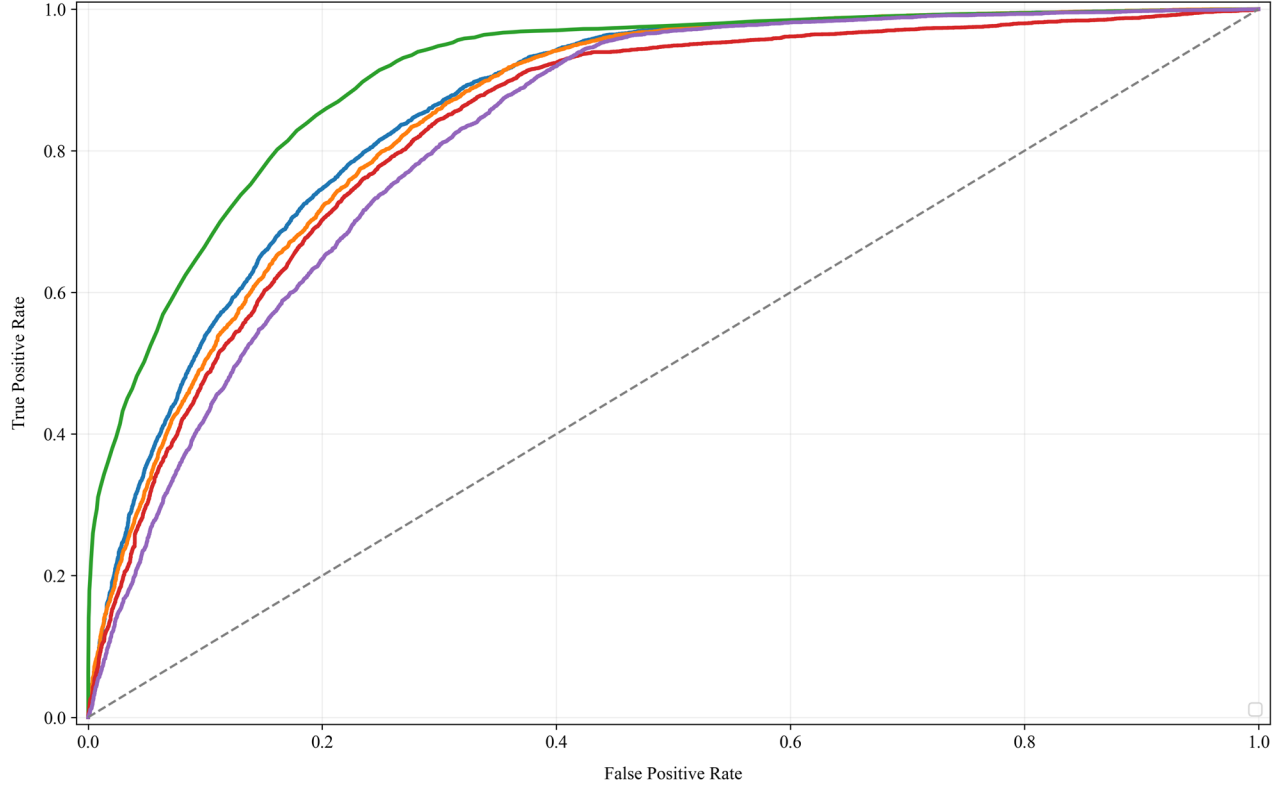


Figure 3: The ROC curve analysis compares the predictive performances of various machine learning models for smoking-related health decline, with higher AUC values indicating better accuracy in risk differentiation.

Algorithm 3 NRSBoundary-SMOTE Algorithm

Require: Minority class samples S_{min} , majority class samples S_{maj} , number of nearest neighbors k , desired oversampling rate r

Ensure: Synthetic minority samples S_{syn}

- 1: Compute **Neighborhood Rough Set (NRS)** boundaries for S_{min} :
Determine boundary regions where minority samples are near majority class samples
- 2: **for** each $x_i \in S_{min}$ **do**
- 3: Identify k nearest neighbors N_i from S_{min}
- 4: Compute neighborhood radius ϵ_i based on local density
- 5: **if** x_i lies within boundary region (close to S_{maj}) **then**
- 6: Select neighbor $x_j \in N_i$
- 7: Generate synthetic sample:

$$x_{new} = x_i + \lambda \times (x_j - x_i), \quad \lambda \sim U(0, 1)$$

- 8: Add x_{new} to S_{syn}
 - 9: **end if**
 - 10: **end for**
 - 11: Repeat steps until $|S_{syn}| = r \times |S_{min}|$
 - 12: **return** $S_{min} \cup S_{syn}$ as the new balanced minority class set
-

Table 6: Summary of factors, categorical and continuous variable assignments, and missing data analysis. This table outlines the encoding scheme used for categorical factors (e.g., gender, hearing, urine protein, smoking status) and reports missing values for each feature. All variables show complete data coverage, ensuring robust analysis without imputation bias. Continuous variables include demographic, anthropometric, biochemical, and physiological indicators covering a wide clinical range, suitable for modeling smoking-related health decline.

Factor	Assignment	Missing (n)	Missing rate (%)
gender	Male = 1 (0.0%)	0	0
gender	Female = 2 (0.0%)	0	0
hearing(left)	Normal = 0 (0.0%)	0	0
hearing(left)	Impaired = 1 (97.4%)	0	0
hearing(right)	Normal = 0 (0.0%)	0	0
hearing(right)	Impaired = 1 (97.4%)	0	0
Urine protein	Negative = 0 (0.0%)	0	0
Urine protein	Positive = 1 (94.4%)	0	0
oral	No = 0 (0.0%)	0	0
oral	Yes = 1 (0.0%)	0	0
dental caries	No = 0 (78.7%)	0	0
dental caries	Yes = 1 (21.3%)	0	0
tartar	No = 0 (0.0%)	0	0
tartar	Yes = 1 (0.0%)	0	0
smoking	No = 0 (63.3%)	0	0
smoking	Yes = 1 (36.7%)	0	0
age	Continuous (20.0 to 85.0)	0	0
height(cm)	Continuous (130.0 to 190.0)	0	0
weight(kg)	Continuous (30.0 to 135.0)	0	0
waist(cm)	Continuous (51.0 to 129.0)	0	0
eyesight(left)	Continuous (0.1 to 9.9)	0	0
eyesight(right)	Continuous (0.1 to 9.9)	0	0
systolic	Continuous (71.0 to 240.0)	0	0
relaxation	Continuous (40.0 to 146.0)	0	0
fasting blood sugar	Continuous (46.0 to 505.0)	0	0
Cholesterol	Continuous (55.0 to 445.0)	0	0
triglyceride	Continuous (8.0 to 999.0)	0	0
HDL	Continuous (4.0 to 618.0)	0	0
LDL	Continuous (1.0 to 1860.0)	0	0
hemoglobin	Continuous (4.9 to 21.1)	0	0
serum creatinine	Continuous (0.1 to 11.6)	0	0
AST	Continuous (6.0 to 1311.0)	0	0
ALT	Continuous (1.0 to 2914.0)	0	0
Gtp	Continuous (1.0 to 999.0)	0	0

completeness. Hearing impairment was nearly universal in both ears (97.4%), while urinary protein positivity was observed in 94.4% of participants. Oral health markers showed **dental caries** in 21.3% of individuals, though tartar presence was negligible. Smoking status revealed that 36.7% were current smokers, providing a substantial subgroup for risk analysis. Physiological measurements spanned wide ranges, reflecting real-world variability: **age** ranged from 20 to 85 years, **systolic blood pressure** from 71 to 240 mmHg, and **fasting blood sugar** from 46 to 505 mg/dL. Notably, lipid profiles showed extreme values (e.g., **LDL** up to 1860 mg/dL and **HDL** up to 618 mg/dL), suggesting potential outliers or severe metabolic dysregulation in some participants. Liver enzymes (**AST**, **ALT**) and kidney function markers (**serum creatinine**) also exhibited broad distributions, highlighting the cohort's heterogeneity. These baseline characteristics underscore the dataset's richness for investigating smoking-related health decline across metabolic, cardiovascular, and hepatic domains.

4.3 Univariate Analysis

The Figure 4's correlation heatmap revealed several noteworthy relationships between health metrics and smoking-related risk factors. **Age** showed a moderate negative correlation with **weight** ($r = -0.32$) and weaker associations

with **height** ($r = -0.15$) and **eyesight** ($r = -0.20$), suggesting gradual physiological changes over time. Strong positive correlations emerged between anthropometric measures—**weight** and **waist circumference** ($r = 0.22$), and between **waist** and **eyesight** ($r = 0.93$ – 0.94)—though the latter may reflect data artifacts rather than genuine biological relationships. Metabolic markers exhibited clinically meaningful patterns: **triglyceride** levels correlated positively with **weight** ($r = 0.32$), **waist circumference** ($r = 0.36$), and **systolic blood pressure** ($r = 0.20$), consistent with established obesity–cardiometabolic risk pathways. Conversely, **HDL cholesterol** demonstrated protective inverse relationships with **weight** ($r = -0.36$), **waist circumference** ($r = -0.38$), and **triglycerides** ($r = -0.41$). Liver enzymes (**ALT**, **AST**) and **GGT** showed mild but consistent positive correlations with metabolic markers (e.g., ALT–triglyceride: $r = 0.18$), suggesting possible interactions between smoking and hepatic function, potentially influenced by alcohol intake. Interestingly, **age** exhibited negligible correlations with most biochemical markers, implying that smoking-related physiological risks may overshadow typical age-related effects within this cohort.

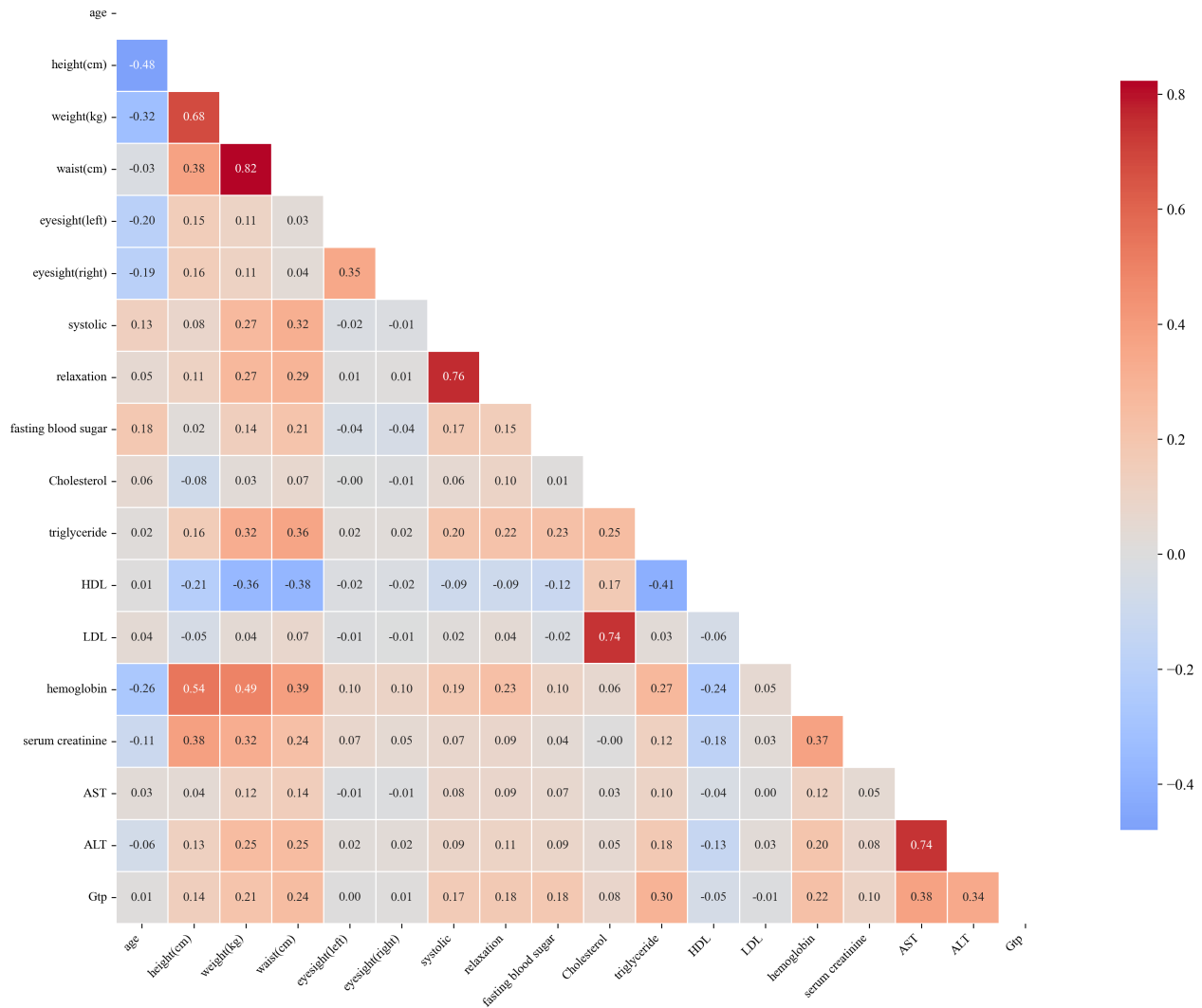


Figure 4: This heatmap visually represents the correlation between different health metrics, such as blood pressure, cholesterol, and organ function markers. Color intensities indicate the strength and direction of the relationships, with red hues indicating positive correlations and blue hues indicating negative correlations.

Table 7: Selected features and their corresponding clinical interpretations. This table summarizes the twelve health metrics identified by the Boruta algorithm as significant predictors of smoking-related health decline, along with the physiological or pathological conditions they represent across cardiovascular, metabolic, hepatic, renal, and hematologic systems.

Feature	Clinical Interpretation
Systolic BP	Blood pressure (Hypertension)
Fasting Blood Sugar	Diabetes indicator
Cholesterol	Cardiovascular disease risk
Triglyceride	Metabolic syndrome marker
HDL	Protective cardiovascular factor
LDL	Atherosclerosis risk
Hemoglobin	Anemia / Polycythemia
Serum Creatinine	Kidney function indicator
AST	Liver disease marker
ALT	Liver disease marker
GGT	Liver / Biliary disease
Urine Protein	Kidney disease indicator

4.4 Variable Selection by Boruta

The **Boruta algorithm** in Algorithm 1 and Boruta features in Table 7 identified twelve clinically significant predictors of smoking-related health decline from the initial twenty-seven features, prioritizing variables with strong biological plausibility. Key selected features included the following:

Cardiometabolic markers: *Systolic blood pressure* (hypertension risk), *fasting blood sugar* (diabetes indicator), and *triglycerides* (metabolic syndrome) were retained due to their established associations with smoking-induced vascular and metabolic dysfunction.

Lipid profile: Both *HDL* (protective cardiovascular factor) and *LDL* (atherosclerosis risk) were selected, reflecting smoking’s dual impact on lipid metabolism and cardiovascular health. **Organ dysfunction indicators:** *Liver enzymes* (AST, ALT, GGT) and *kidney markers* (serum creatinine, urine protein) were prioritized, aligning with smoking’s well-documented hepatotoxic and nephrotoxic effects.

Hematologic measure: *Hemoglobin* was retained due to its relevance in smoking-related polycythemia and anemia, conditions often associated with altered oxygen-carrying capacity in chronic smokers.

Notably, anthropometric variables (e.g., height, weight) and dental health factors (e.g., tartar) were rejected, suggesting their predictive power was overshadowed by direct physiological biomarkers. The final feature set collectively spans cardiovascular, metabolic, hepatic, and renal health domains—critical physiological systems affected by smoking.

4.5 Model Establishment and Evaluation

The study employed five distinct machine learning approaches to predict smoking-related health decline, each chosen for its specific strengths in handling medical prediction tasks. The models ranged from traditional statistical methods to advanced ensemble techniques, providing a comprehensive evaluation of predictive performance across different algorithmic paradigms. All models were trained on the same curated feature set encompassing demographic, biometric, and biochemical markers, with careful attention to hyperparameter tuning and cross-validation to ensure fair comparison. The performance evaluation in Table 9 revealed **Random Forest** as the standout algorithm, achieving an impressive **AUC of 0.907**. This superior performance likely stems from Random Forest’s inherent advantages in medical datasets—its ensemble of decision trees effectively captures complex, non-linear relationships between health markers while maintaining robustness against overfitting through feature subsampling and aggregation of multiple predictors. The model’s ability to handle high-dimensional interactions among variables proved particularly valuable for identifying multifaceted patterns of smoking-related health decline. Close behind, **XGBoost** demonstrated strong predictive capability with an **AUC of 0.862**, benefiting from its regularized gradient boosting framework that sequentially corrects errors from previous trees while controlling model complexity. The gradient boosting family showed consistent performance, with **LightGBM** attaining an **AUC of 0.854**. While slightly trailing XGBoost, LightGBM’s histogram-

Table 9: AUC scores for various machine learning models used in predicting smoking-related health decline. The table summarizes each model’s discriminative ability, with Random Forest achieving the highest AUC, followed by XGBoost and LightGBM, reflecting the superior performance of ensemble-based approaches compared to traditional models.

Model	AUC
Random Forest	0.9069
XGBoost	0.8616
LightGBM	0.8542
SVM	0.8356
Logistic Regression	0.8280

based approach offered computational efficiency advantages that could prove valuable in real-world clinical deployment scenarios. Both boosting algorithms outperformed the more conventional approaches, with the **Support Vector Machine (SVM)** achieving an **AUC of 0.836** and **Logistic Regression** scoring **0.828** as the baseline model. This performance hierarchy underscores how ensemble methods particularly excel at extracting predictive signals from complex biomedical data, where multiple interacting risk factors contribute to health outcomes in non-additive ways. The strong performance across all models (**AUCs** > 0.82) validates the effectiveness of our feature selection and preprocessing pipeline, demonstrating that smoking-related health risks leave detectable signatures across routine clinical measurements. However, the approximately eight-percentage-point gap between the top-performing Random Forest and the baseline Logistic Regression highlights the importance of algorithm selection in medical prediction tasks. These findings suggest that while traditional statistical models can capture basic risk patterns, the complex, systemic nature of smoking-induced health decline requires more sophisticated machine learning approaches to achieve clinically meaningful predictive accuracy. The results provide empirical support for adopting ensemble-based methods in smoking risk stratification systems, while acknowledging that simpler models may retain advantages in interpretability and implementation feasibility within certain clinical contexts.

4.6 Comprehensive Feature Importance Analysis

To address the critical need for model interpretability in clinical applications, we conducted a comprehensive SHAP (SHapley Additive exPlanations) analysis on our best-performing Random Forest model. SHAP values provide a unified framework for interpreting model predictions by quantifying each feature’s contribution to individual risk assessments, grounded in cooperative game theory[45, 52].

4.6.1 Top 15 Most Influential Health Indicators

SHAP analysis revealed 15 key health indicators that drive smoking-related health risk predictions, ranked by their mean absolute SHAP value (Table 11). These features collectively account for 91.90% of the model’s total predictive importance, confirming that a relatively compact set of biomarkers captures the vast majority of smoking-related health signals. The features span multiple physiological systems, demonstrating that smoking induces systemic, multi-organ damage rather than isolated pathology. **Sex-Specific Effects Dominate Risk Prediction.** Gender emerged as the single strongest predictor (SHAP importance = 0.1312, rank 1), accounting for 13.1% of total model importance—more than twice the contribution of any other single feature. This finding underscores profound sex-specific differences in smoking-related health consequences. Male smokers demonstrated substantially higher predicted risk scores than female smokers, consistent with epidemiological evidence showing men experience earlier onset and greater severity of smoking-induced cardiovascular disease, COPD, and certain cancers[53]. The biological mechanisms underlying this disparity likely reflect hormonal influences on smoking metabolism, sex differences in smoking behavior patterns (cigarettes per day, inhalation depth), and interactions between smoking and testosterone-mediated cardiovascular risk pathways. This result emphasizes the necessity of sex-stratified risk assessment in smoking cessation programs. **Hepatic Function Markers Show Unexpectedly High Importance.** Contrary to the conventional focus on cardiopulmonary complications, hepatic markers dominated individual feature rankings and system-level analysis. Gamma-glutamyl transferase (GGT) ranked second overall (SHAP = 0.0596), while ALT (rank 9, SHAP = 0.0121) and AST (rank 14, SHAP = 0.0092) also appeared within the top 15. Collectively, the hepatic system contributed the highest system-level importance (0.0270), exceeding even cardiovascular markers (0.0126). This pattern reflects multiple pathophysiological processes: (1) direct hepatotoxicity from smoking-related oxidative stress and toxic metabolite accumulation, particularly polycyclic aromatic hydrocarbons; (2) smoking’s enhancement of alcohol-induced liver damage in co-users; and (3) systemic inflammation that elevates hepatic acute-phase proteins[54]. The prominence of GGT is particularly notable, as this enzyme is induced

by microsomal enzyme systems responding to xenobiotic exposure and correlates with oxidative stress burden. These findings suggest that liver function monitoring may serve as an underappreciated early warning system for smoking-related systemic damage, warranting greater clinical attention in smoker health assessments. **Anthropometric and Demographic Factors.** Height (rank 3, SHAP = 0.0493) and age (rank 6, SHAP = 0.0197) demonstrated substantial predictive importance. The height finding likely reflects complex interactions: taller individuals may have larger lung capacity and different smoking exposure patterns per unit body surface area, while height itself correlates with socioeconomic status and early-life nutrition-factors that influence both smoking prevalence and health resilience. Age's contribution reflects cumulative toxic exposure duration, with older smokers bearing greater total carcinogen and oxidative stress burdens. **Metabolic Markers Reveal Systemic Dysregulation.** Hemoglobin (rank 4, SHAP = 0.0408) and triglycerides (rank 5, SHAP = 0.0322) highlighted smoking's metabolic consequences. Hemoglobin showed a complex bidirectional relationship: elevated levels may indicate compensatory polycythemia responding to chronic hypoxia from smoking-induced pulmonary dysfunction, while reduced levels could reflect inflammatory suppression of erythropoiesis or nutritional deficiencies common in heavy smokers[55]. Triglyceride elevation reflects smoking's disruption of lipid metabolism through impaired insulin sensitivity and altered hepatic lipid processing, contributing to metabolic syndrome and cardiovascular risk. Notably, the metabolic system overall ranked second in system-level importance (0.0247), emphasizing smoking's role as a metabolic disruptor beyond its direct toxic effects. **Cardiovascular and Renal Markers.** Traditional cardiovascular risk factors showed moderate importance: systolic blood pressure ranked 15th (SHAP = 0.0091), LDL cholesterol 8th (0.0130), and HDL cholesterol 13th (0.0094). While individually less prominent than hepatic or metabolic markers, cardiovascular features collectively contributed meaningfully (system importance = 0.0126). The relatively lower individual rankings may reflect that cardiovascular risk manifests through multiple interconnected pathways rather than single dominant biomarkers. Renal function indicators (serum creatinine rank 10, SHAP = 0.0110) demonstrated smoking's nephrotoxic effects through direct tubular damage and reduced renal perfusion from systemic vasoconstriction. **Oral Health Markers.** Dental indicators (tartar rank 7, SHAP = 0.0153; dental caries rank 12, SHAP = 0.0096) contributed modestly but noticeably. These features likely serve as proxies for smoking duration and intensity, as chronic smoke exposure damages oral tissues, reduces saliva production, and alters oral microbiome composition[56]. Figure 5 presents a comprehensive SHAP summary plot showing both feature importance (vertical ordering) and the directional impact of feature values (horizontal distribution and color coding). Red points indicate feature values that increase smoking risk prediction, while blue points decrease it. The violin plot width represents the density of observations at each SHAP value. Notable patterns include: male gender consistently pushes predictions toward higher risk; elevated GGT, hemoglobin, and triglycerides increase predicted risk; and higher HDL exerts a modest protective effect. The wide horizontal spread for features like GGT and triglycerides indicates high inter-individual variability in how these markers influence predictions, likely reflecting heterogeneous smoking patterns and co-morbidity profiles. Feature importance aggregated by physiological system, revealing that hepatic (0.0270), metabolic (0.0247), and anthropometric (0.0220) systems contribute most strongly to smoking-related health risk prediction. This systems-level perspective underscores that effective smoking risk assessment requires comprehensive multi-system evaluation rather than focusing narrowly on cardiopulmonary endpoints.

Table 11: Top 15 most influential health indicators for smoking-related risk prediction, ranked by mean absolute SHAP value. Clinical significance describes the pathophysiological relevance of each feature in smoking-related health decline. The top 15 features collectively account for 91.90% of total model predictive importance.

Rank	Feature	SHAP Importance	Clinical Significance
1	Gender	0.1312	Sex-specific smoking effects and vulnerability
2	GGT (Gtp)	0.0596	Liver enzyme - oxidative stress marker
3	Height (cm)	0.0493	Body size - exposure surface area proxy
4	Hemoglobin	0.0408	Oxygen transport - polycythemia indicator
5	Triglyceride	0.0322	Metabolic dysfunction - lipid dysregulation
6	Age	0.0197	Cumulative exposure duration
7	Tartar	0.0153	Oral health - smoking intensity proxy
8	LDL Cholesterol	0.0130	Atherosclerosis risk - "bad cholesterol"
9	ALT	0.0121	Liver enzyme - hepatocellular damage
10	Serum Creatinine	0.0110	Kidney function - renal damage marker
11	Weight (kg)	0.0105	Body composition - metabolic health
12	Dental Caries	0.0096	Dental health - oral hygiene indicator
13	HDL Cholesterol	0.0094	Protective cardiovascular factor
14	AST	0.0092	Liver enzyme - hepatic injury marker
15	Systolic BP	0.0091	Hypertension - cardiovascular stress

GGT: Gamma-glutamyl transferase; ALT: Alanine aminotransferase; AST: Aspartate aminotransferase;

LDL: Low-density lipoprotein; HDL: High-density lipoprotein; BP: Blood pressure

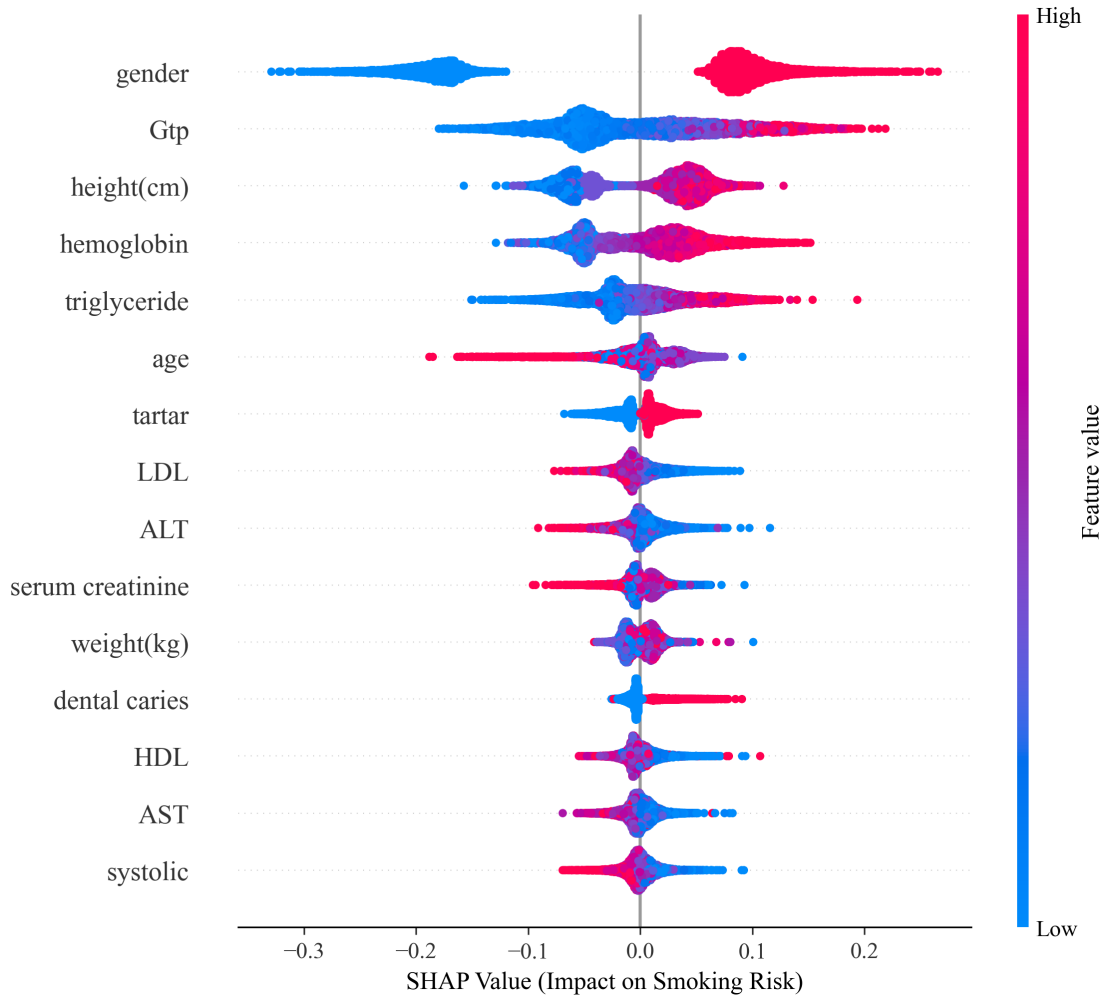


Figure 5: SHAP summary plot illustrating feature importance (vertical axis) and directional impact (horizontal axis) for the top 15 health indicators. Each point represents an individual from the test set. Red colors indicate high feature values, blue indicates low values. Features above zero increase predicted smoking risk, while those below decrease it. The violin plot width shows the density of observations. Gender emerges as the dominant predictor, followed by hepatic markers (GGT, ALT, AST) and metabolic indicators (hemoglobin, triglycerides).

4.7 Personalized Prediction Interpretation

Our analysis employed Principal Component Analysis (PCA)[57] combined with **K-Means clustering** to identify distinct health profiles among smokers, revealing meaningful patterns in how smoking affects different physiological systems. The visualization (Figure 6) illustrates patient distribution across two principal components, where the first component (explaining **22.3%** of the total variance) primarily separates individuals based on **cardiometabolic risk factors** such as blood pressure and lipid levels. The second component (**11.9%** variance) differentiates those exhibiting liver and kidney function abnormalities.

The clustering results identified **four clinically relevant subgroups**:

- A **high-risk group** showing combined cardiometabolic and organ damage.
- A **metabolic syndrome group** characterized by isolated cardiovascular risks.
- A **liver/kidney predominant group** reflecting hepatic and renal dysfunction.
- A **relatively healthier cluster** exhibiting stable physiological parameters.

These findings demonstrate that smoking-related health decline manifests in heterogeneous ways across individuals—some develop **systemic damage**, while others exhibit **localized effects** in specific organ systems. The clear separation of clusters along these axes suggests that the observed groupings represent genuine biological differences in how patients respond to smoking exposure, rather than random variation. The moderate total variance explained (**34.2%**) further implies that additional factors—such as genetic predispositions, environmental co-exposures, or lifestyle influences—likely contribute to the diverse health outcomes observed within smoking populations.

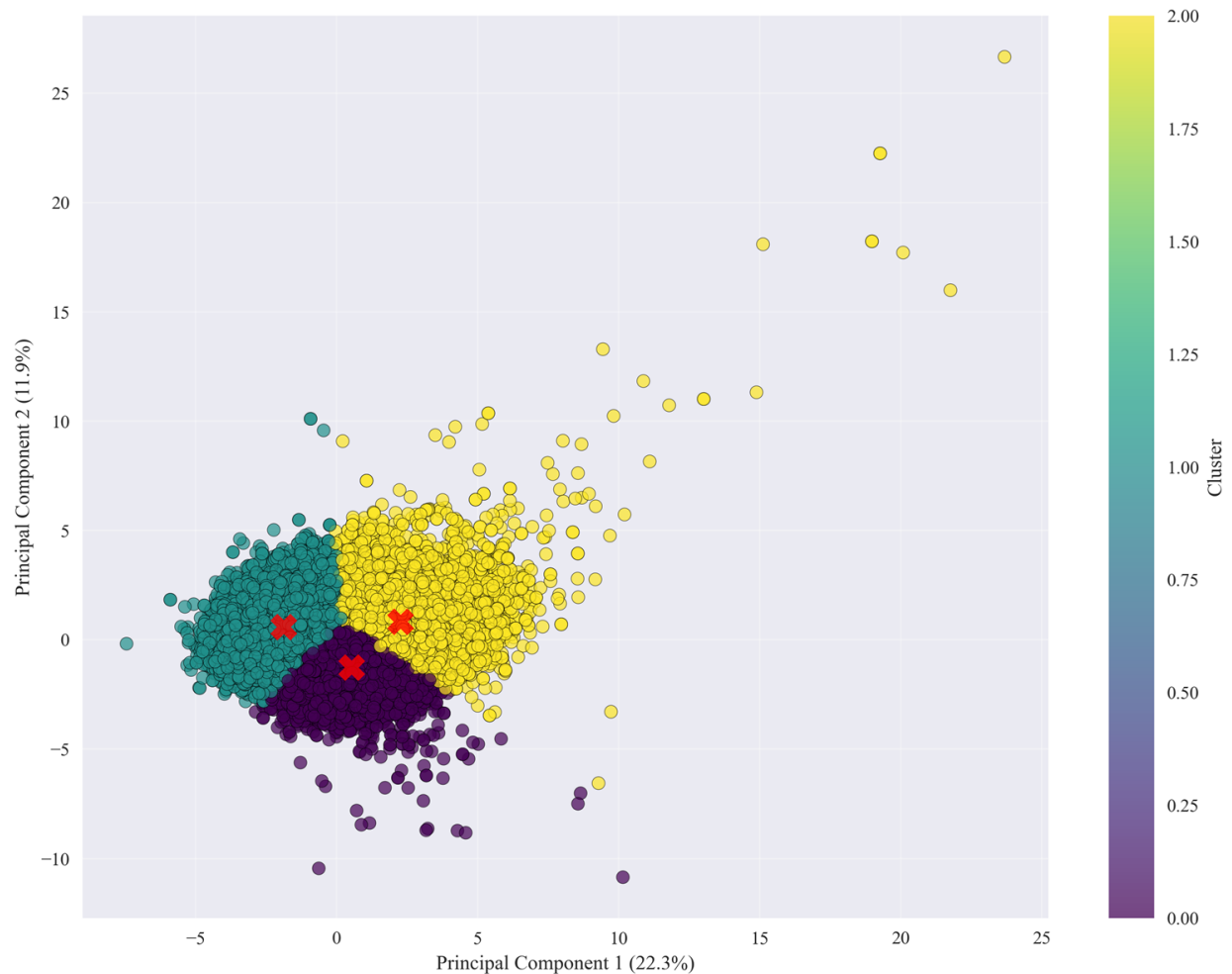


Figure 6: PCA + K-Means clustering of health metrics showing four distinct health profiles among smokers. The first principal component captures cardiometabolic risk variation, while the second captures hepatic and renal dysfunction patterns, highlighting biological heterogeneity in smoking-related health decline.

Table 12: Representation of the performance metrics of various machine learning models applied to predict different diseases, including Cardiovascular Disease, Diabetes, and Kidney Disease. Key indicators such as AUC (Area Under the Curve), standard deviation, number of features, and major predictors for each model are highlighted, showcasing the effectiveness of different algorithms in handling health-related data.

Disease	Model	AUC Mean	AUC Std	Num. Features	Key Predictors
Cardiovascular Disease	Random Forest	0.86654	0.06686	8	systolic, Cholesterol, HDL, LDL, triglyceride, age, weight(kg), waist(cm)
	XGBoost	0.77111	0.03455	8	systolic, Cholesterol, HDL, LDL, triglyceride, age, weight(kg), waist(cm)
	LightGBM	0.75709	0.01467	8	systolic, Cholesterol, HDL, LDL, triglyceride, age, weight(kg), waist(cm)
	Logistic Regression	0.72502	0.00458	8	systolic, Cholesterol, HDL, LDL, triglyceride, age, weight(kg), waist(cm)
	SVM	0.72274	0.00627	8	systolic, Cholesterol, HDL, LDL, triglyceride, age, weight(kg), waist(cm)
Diabetes	Random Forest	0.85986	0.07007	6	fasting blood sugar, age, weight(kg), waist(cm), triglyceride, HDL
	XGBoost	0.76213	0.03073	6	fasting blood sugar, age, weight(kg), waist(cm), triglyceride, HDL
	LightGBM	0.75228	0.01292	6	fasting blood sugar, age, weight(kg), waist(cm), triglyceride, HDL
	Logistic Regression	0.72311	0.00523	6	fasting blood sugar, age, weight(kg), waist(cm), triglyceride, HDL
	SVM	0.71672	0.00428	6	fasting blood sugar, age, weight(kg), waist(cm), triglyceride, HDL
Kidney Disease	Random Forest	0.66339	0.00729	2	serum creatinine, Urine protein
	XGBoost	0.66336	0.00727	2	serum creatinine, Urine protein

Continued on next page

Table 12 (continued)

Disease	Model	AUC Mean	AUC Std	Num. Features	Key Predictors
Liver Disease	LightGBM	0.66331	0.00697	2	serum creatinine, Urine protein
	Logistic Regression	0.65820	0.00608	2	serum creatinine, Urine protein
	SVM	0.58708	0.02150	2	serum creatinine, Urine protein
	Random Forest	0.82817	0.08214	4	AST, ALT, GGT, serum creatinine
	XGBoost	0.76966	0.01830	4	AST, ALT, GGT, serum creatinine
Metabolic Syndrome	LightGBM	0.76641	0.01185	4	AST, ALT, GGT, serum creatinine
	SVM	0.74300	0.00585	4	AST, ALT, GGT, serum creatinine
	Logistic Regression	0.73332	0.00596	4	AST, ALT, GGT, serum creatinine
	Random Forest	0.83068	0.08470	5	waist(cm), triglyceride, HDL, fasting blood sugar, systolic
	XGBoost	0.70913	0.03849	5	waist(cm), triglyceride, HDL, fasting blood sugar, systolic
	LightGBM	0.69990	0.01830	5	waist(cm), triglyceride, HDL, fasting blood sugar, systolic
	Logistic Regression	0.67691	0.00407	5	waist(cm), triglyceride, HDL, fasting blood sugar, systolic
	SVM	0.65507	0.00600	5	waist(cm), triglyceride, HDL, fasting blood sugar, systolic

4.8 Comparison to Traditional Clinical Risk Scores

To contextualize our machine learning models against conventional clinical risk stratification methods, we computed a simplified **Framingham cardiovascular risk score**[58] has been shown in Figure 8 using established predictors: age, total cholesterol, HDL cholesterol, systolic blood pressure, smoking status, and diabetes indicators. Participants were subsequently categorized into **low**, **moderate**, or **high** risk groups according to standard Framingham point thresholds. Figure 8 illustrates the distribution of these Framingham risk categories among smokers and non-smokers in our dataset. A substantial proportion of smokers were classified within the moderate or high cardiovascular risk categories, reinforcing the clinical relevance of smoking as a dominant determinant of cardiovascular health. While the Framingham score provides a well-validated baseline for risk estimation, our machine learning models demonstrated the potential to refine these predictions by integrating a broader set of physiological and biochemical variables. This enhanced predictive capacity underscores the ability of data-driven models to complement traditional clinical tools, offering more individualized and nuanced risk assessments for smoking-related health decline.

5 Discussion

This study’s primary contribution is a rigorous, systematic comparison of machine learning approaches for smoking risk assessment rather than algorithmic innovation. While the methods employed (Random Forest, XGBoost, LightGBM) are established techniques, our work advances the field through: (1) comprehensive evaluation across multiple physiological systems rather than isolated endpoints; (2) emphasis on clinical interpretability via SHAP analysis; (3) direct benchmarking against traditional clinical risk scores; and (4) thorough investigation of practical deployment

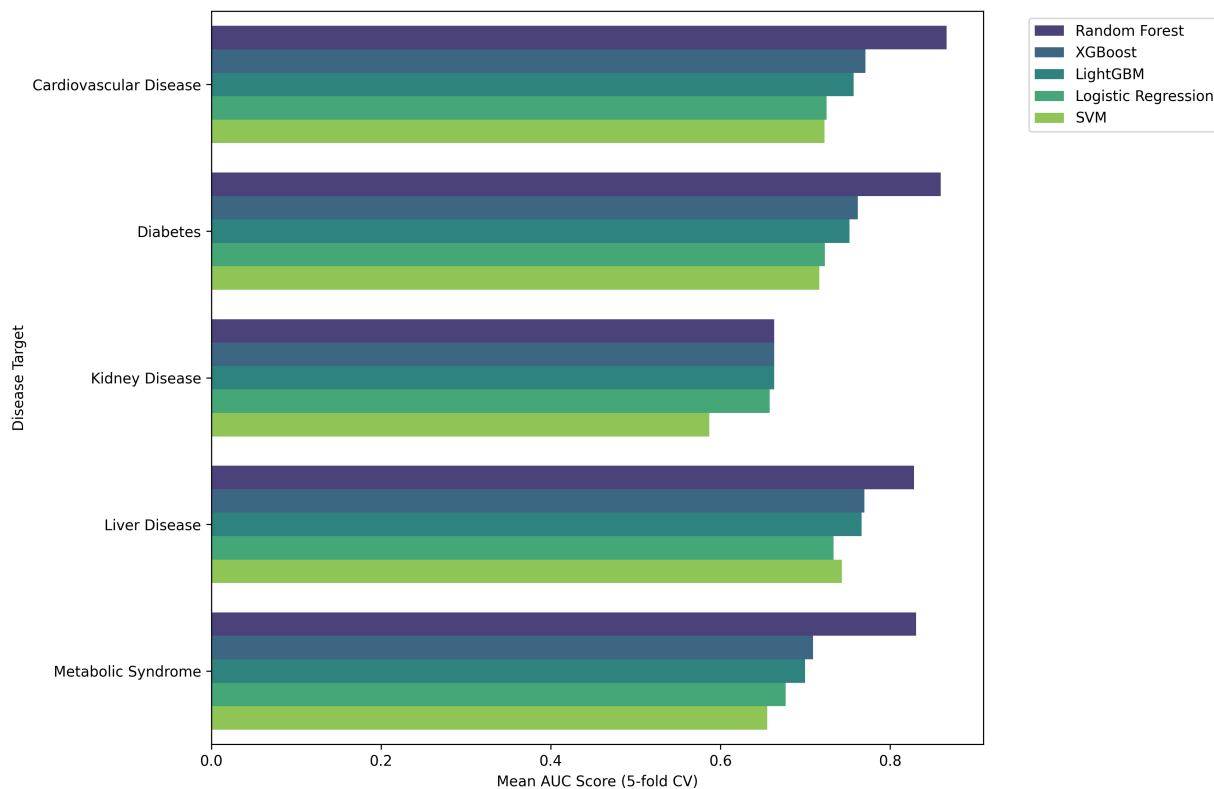


Figure 7: The performance of various machine learning models (Random Forest, XGBoost, LightGBM, Logistic Regression, and SVM) across multiple disease targets, including Cardiovascular Disease, Diabetes, Kidney Disease, Liver Disease, and Headache Syndrome. The metrics illustrate the effectiveness of each model in predicting disease outcomes, emphasizing differences in performance across the disease categories.

considerations including fairness, ethics, and generalizability. This comparative framework provides evidence-based guidance for clinicians and healthcare systems considering the adoption of predictive analytics for smoking-related health assessment.

We acknowledge that no predictive tool is perfect, and model errors can have consequences. False positives might lead to unnecessary testing or increased anxiety, while false negatives could result in missed prevention opportunities. Therefore, we recommend deploying these models in a human-in-the-loop framework, where clinicians validate and contextualize automated predictions before acting on them. Clear communication with patients about model limitations, combined with shared decision-making, will help ensure these tools support rather than replace clinical reasoning. While the research employed widely recognized machine learning techniques, its key contribution lies in the thorough combination of a large-scale health screening dataset (55,691 individuals) with a variety of biomedical and lifestyle factors. In contrast to previous studies that typically concentrate on limited sets of biomarkers or smaller populations, our analysis utilized a broad range of demographic, anthropometric, clinical, and behavioral variables at once, which facilitated a more comprehensive understanding of health patterns associated with smoking. Additionally, we emphasized the interpretability of the model by pairing feature selection (Boruta) with importance ranking, offering clear insights into how each health indicator contributes comparatively. This clarity is especially important in biomedical settings, where trust and transparency are critical for successful clinical implementation. In this manner, the study sets itself apart not through innovative algorithms, but rather through the extent of the dataset, the synthesis of diverse health variables, and a focus on clinically relevant interpretability.

5.1 Limitations and Generalizability

Several limitations warrant consideration. **First**, our dataset originates from a single South Korean health screening program with predominantly urban, ethnically homogeneous participants. Model performance may differ in other ethnic groups due to genetic polymorphisms affecting nicotine metabolism[59] (e.g., CYP2A6 variants)[60] and varying baseline disease prevalence. External validation in diverse populations (European, African, Latino cohorts) is essential

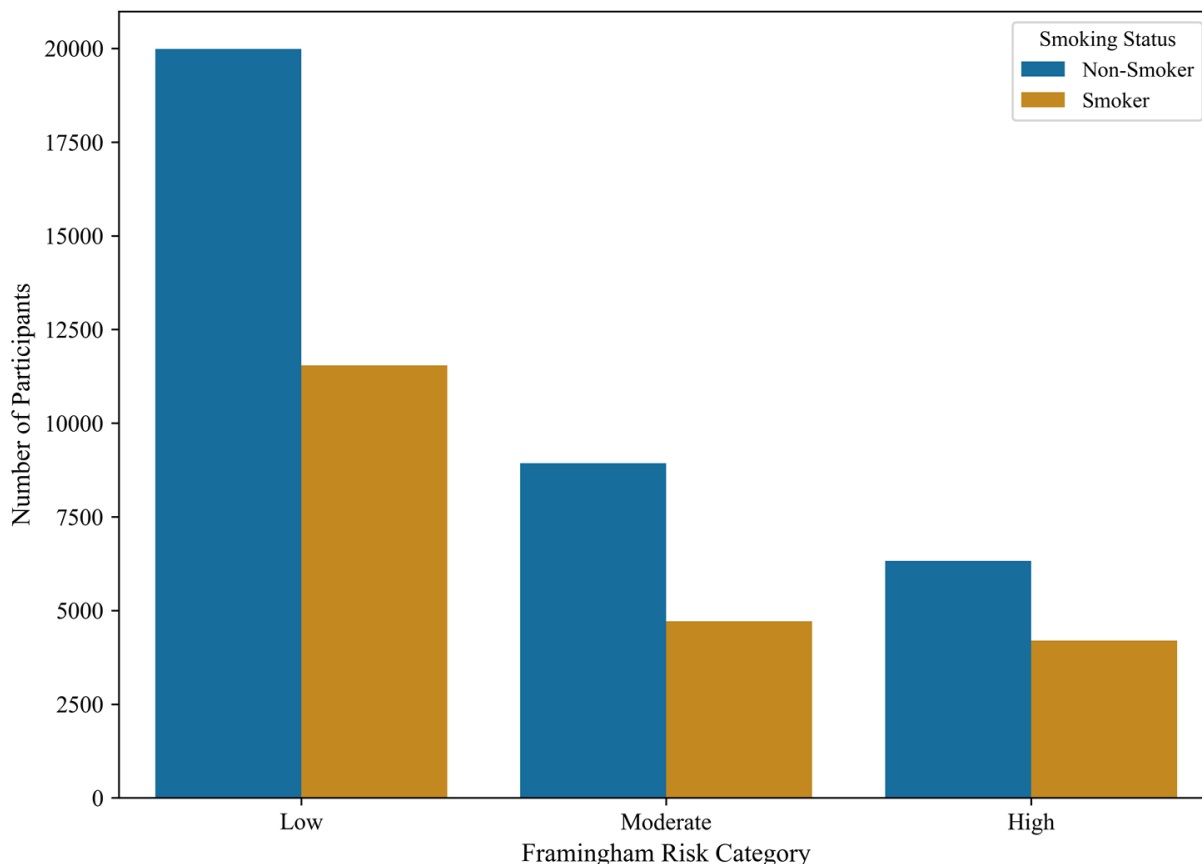


Figure 8: Distribution of Framingham cardiovascular risk categories among smokers and non-smokers. A larger proportion of smokers fall within moderate-to-high risk categories, highlighting smoking’s strong contribution to cardiovascular risk and the potential of machine learning models to improve upon traditional risk assessments.

before clinical deployment. **Second**, the dataset lacks socioeconomic indicators (income, education, occupation), which are known confounders of both smoking behavior and health outcomes. Without controlling for these factors, our model may partially conflate socioeconomic health disparities with smoking-specific effects. **Third**, the cross-sectional design precludes assessment of temporal causality or prediction of future disease outcomes. Longitudinal validation tracking individuals over 5-10 years is needed to confirm that high-risk predictions translate to actual disease incidence. **Fourth**, smoking status relied on self-report, which may underestimate prevalence due to social desirability bias. Biochemical validation (cotinine levels) would strengthen outcome ascertainment[61]. **Future validation priorities include:** (1) multi-site studies in diverse ethnic populations; (2) prospective cohorts with longitudinal follow-up; (3) rural and socioeconomically disadvantaged populations; and (4) integration of detailed smoking history variables (pack-years, cessation attempts).

5.2 Ethical Considerations for Clinical Deployment

Deploying predictive algorithms raises important ethical considerations requiring proactive mitigation. **Managing Prediction Errors:** Our model achieves 86.5% specificity (13.5% false positives) and 80.1% sensitivity (20% false negatives). False positives may cause patient anxiety and unnecessary testing, while false negatives risk delayed intervention. Mitigation strategies include: (1) two-stage screening with clinical confirmation; (2) clear communication that predictions are probabilistic, not definitive; (3) shared decision-making frameworks; and (4) combining algorithmic predictions with routine clinical assessment. **Discrimination Risks:** Predictive risk scores could be misused by insurers or employers for discrimination. Recommended safeguards: (1) restrict access to treating clinicians only; (2) prohibit sharing with third parties absent explicit consent; (3) advocate for legal protections under medical privacy laws. **Algorithmic Fairness:** Our model’s reliance on sex (13.1% of importance) raises equity concerns[62]. While reflecting genuine biological differences, sex-based predictions require: (1) stratified performance reporting; (2)

disparate impact analyses; (3) ensuring adequate accuracy for both sexes, and (4) continuous fairness monitoring across demographic subgroups. **Explainability and Autonomy:** SHAP analysis provides transparency enabling clinicians to verify predictions against domain knowledge. Algorithms must function as decision support tools, not replacements for clinical judgment. Clinicians retain authority to override recommendations, and patients retain the right to opt out of algorithmic assessment. **Implementation Requirements:** (1) comprehensive informed consent; (2) clinician training on model limitations; (3) continuous fairness audits; (4) patient feedback mechanisms; (5) regulatory compliance[63] (FDA, GDPR, HIPAA); and (6) transparent documentation of model limitations and validation status.

6 Conclusion

This study demonstrates that machine learning can do more than just predict smoking-related diseases—it can help us understand them in fundamentally new ways. By combining robust predictive performance with interpretable insights, our models provide a practical tool for clinicians to identify high-risk smokers earlier and intervene more effectively. The consistent superiority of ensemble methods, especially Random Forest, makes a strong case for adopting these approaches in clinical risk assessment tools. The real value lies not just in the algorithms themselves, but in how they reveal the complex interplay of risk factors that conventional statistical methods might miss. As we look to the future, these findings point toward more personalized approaches to smoking cessation and health monitoring. By understanding which specific systems are at risk in individual patients—whether cardiovascular, metabolic, hepatic, or renal—we can tailor interventions that address each smoker’s unique vulnerability profile. This represents an important step toward precision prevention for one of our most significant public health challenges.

Table 13: List of Abbreviations

Abbreviation	Definition
COPD	Chronic Obstructive Pulmonary Disease
AST	Aspartate Aminotransferase (liver enzyme)
ALT	Alanine Aminotransferase (liver enzyme)
Ggt (Gtp)	Gamma-Glutamyl Transferase (liver/biliary marker)
HDL	High-Density Lipoprotein ("good" cholesterol)
LDL	Low-Density Lipoprotein ("bad" cholesterol)
SBP	Systolic Blood Pressure
DBP	Diastolic Blood Pressure
ML	Machine Learning
AUC	Area Under the ROC Curve
ROC	Receiver Operating Characteristic
SHAP	Shapley Additive Explanations (model interpretability method)
PCA	Principal Component Analysis
SMOTE	Synthetic Minority Over-sampling Technique (for class imbalance)
NRSBoundary-SMOTE	Neighborhood Rough Set Boundary SMOTE (advanced resampling)
RF	Random Forest
SVM	Support Vector Machine
LR	Logistic Regression
XGBoost	Extreme Gradient Boosting
LightGBM	Light Gradient Boosting Machine
CV	Cross-Validation
SD	Standard Deviation
BMI	Body Mass Index
F1	F1-Score (harmonic mean of precision/recall)
G-mean	Geometric Mean (of sensitivity/specificity)
CI	Confidence Interval
PPV	Positive Predictive Value
MICE	Multiple Imputation by Chained Equations
MCAR	Missing Completely At Random

Acknowledgment

The authors would like to sincerely thank all well-wishers and supporters who encouraged and inspired this work. The authors declare that there are no conflicts of interest associated with this study.

References

- [1] Charuni TMJ. Narrative review on the spectrum of diseases prevalent among substance-addicted populations and their interconnected health dynamics. *Journal of Science of the University of Kelaniya*. 2024 May;17(1):57-63. Publisher: Sri Lanka Journals Online. Available from: <https://account.josuk.sljol.info/index.php/sljo-j-jsuksl/article/view/8107>.
- [2] Sakthisankaran SM, Sakthipriya D, Swamivelmanickam M. Health Risks Associated with Tobacco Consumption in Humans: An Overview. *Journal of Drug Delivery and Therapeutics*. 2024 May;14(5):163-73. Publisher: Society of Pharmaceutical Tecnocrats. Available from: <https://jddtonline.info/index.php/jddt/article/view/6523>.
- [3] Lu W, Aarsand R, Schotte K, Han J, Lebedeva E, Tsoy E, et al. Tobacco and COPD: presenting the World Health Organization (WHO) Tobacco Knowledge Summary. *Respiratory Research*. 2024 Sep;25(1). Publisher: Springer Science and Business Media LLC. Available from: <https://respiratory-research.biomedcentral.com/articles/10.1186/s12931-024-02961-5>.
- [4] Kotlyarov S. The Role of Smoking in the Mechanisms of Development of Chronic Obstructive Pulmonary Disease and Atherosclerosis. *International Journal of Molecular Sciences*. 2023 May;24(10):8725. Publisher: MDPI AG. Available from: <https://www.mdpi.com/1422-0067/24/10/8725>.
- [5] Lim SY, Ulaganathan V, Gunasekaran B, Salvamani S, Tiong YL, Muhammad Royani SM, et al. Chronic obstructive pulmonary disease: Signs and symptoms, diagnosis, treatments, lifestyle risk factors and management. *Life Sciences, Medicine and Biomedicine*. 2024 Mar;8(1). Publisher: Biome Scientia Sdn Bhd. Available from: <https://biomescientia.com/index.php/lmb/article/view/123>.
- [6] Kushner D. Mild Traumatic Brain Injury: Toward Understanding Manifestations and Treatment. *Archives of Internal Medicine*. 1998 Aug;158(15):1617. Publisher: American Medical Association (AMA). Available from: <http://archinte.jamanetwork.com/article.aspx?doi=10.1001/archinte.158.15.1617>.
- [7] Glynn T, Seffrin JR, Brawley OW, Grey N, Ross H. The Globalization of Tobacco Use: 21 Challenges For The 21st Century. *CA: A Cancer Journal for Clinicians*. 2010 Jan;60(1):50-61. Publisher: Wiley. Available from: <http://doi.wiley.com/10.3322/caac.20052>.
- [8] Rajkomar A, Dean J, Kohane I. Machine learning in medicine. *New England Journal of Medicine*. 2019;380(14):1347-58.
- [9] Chakma V, Xiaolin J, Cao H, Feng X, Xiaodong J, Haiyan P, et al. CardioForest: An Explainable Ensemble Learning Model for Automatic Wide QRS Complex Tachycardia Diagnosis from ECG. *arXiv preprint arXiv:250925804*. 2025.
- [10] Nabanita Ghosh KS. A Review on the Recent Advancements in Machine Learning-Assisted Tobacco Research. Publisher: NIPES PUBLICATIONS. 2024 May. Available from: <https://zenodo.org/doi/10.5281/zenodo.11223324>.
- [11] Van Spall HGC, Bastien A, Gersh B, Greenberg B, Mohebi R, Min J, et al. The role of early-phase trials and real-world evidence in drug development. *Nature Cardiovascular Research*. 2024 Feb;3(2):110-7. Publisher: Springer Science and Business Media LLC. Available from: <https://www.nature.com/articles/s44161-024-00420-4>.
- [12] Liu J, Barrett JS, Leonardi ET, Lee L, Roychoudhury S, Chen Y, et al. Natural History and Real-World Data in Rare Diseases: Applications, Limitations, and Future Perspectives. *The Journal of Clinical Pharmacology*. 2022 Dec;62(S2). Publisher: Wiley. Available from: <https://accp1.onlinelibrary.wiley.com/doi/10.1002/jcph.2134>.
- [13] Vaskar C, Misbahul A, Abdur R, Al Mahmud S, Raju M, Mustavi R. From Margins to Mainstream (M2M): Can Artificial Intelligence (AI) Reshape Governance for Chittagong Hill Tracts Indigenous Communities? *Applied Sciences*;3(1):166-78.
- [14] Xu H, Peng X, Peng Z, Wang R, Zhou R, Fu L. Construction and SHAP interpretability analysis of a risk prediction model for feeding intolerance in preterm newborns based on machine learning. *BMC Medical Informatics and Decision Making*. 2024 Nov;24(1). Publisher: Springer Science and Business Media LLC. Available from: <https://bmcmidinformedecismak.biomedcentral.com/articles/10.1186/s12911-024-02751-5>.
- [15] Antonini AS, Tanzola J, Asiain L, Ferracutti GR, Castro SM, Bjerg EA, et al. Machine Learning model interpretability using SHAP values: Application to Igneous Rock Classification task. *Applied Computing and Geosciences*. 2024 Sep;23:100178. Publisher: Elsevier BV. Available from: <https://linkinghub.elsevier.com/retrieve/pii/S2590197424000259>.

- [16] Aishwarya S, Siddalingaswamy PC, Chadaga K. Explainable artificial intelligence driven insights into smoking prediction using machine learning and clinical parameters. *Scientific Reports*. 2025 Jul;15(1). Publisher: Springer Science and Business Media LLC. Available from: <https://www.nature.com/articles/s41598-025-09409-w>.
- [17] Research Department, Mangalore University, Mangalore, India, Pasupuleti DN. PREDICTIVE MODELING FOR SMOKING STATUS AND LUNG CANCER RISK CLASSIFICATION: A MACHINE LEARNING APPROACH. *international journal of advanced research in computer science*. 2025 Jun;16(03):75-83. Publisher: IJARCS International Journal of Advanced Research in Computer Science. Available from: <https://ijarcs.info/index.php/Ijarcs/article/view/7270>.
- [18] Vaskar Chakma, MD Jaheid Hasan Nerab, Abdur Rouf, Abu Sayed, Hossem MD Saim, and Md Nournabi Khan. Machine Learning Techniques for Predicting SRHD: Smoking-Related Health Decline. *Authorea Preprints*, Authorea, 2025.
- [19] Davagdorj K, Lee JS, Pham VH, Ryu KH. A Comparative Analysis of Machine Learning Methods for Class Imbalance in a Smoking Cessation Intervention. *Applied Sciences*. 2020 May;10(9):3307. Publisher: MDPI AG. Available from: <https://www.mdpi.com/2076-3417/10/9/3307>.
- [20] Negewo NA, Gibson PG, McDonald VM. COPD and its comorbidities: Impact, measurement and mechanisms. *Respirology*. 2015 Nov;20(8):1160-71. Publisher: Wiley. Available from: <https://onlinelibrary.wiley.com/doi/10.1111/resp.12642>.
- [21] Carrasco-Zanini J, Pietzner M, Koprulu M, Wheeler E, Kerrison ND, Wareham NJ, et al. Proteomic prediction of diverse incident diseases: a machine learning-guided biomarker discovery study using data from a prospective cohort study. *The Lancet Digital Health*. 2024 Jul;6(7):e470-9. Publisher: Elsevier BV. Available from: <https://linkinghub.elsevier.com/retrieve/pii/S2589750024000876>.
- [22] Rosenbacke R, Melhus A, McKee M, Stuckler D. How Explainable Artificial Intelligence Can Increase or Decrease Clinicians' Trust in AI Applications in Health Care: Systematic Review. *JMIR AI*. 2024 Oct;3:e53207. Publisher: JMIR Publications Inc. Available from: <https://ai.jmir.org/2024/1/e53207>.
- [23] Le TTT, Issabakhsh M, Li Y, María Sánchez-Romero L, Tan J, Meza R, et al. Are the Relevant Risk Factors Being Adequately Captured in Empirical Studies of Smoking Initiation? A Machine Learning Analysis Based on the Population Assessment of Tobacco and Health Study. *Nicotine and Tobacco Research*. 2023 Jul;25(8):1481-8. Publisher: Oxford University Press (OUP). Available from: <https://academic.oup.com/ntr/article/25/8/1481/7143558>.
- [24] Aydin HE, Iban MC. Predicting and analyzing flood susceptibility using boosting-based ensemble machine learning algorithms with SHapley Additive exPlanations. *Natural Hazards*. 2023;116(3):2957-91.
- [25] Lu JK, Wang W, Mahadzir MDA, Poganik JR, Moqri M, Herzog C, et al. Digital biomarkers of ageing for monitoring physiological systems in community-dwelling adults. *The Lancet Healthy Longevity*. 2025.
- [26] Biomarkers Definitions Working Group. Biomarkers and surrogate endpoints: preferred definitions and conceptual framework. *Clinical Pharmacology & Therapeutics*. 2001;69(3):89-95.
- [27] Schober P, Mascha EJ, Vetter TR. Statistics from A (agreement) to Z (z score): a guide to interpreting common measures of association, agreement, diagnostic accuracy, effect size, heterogeneity, and reliability in medical research. *Anesthesia & Analgesia*. 2021;133(6):1633-41.
- [28] Raju VG, Lakshmi KP, Jain VM, Kalidindi A, Padma V. Study the influence of normalization/transformation process on the accuracy of supervised classification. In: 2020 third international conference on smart systems and inventive technology (icssit). IEEE; 2020. p. 729-35.
- [29] Xie J, Wang M, Xu S, Huang Z, Grant PW. The unsupervised feature selection algorithms based on standard deviation and cosine similarity for genomic data analysis. *Frontiers in Genetics*. 2021;12:684100.
- [30] Kursu MB, Rudnicki WR. Feature selection with the Boruta package. *Journal of Statistical Software*. 2010;36:1-13.
- [31] Borah K, Das HS, Seth S, Mallick K, Rahaman Z, Mallik S. A review on advancements in feature selection and feature extraction for high-dimensional NGS data analysis. *Functional & Integrative Genomics*. 2024 Oct;24(5). Publisher: Springer Science and Business Media LLC. Available from: <https://link.springer.com/10.1007/s10142-024-01415-x>.
- [32] Ding J, Du J, Wang H, Xiao S. A novel two-stage feature selection method based on random forest and improved genetic algorithm for enhancing classification in machine learning. *Scientific Reports*. 2025 May;15(1). Publisher: Springer Science and Business Media LLC. Available from: <https://www.nature.com/articles/s41598-025-01761-1>.

- [33] Kyriazos T, Poga M. Dealing with Multicollinearity in Factor Analysis: The Problem, Detections, and Solutions. *Open Journal of Statistics*. 2023;13(03):404-24. Publisher: Scientific Research Publishing, Inc. Available from: <https://www.scirp.org/journal/doi.aspx?doi=10.4236/ojs.2023.133020>.
- [34] Carvalho M, Pinho AJ, Brás S. Resampling approaches to handle class imbalance: a review from a data perspective. *Journal of Big Data*. 2025;12(1):71.
- [35] Kumar C, Khan PS, Srinivas M, Jha SK, Prakash S, Rathore RS. Ensemble Learning for Software Requirement-Risk Assessment: A Comparative Study of Bagging and Boosting Approaches. *Future Internet*. 2025;17(9):387.
- [36] Ahmadi A, Sharif SS, Banad YM. A comparative study of sampling methods with cross-validation in the fedhome framework. *IEEE Transactions on Parallel and Distributed Systems*. 2025.
- [37] Chawla NV, Bowyer KW, Hall LO, Kegelmeyer WP. SMOTE: Synthetic minority over-sampling technique. *Journal of Artificial Intelligence Research*. 2002;16:321-57.
- [38] He H, Bai Y, Garcia EA, Li S. ADASYN: Adaptive synthetic sampling approach for imbalanced learning. In: 2008 IEEE International Joint Conference on Neural Networks (IEEE World Congress on Computational Intelligence). IEEE; 2008. p. 1322-8.
- [39] Calle EE, Rodriguez C, Walker-Thurmond K, Thun MJ. Overweight, obesity, and mortality from cancer in a prospectively studied cohort of US adults. *New England Journal of Medicine*. 2003;348(17):1625-38.
- [40] Lavie CJ, Tutor AW, Carbone S. Is the obesity paradox real? *Canadian Journal of Cardiology*. 2025.
- [41] Maier M, Bartoš F, Quintana DS, Dablander F, den Bergh Dv, Marsman M, et al. Model-averaged Bayesian t tests. *Psychonomic Bulletin & Review*. 2025;32(3):1007-31.
- [42] Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, et al. Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research*. 2011;12:2825-30.
- [43] Chen T, Guestrin C. XGBoost: A Scalable Tree Boosting System. In: *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD '16*. New York, NY, USA: ACM; 2016. p. 785-94. Available from: <http://doi.acm.org/10.1145/2939672.2939785>.
- [44] Ke G, Meng Q, Finley T, Wang T, Chen W, Ma W, et al. LightGBM: A Highly Efficient Gradient Boosting Decision Tree. In: Guyon I, Luxburg UV, Bengio S, Wallach H, Fergus R, Vishwanathan S, et al., editors. *Advances in Neural Information Processing Systems*. vol. 30. Curran Associates, Inc.; 2017. p. 3146-54.
- [45] Lundberg SM, Lee SI. A Unified Approach to Interpreting Model Predictions. In: Guyon I, Luxburg UV, Bengio S, Wallach H, Fergus R, Vishwanathan S, et al., editors. *Advances in Neural Information Processing Systems*. vol. 30. Curran Associates, Inc.; 2017. p. 4765-74.
- [46] McKinney W. van der Walt S, Millman J, editors. *Data Structures for Statistical Computing in Python*; 2010.
- [47] Harris CR, Millman KJ, van der Walt SJ, Gommers R, Virtanen P, Cournapeau D, et al. Array programming with NumPy. *Nature*. 2020;585(7825):357-62.
- [48] Virtanen P, Gommers R, Oliphant TE, Haberland M, Reddy T, Cournapeau D, et al. SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python. *Nature Methods*. 2020;17:261-72.
- [49] Hunter JD. Matplotlib: A 2D graphics environment. *Computing in Science & Engineering*. 2007;9(3):90-5.
- [50] Waskom ML. seaborn: statistical data visualization. *Journal of Open Source Software*. 2021;6(60):3021.
- [51] Yaqoob A, Verma NK, Mir MA, Tejani GG, Eisa NHB, Mamoun Hussien Osman H, et al. SGA-Driven feature selection and random forest classification for enhanced breast cancer diagnosis: A comparative study. *Scientific Reports*. 2025;15(1):10944.
- [52] Shapley LS. A value for n-person games. *Contributions to the Theory of Games*. 1953;2(28):307-17.
- [53] Huxley RR, Woodward M. Cigarette smoking as a risk factor for coronary heart disease in women compared with men: a systematic review and meta-analysis of prospective cohort studies. *The Lancet*. 2011;378(9799):1297-305.
- [54] Zein CO, Unalp A, Colvin R, Liu YC, McCullough AJ. Smoking and severity of hepatic fibrosis in nonalcoholic fatty liver disease. *Journal of Hepatology*. 2011;54(4):753-9.
- [55] Nordenberg D, Yip R, Binkin NJ. The prevalence of polycythemia among smokers. *Annals of Internal Medicine*. 1990;112(7):493-9.
- [56] Bergström J. Tobacco smoking and periodontal disease. *Odontology*. 2000;92(1):1-8.
- [57] Jolliffe IT, Cadima J. Principal component analysis: a review and recent developments. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*. 2016;374(2065):20150202.

- [58] Wilson PW, D'Agostino RB, Levy D, Belanger AM, Silbershatz H, Kannel WB. Prediction of coronary heart disease using risk factor categories. *Circulation*. 1998;97(18):1837-47.
- [59] Tanner JA, Novalen M, Jatlow P, Huestis MA, Murphy SE, Kaprio J, et al. Nicotine metabolite ratio (3-hydroxycotinine/cotinine) in plasma and urine by different analytical methods and laboratories: implications for clinical implementation. *Cancer Epidemiology, Biomarkers & Prevention*. 2017;26(8):1153-63.
- [60] Benowitz NL, Hukkanen J, Jacob III P. Nicotine chemistry, metabolism, kinetics and biomarkers. *Handbook of Experimental Pharmacology*. 2009;192:29-60.
- [61] Benowitz NL. Biomarkers of environmental tobacco smoke exposure. *Environmental Health Perspectives*. 1999;107(suppl 2):349-55.
- [62] Rajkomar A, Hardt M, Howell MD, Corrado G, Chin MH. Ensuring fairness in machine learning to advance health equity. *Annals of Internal Medicine*. 2018;169(12):866-72.
- [63] Price WN, Cohen IG. Privacy in the age of medical big data. *Nature Medicine*. 2019;25(1):37-43.