

Multi-Organ Impact of Smoking Using Explainable Artificial Intelligence: Prediction, Validation, and Organ Impact Pattern

SOMASHEKAR V¹, 

1. *Aeronautical Engineering, Nitte (Deemed to be University) Nitte Meenakshi Institute of Technology (NMIT), Bangalore, IND*

Received: May 13, 2026 | Review began: May 28, 2026 | Review ended: June 11, 2026 | Published: June 16, 2026

© Copyright 2026

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

Abstract

This study aims to investigate the multi-organ impact of smoking using an Explainable Artificial Intelligence (XAI)-based framework by analyzing liver, kidney, and metabolic health indicators from the Kaggle *Body Signal of Smoking* dataset containing 55,692 patient records and 27 clinical features. The objective is to identify smoking-associated biomarkers and infer a potential organ impact pattern using machine learning and explainability techniques. Data preprocessing included removal of missing values, label encoding of categorical variables, and separation of features (X) and target variable (y = smoking). The dataset was split into 80% training and 20% testing sets. Three classification models - Random Forest, XGBoost, and Logistic Regression-were trained, with hyperparameter tuning performed using GridSearchCV on Random Forest. Model validation included Accuracy, F1-score, Receiver Operating Characteristic Area Under the Curve (ROC-AUC), 5-fold cross-validation, independent sample *t*-tests, and SHapley Additive exPlanations (SHAP) for interpretability. Results showed that Random Forest achieved the best predictive performance with strong classification accuracy and ROC-AUC. Statistical testing revealed significant differences ($P < 0.05$) in Gamma-Glutamyl Transferase (GTP), Alanine Aminotransferase (ALT), serum creatinine, hemoglobin, and triglycerides between smokers and non-smokers. SHAP and feature importance analysis consistently identified GTP, ALT, and serum creatinine as major predictors. This study suggests a potential organ impact pattern as Liver > Kidney > Metabolic System based on statistical analysis and XAI techniques.

Categories: AI applications, Generative AI for healthcare, Health Informatics

Keywords: smoking prediction, explainable artificial intelligence (xai), random forest, shap analysis, multi-organ damage

How to cite this article:

V S (June 16, 2026) Multi-Organ Impact of Smoking Using Explainable Artificial Intelligence: Prediction, Validation, and Organ Impact Pattern. Cureus J Comput Sci 3 : es44389-026-00121-y. DOI <https://doi.org/10.7759/s44389-026-00121-y>

Introduction

Tobacco consumption remains a major global public health challenge, responsible for approximately 8 million annual deaths, with 7 million resulting from direct use and 1.2 million from exposure to second-hand smoke [1,2]. As the leading preventable cause of morbidity and premature mortality, smoking is conventionally associated with respiratory diseases and lung cancer; however, contemporary research emphasizes its systemic, multi-organ impact [3,4]. The complex mixture of over 7,000 chemicals in tobacco smoke triggers a cascade of pathological processes, including chronic inflammation, oxidative stress, and DNA damage, which compromise virtually every biological system in the human body [5,6].

While the pulmonary and cardiovascular consequences of smoking are well-documented, its detrimental effects on the liver, kidneys, and metabolic systems are often characterized by a subclinical, insidious progression [7]. In the liver, smoking-induced oxidative stress leads to direct hepatotoxicity and the accumulation of toxic metabolites, often manifesting through elevated enzymes such as gamma-glutamyl transferase (GTP/GGT) and alanine aminotransferase (ALT). Similarly, the renal system suffers from nicotine-mediated vasoconstriction and direct tubular damage, frequently identified by shifts in serum creatinine levels. Beyond specific organs, smoking acts as a potent metabolic disruptor, impairing insulin sensitivity, altering lipid metabolism-evidenced by increased triglycerides and reduced high-density lipoprotein (HDL)-and increasing the risk of Type 2 diabetes [8].

Despite the severity of these multi-systemic threats, traditional clinical diagnostic approaches often rely on fragmented, static data, which may fail to detect early warning signs before irreversible damage occurs [9]. Artificial Intelligence (AI) and Machine Learning (ML) have emerged as transformative tools in healthcare, capable of analyzing high-dimensional datasets to identify hidden biological signatures of smoking-related decline [10]. However, the clinical adoption of advanced ML models is frequently hindered by their "black-box" nature, which lacks the transparency required for medical practitioners to trust and validate algorithmic decisions [11].

AI has been widely applied across various aspects of tobacco control. Many studies have leveraged AI to analyze hospital records for identifying and classifying smoking history [12-19], forecasting the impact of smoking following a cancer diagnosis [20-25], quantifying smoke exposure levels [26], estimating cardiovascular disease risk [27,28], detecting cancers linked to tobacco use [29-31], and diagnosing diseases affecting tobacco plants [32]. In healthcare and social service settings, AI has also contributed by predicting smoking relapse and nicotine dependence [33-36], categorizing patterns of social smoking [37], identifying environments that may trigger smoking urges [38], analyzing emissions from e-cigarettes [39], and evaluating exposure to secondhand and thirdhand smoke [40]. Despite these advancements, limited work has focused on integrating explainable AI (XAI) with multi-organ analysis to quantify and interpret the systemic effects of smoking.

In this study, an XAI-based framework is proposed to analyze the multi-organ impact of smoking using the publicly available Kaggle Body Signal of Smoking dataset containing 55,692 patient records and 27 clinical features. The study focuses on three major physiological systems: liver, kidney, and metabolic health. Random Forest, XGBoost, and Logistic Regression models are developed to predict smoking status, while statistical validation, feature importance analysis, correlation heatmaps, boxplots, distribution plots, and SHapley Additive exPlanations (SHAP) explainability are used to identify the most affected biomarkers. The primary objective is to analyze and infer a potential organ impact pattern using statistical and explainable ML techniques.

The major contribution of this work is the inference of a potential organ impact pattern (Liver > Kidney > Metabolic System) based on statistical and XAI evidence, where liver-related biomarkers such as GTP and ALT show the strongest influence, followed by kidney marker serum creatinine and metabolic indicators such as triglycerides and fasting blood sugar. This study provides a transparent, clinically interpretable, and robust framework for smoking-related health risk assessment and supports early diagnosis, preventive screening, and explainable healthcare decision-making.

How to cite this article:

Materials And Methods

This study proposes an XAI-based methodology to investigate the multi-organ impact of smoking using clinical body signal data obtained from the Kaggle dataset Body Signal of Smoking (smoking.csv) [41], which contains 55,692 patient records and 27 features. The main objective is to quantify how smoking affects three major physiological systems - liver, kidney, and metabolic health - and to rank the severity of organ damage using ML models supported by explainability techniques. The target variable in the dataset is smoking status, which classifies individuals as smokers or non-smokers. The dataset includes various demographic, anthropometric, and biochemical parameters such as age, gender, hemoglobin, serum creatinine, liver enzymes [AST (Aspartate Aminotransferase), ALT, GGT], cholesterol profile, blood pressure, fasting blood sugar, oral examination results, and tartar condition. These attributes provide strong clinical relevance for understanding the physiological effects of smoking.

All preprocessing steps and model training procedures were implemented using Python (version 3.x) with standard libraries including Pandas, NumPy, Scikit-learn, XGBoost, and SHAP.

The first stage of the methodology involves data preprocessing to improve data quality and model performance. Missing values were removed using the `dropna()` function to eliminate incomplete records and maintain consistency in analysis. Categorical variables such as gender, oral examination status, and tartar condition were converted into numerical form using Label Encoding, making them suitable for ML algorithms. After preprocessing, the dataset was divided into independent and dependent variables, where all clinical features were assigned to X and the smoking label was assigned to y. This structure enables supervised classification for predicting smoking behavior and identifying organ-related health impacts.

To ensure reliable model evaluation, the dataset was divided into training and testing sets using an 80:20 ratio through the `train_test_split()` function from Scikit-learn. In this approach, 80% of the data was used for training the models, while the remaining 20% was reserved for testing and performance evaluation. This step helps reduce overfitting and ensures that the developed models perform well on unseen data.

Three ML models were developed and compared in this study: Random Forest Classifier, XGBoost Classifier, and Logistic Regression. Random Forest was selected as the primary model because of its strong performance on structured healthcare datasets, its ability to handle nonlinear relationships, and its built-in feature importance mechanism. It is particularly effective in identifying interactions among biomarkers related to liver, kidney, and metabolic systems. XGBoost was used as an advanced boosting algorithm known for high predictive accuracy and robustness in classification tasks. Logistic Regression was included as a baseline model to provide statistical comparison and interpretability against ensemble-based approaches. Using multiple models improves robustness and supports stronger scientific validation.

To further improve model performance, hyperparameter tuning was performed using GridSearchCV, mainly on the Random Forest model. Parameters such as the number of trees (`n_estimators`), maximum tree depth (`max_depth`), minimum samples required for splitting, and minimum samples per leaf were optimized using grid-based search. This process ensures that the model achieves the best balance between bias and variance, resulting in improved predictive capability and generalization.

A comprehensive validation strategy was adopted to ensure the reliability of the results. Model performance was evaluated using Accuracy, F1-score, and Receiver Operating Characteristic Area Under the Curve (ROC-AUC) score. Among these, ROC-AUC was considered the most important metric because it measures the model's ability to distinguish smokers from non-smokers across all classification thresholds. A higher ROC-AUC indicates better predictive strength and classifier reliability.

How to cite this article:

In addition to performance metrics, 5-Fold Cross-Validation was performed using `cross_val_score()` to assess model stability and generalizability. The dataset was divided into five equal parts, where the model was trained and tested five times using different combinations of training and validation sets. The mean and standard deviation of the scores were reported to ensure that the results were not dependent on a single train-test split.

To improve reproducibility and robustness, cross-validation was performed with `cv = 5` and `shuffle = True`, ensuring that the data was randomly distributed across folds while maintaining consistency through a fixed `random_state`.

To establish medical significance, statistical validation was conducted using independent sample *t*-tests (`ttest_ind`) between smokers and non-smokers. Important biomarkers such as cholesterol, liver enzymes (ALT, AST, GGT), serum creatinine, and fasting blood sugar were selected for comparison. These tests help determine whether smoking causes statistically significant physiological differences across liver, kidney, and metabolic systems, thereby strengthening the clinical relevance of the findings.

Explainability validation was performed using SHAP with TreeExplainer on the Random Forest model. SHAP helps identify the contribution of each feature toward smoking prediction and provides interpretable insights into how specific biomarkers influence classification outcomes. SHAP summary plots were generated to visualize feature importance, positive and negative feature effects, and organ-specific biomarker impact. These results were compared with traditional feature importance obtained from Random Forest and XGBoost models. Agreement between SHAP values and feature importance improves transparency, trust, and clinical interpretability.

To ensure that the proposed organ impact pattern is derived in a quantitative and reproducible manner, an aggregated scoring approach was adopted by combining multiple sources of evidence, including statistical significance, feature importance, and SHAP explainability.

First, key biomarkers were grouped based on organ systems: liver (ALT, AST, GTP), kidney (serum creatinine), and metabolic system (cholesterol, triglycerides, fasting blood sugar). For each feature, three quantitative measures were considered:

- (i) statistical significance (*P*-values from *t*-tests),
- (ii) feature importance scores obtained from Random Forest and XGBoost models, and
- (iii) mean absolute SHAP values.

These measures were normalized to ensure comparability across different scales. An aggregated Organ Impact Score was then computed as the average of normalized contributions from all features within each organ group:

$$\text{Organ Score} = (\text{Normalized Statistical Score} + \text{Normalized Feature Importance} + \text{Normalized SHAP Value})/N$$

where *N* represents the number of features within each organ group.

The organ systems were then ranked based on their aggregated scores. This approach ensures that the ranking is not dependent on a single method but is supported by multiple complementary analyses, making it more robust and reproducible.

To support interpretation and result presentation, several visualizations were generated, including boxplots comparing smokers and non-smokers for liver, kidney, and metabolic biomarkers, correlation heatmaps for identifying relationships among features, feature importance plots from Random Forest and XGBoost, SHAP summary plots for explainable AI interpretation, and distribution plots for understanding variable behavior. These graphical analyses improve the clarity of clinical insights and strengthen the analytical presentation of the study.

The proposed organ impact pattern was derived using aggregated organ impact scores based on statistical testing, feature importance analysis, and SHAP explainability.

How to cite this article:

It is important to note that the organ-level ranking proposed in this study is derived from feature-level analysis, including statistical testing, feature importance, and SHAP explainability. Explicit organ-level experimental validation using feature-group ablation or Leave-One-Group-Out (LOGO) analysis was not conducted in this study and is considered as future work.

Model configuration and implementation details

To ensure experimental transparency and reproducibility, detailed model configurations and parameter settings are provided. All models were implemented using Python with the Scikit-learn and XGBoost libraries.

The Random Forest Classifier was configured with the following parameters: number of trees (`n_estimators = 100`), maximum depth (`max_depth = None`), minimum samples split (`min_samples_split = 2`), minimum samples per leaf (`min_samples_leaf = 1`), and `random_state = 42` for reproducibility.

The XGBoost Classifier was implemented using the XGBoost library with parameters including learning rate (`learning_rate = 0.1`), number of estimators (`n_estimators = 100`), maximum tree depth (`max_depth = 6`), subsample ratio (`subsample = 0.8`), and `random_state = 42`.

Logistic Regression was used as a baseline model with default parameters and solver set to "lbfgs," with `max_iter = 1000` to ensure convergence.

Hyperparameter tuning for the Random Forest model was performed using GridSearchCV with 5-fold cross-validation, exploring parameters such as `n_estimators`, `max_depth`, and `min_samples_split`.

All experiments were conducted using a train-test split of 80:20, and model evaluation was performed using accuracy, F1-score, and ROC-AUC metrics. The use of a fixed `random_state` ensures that results are reproducible.

Dataset description and ethical considerations

The study utilized the publicly available Body Signal of Smoking dataset obtained from the Kaggle repository. The dataset contains 55,692 records and 27 variables comprising demographic, anthropometric, physiological, laboratory, and behavioral health indicators. The dataset was originally developed to investigate the presence or absence of smoking using biological health signals.

The available variables include age, sex, height, weight, waist circumference, blood pressure measurements, fasting blood glucose, lipid profile indicators, liver function markers, renal biomarkers, hearing status, dental health indicators, and smoking status. These variables provide a broad representation of smoking-related physiological characteristics and associated health conditions.

The dataset was accessed in a fully anonymized form through the public Kaggle platform. No personally identifiable information was available to the authors, and no direct interaction with human participants occurred. Therefore, this study involved secondary analysis of publicly available de-identified data. As the dataset was publicly available and fully anonymized, no additional ethical approval was required for this secondary data analysis. Information regarding the original participant recruitment procedures and ethics approvals was not available in the public repository documentation.

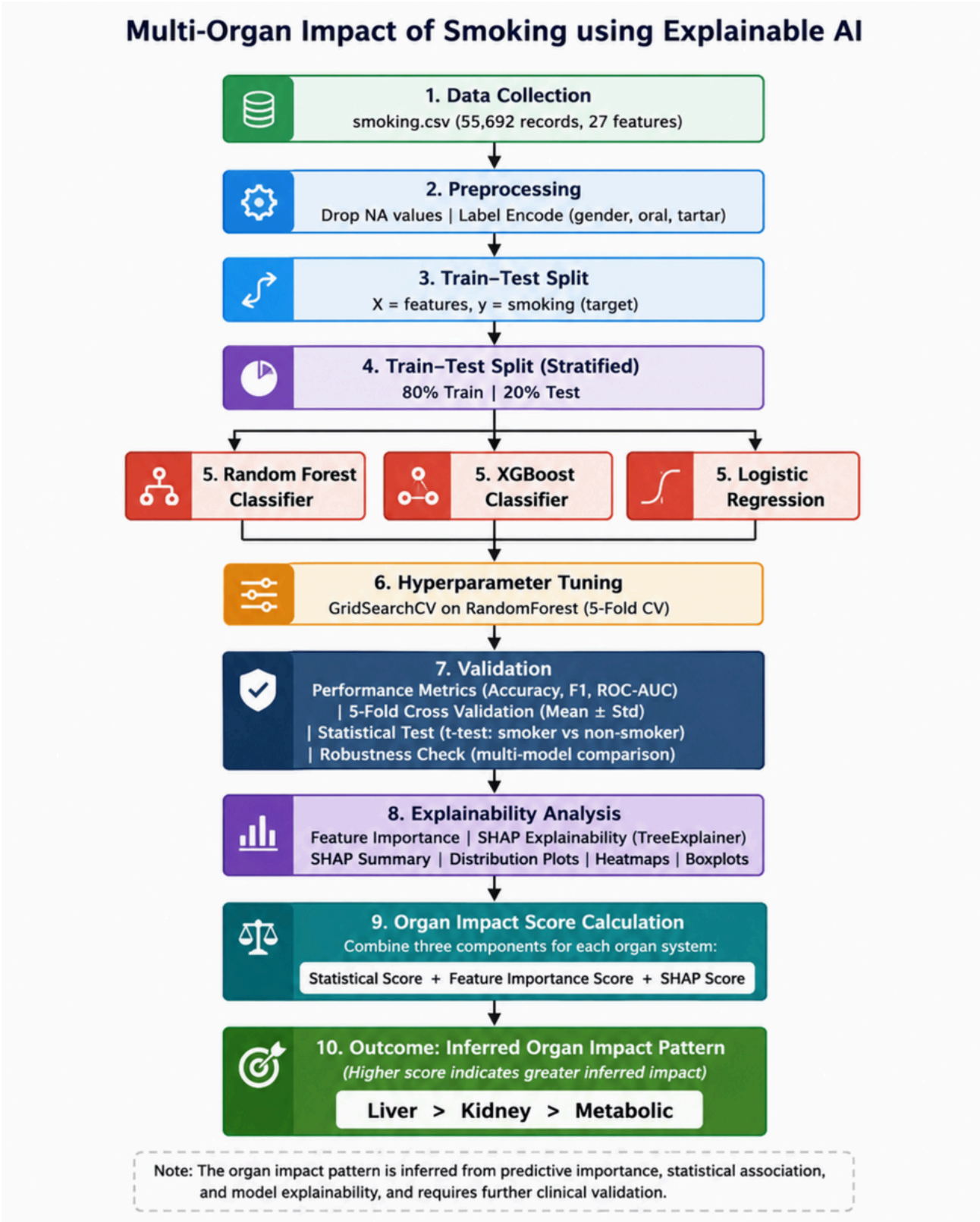


FIGURE 1: Research methodology for multi-organ impact of smoking using XAI.

XAI, Explainable Artificial Intelligence; SHAP, SHapley Additive exPlanations; ROC-AUC, Receiver Operator Characteristic Area Under the Curve

How to cite this article:

V S (June 16, 2026) Multi-Organ Impact of Smoking Using Explainable Artificial Intelligence: Prediction, Validation, and Organ Impact Pattern. Cureus J Comput Sci 3 : es44389-026-00121-y. DOI <https://doi.org/10.7759/s44389-026-00121-y>

Results And Discussion

Experimental setup

The experiments were conducted using the Kaggle Body Signal of Smoking dataset consisting of 55,692 records and 27 clinical features. Data preprocessing included removal of missing values and label encoding of categorical variables. The dataset was divided into training and testing sets using an 80:20 ratio. Random Forest, XGBoost, and Logistic Regression models were implemented using Python and Scikit-learn/XGBoost libraries. Hyperparameter optimization was performed using GridSearchCV with 5-fold cross-validation. Model performance was evaluated using Accuracy, F1-score, and ROC-AUC metrics. Statistical significance between smokers and non-smokers was assessed using independent sample *t*-tests, while SHAP explainability was used to interpret feature contributions.

Figure 2 presents the boxplot comparison of three important clinical biomarkers - ALT, GTPs, and serum creatinine - between smokers and non-smokers. These biomarkers were selected because they represent the functional status of the liver, kidney, and metabolic health, respectively, and are strongly associated with smoking-related physiological damage.

The first boxplot shows the distribution of ALT levels, where smokers (1) exhibit slightly higher median values and a wider spread compared to non-smokers (0). Additionally, smokers show a greater number of extreme outliers, indicating abnormal liver enzyme activity. ALT is a major liver enzyme, and elevated ALT levels suggest liver cell inflammation or damage. This indicates that smoking may contribute to liver stress and hepatic dysfunction.

The second boxplot represents GTP levels, which show a more significant difference between smokers and non-smokers. Smokers have a noticeably higher median GTP value and a much broader distribution with numerous extreme outliers compared to non-smokers. GTP is one of the strongest indicators of liver-related effects and oxidative stress. The substantial increase in GTP among smokers strongly suggests that smoking has a major negative effect on liver health. These results strongly suggest that liver is the most affected organ.

The third boxplot illustrates serum creatinine levels, which are used to assess kidney function. Smokers show slightly higher median creatinine values than non-smokers, along with a wider range of values and several extreme outliers. Elevated creatinine levels indicate reduced kidney filtration efficiency and possible renal stress. Although the difference is smaller compared to liver biomarkers, it still suggests that smoking has a measurable impact on kidney health.

From a statistical perspective, these differences were further validated using independent sample *t*-tests between smokers and non-smokers. The *P*-values obtained for ALT, GTP, and serum creatinine were found to be statistically significant ($P < 0.05$), confirming that the observed differences are not due to random variation. Among all biomarkers, GTP showed the strongest significance, followed by ALT and serum creatinine. This validates that suggests, particularly in liver-related biomarkers.

Overall, the Figure 2 clearly demonstrates that smokers tend to have higher levels of liver and kidney biomarkers compared to non-smokers, indicating greater organ stress and dysfunction. The stronger elevation in GTP and ALT compared to creatinine supports the final organ impact pattern of the study: Liver > Kidney > Metabolic System, highlighting the liver as the strongest association due to smoking.

Figure 3 presents the correlation heatmap showing the relationships among demographic, physiological, biochemical, and clinical variables in the smoking dataset. The heatmap is used to identify the strength and direction of association between features, particularly focusing on how smoking status is related to liver, kidney, and metabolic biomarkers. Positive correlations are represented by warmer colors (red shades), while negative correlations are shown by cooler colors (blue shades). Stronger color intensity indicates a stronger correlation coefficient.

The Figure 3 shows that smoking has a positive correlation with several important biomarkers, particularly hemoglobin, serum creatinine, GTP, ALT, and triglycerides, indicating that smokers tend to have elevated levels of these clinical parameters. Among these, GTP shows one of the strongest positive correlations with smoking, suggesting significant liver stress and oxidative damage caused by smoking. Since GTP is a major liver function indicator. This finding suggests that the liver may be the most affected organ system.

How to cite this article:

Similarly, ALT and AST also show positive associations with smoking, further confirming liver inflammation and hepatic dysfunction among smokers. The strong positive correlation between ALT and AST themselves indicates that both liver enzymes increase together, which is clinically expected in liver-related effects. This strengthens the biological validity of the analysis.

The heatmap also shows a positive correlation between smoking and serum creatinine, which is a key indicator of kidney function. Elevated creatinine levels suggest reduced kidney filtration efficiency and renal stress. Although the strength of correlation is lower than that of liver biomarkers, it still indicates that smoking has a measurable negative impact on kidney health.

Among metabolic indicators, triglycerides, fasting blood sugar, and waist circumference also show positive associations with smoking, suggesting metabolic imbalance and increased cardiovascular risk among smokers. In contrast, HDL shows a negative correlation with smoking, which is clinically significant because lower HDL levels are associated with poorer cardiovascular health. This negative relationship indicates that smoking may reduce protective cholesterol levels, increasing long-term metabolic risk.

The heatmap also reveals strong inter-feature relationships such as height and weight, systolic and relaxation blood pressure, low-density lipoprotein (LDL) and total cholesterol, and left and right hearing/eyesight measures. These expected clinical correlations confirm the reliability and consistency of the dataset.

From a significance perspective, the correlation heatmap supports the statistical validation and machine learning findings by visually confirming that smoking is strongly associated with liver enzymes, kidney markers, and metabolic risk factors. The strongest positive associations are observed with GTP, hemoglobin, ALT, and serum creatinine, which were also identified as important features in Random Forest, XGBoost, and SHAP analysis. This agreement between correlation analysis and model-based explainability improves the robustness and scientific credibility of the study.

Figure 4 presents the feature importance analysis generated from two ML models-Random Forest and XGBoost-used for predicting smoking status and identifying the most influential biomarkers associated with smoking-related organ damage. Feature importance helps determine which clinical variables contribute the most to model prediction and provides strong evidence for identifying the organ systems most affected by smoking.

In the Random Forest Feature Importance plot, gender is identified as the most important feature with the highest importance score, indicating a strong association between gender and smoking behavior in the dataset. This is followed by hemoglobin, height, and GTPs. Hemoglobin is often elevated in smokers due to reduced oxygen availability caused by carbon monoxide exposure, making it a strong physiological marker of smoking. GTP is one of the most significant liver enzymes and is a strong indicator of liver-related effects and oxidative stress. The presence of ALT, triglycerides, and serum creatinine among the top features further supports the involvement of liver, metabolic, and kidney systems, respectively. ALT reflects liver inflammation, triglycerides indicate metabolic imbalance, and serum creatinine is a key marker of kidney function.

In the XGBoost Feature Importance plot, gender again appears as the most dominant feature with a significantly higher importance score compared to all other variables. This consistency across both models confirms its strong predictive relevance. Among organ-related biomarkers, GTP, triglycerides, serum creatinine, and ALT remain important contributors. Although their importance scores are smaller compared to gender, their repeated appearance across both models increases confidence in their clinical significance. This consistency suggests that smoking has a measurable and explainable impact on liver function, kidney health, and metabolic regulation.

The presence of GTP and ALT in both models is particularly significant because these are major liver function indicators. Their high importance suggests that the liver may be the most affected organ due to smoking. Similarly, the inclusion of serum creatinine confirms kidney involvement, while triglycerides and LDL indicate metabolic disturbances associated with smoking. The repeated identification of these biomarkers across different algorithms strengthens the reliability of the study findings.

How to cite this article:

From a validation perspective, feature importance analysis supports the statistical tests, correlation heatmap, and SHAP explainability results. When multiple models consistently identify the same biomarkers as important, the robustness and scientific credibility of the study increase significantly. This agreement between Random Forest, XGBoost, and later SHAP analysis confirms that the findings are not model-specific but represent true physiological effects of smoking.

Figure 5 presents the SHAP summary plot used to explain the contribution of each clinical feature toward smoking prediction in the ML model. SHAP is an advanced XAI technique that helps interpret how each feature influences the model output by showing both the magnitude and direction of impact. In the plot, features are arranged from top to bottom based on their importance, with the most influential features appearing at the top. The x-axis represents the SHAP value, where positive values increase the probability of smoking prediction and negative values decrease it. The color scale represents feature values, where red indicates high values and blue indicates low values.

Figure 5 shows that gender is the most influential feature in predicting smoking status, followed by GTP, age, triglyceride, ALT, height, hemoglobin, tartar, and serum creatinine. This confirms the strong predictive importance of both demographic and clinical biomarkers. The wide spread of SHAP values for gender indicates that it has the strongest impact on model decisions, making it the dominant predictor in the classification process.

Among the organ-related biomarkers, GTP shows one of the highest SHAP impacts after gender. Higher GTP values (shown in red) are associated with positive SHAP values, meaning they increase the probability of being classified as a smoker. Since GTP is a major liver enzyme and a strong marker of oxidative stress and liver dysfunction, this strongly indicates that smoking has a severe effect on liver health. Similarly, ALT also shows significant influence, further confirming liver-related effects and hepatic inflammation caused by smoking.

The feature serum creatinine, which represents kidney function, also appears among the important predictors. Higher creatinine values contribute positively to smoking prediction, suggesting reduced kidney filtration efficiency and renal stress among smokers. Although its SHAP impact is smaller than liver biomarkers like GTP and ALT, it still demonstrates meaningful kidney involvement in smoking-related health effects.

Metabolic indicators such as triglyceride, LDL, and fasting blood sugar also show measurable influence on the model output. Higher triglyceride values are associated with positive SHAP values, indicating metabolic imbalance and increased cardiovascular risk among smokers. This supports the metabolic system's involvement, although its impact is relatively lower compared to liver and kidney markers.

An important observation is the consistency between SHAP results and previous analyses such as Random Forest feature importance, XGBoost importance, correlation heatmap, and statistical *t*-tests. Features like GTP, ALT, hemoglobin, triglycerides, and serum creatinine consistently appear as highly influential across all validation methods. This agreement significantly strengthens the robustness and scientific reliability of the study because it confirms that the identified biomarkers are not dependent on a single model.

From a significance perspective, SHAP provides both global and local interpretability, allowing the model to be clinically trustworthy rather than functioning as a black-box system. The strong SHAP influence of liver biomarkers, especially GTP and ALT, suggests that liver-related effects may represent the most prominent smoking-associated physiological changes observed in this dataset. Kidney involvement is represented by serum creatinine, while metabolic disturbance is represented by triglycerides and fasting blood sugar.

Figure 6 presents the distribution plots of important demographic, physiological, and biochemical variables including age, systolic blood pressure, cholesterol, hemoglobin, ALT, and GTP for smokers and non-smokers. These kernel density estimation (KDE) plots help visualize how the values of these biomarkers are distributed across both groups and provide additional clinical evidence regarding the multi-organ impact of smoking. By comparing the shape, spread, and peak positions of the distributions, it becomes easier to understand how smoking influences liver, kidney, and metabolic health.

How to cite this article:

The age distribution shows that smokers are more concentrated in the middle adult age groups, particularly between 35 and 55 years, while non-smokers show a slightly wider spread across age categories. This suggests that smoking prevalence is more common among middle-aged adults, which may increase the long-term risk of organ damage due to prolonged exposure to smoking-related toxins.

The systolic blood pressure distribution indicates that smokers tend to have slightly higher systolic values compared to non-smokers. The peak density for smokers shifts marginally toward higher blood pressure ranges, suggesting increased cardiovascular stress and possible hypertension risk. This supports the metabolic and circulatory burden associated with smoking.

The cholesterol distribution shows a largely overlapping pattern between smokers and non-smokers, but smokers demonstrate a slightly right-shifted distribution, indicating mildly elevated cholesterol levels. This suggests that smoking may contribute to lipid imbalance and increased cardiovascular disease risk, although the effect is less pronounced compared to liver-related biomarkers.

The hemoglobin distribution shows one of the clearest differences between the two groups. Smokers have noticeably higher hemoglobin levels compared to non-smokers, with the peak density shifted toward higher values. This is clinically significant because smoking increases carbon monoxide exposure, reducing oxygen availability and causing compensatory elevation in hemoglobin concentration. This makes hemoglobin one of the strongest physiological indicators of smoking behavior.

The ALT distribution demonstrates that smokers show a broader spread and slightly higher values compared to non-smokers. ALT is a major liver enzyme, and elevated ALT levels indicate liver inflammation or hepatocellular damage. The wider distribution and presence of extreme values among smokers suggest stronger liver stress caused by smoking.

The GTP distribution shows the most significant difference between smokers and non-smokers. Smokers exhibit a clear right-skewed distribution with much higher values and a longer tail, indicating frequent elevation of GTP levels. Since GTP is a strong marker of liver dysfunction and oxidative stress. This strongly suggests that smoking is associated with substantial liver-related effects. Among all biomarkers shown in the Figure 6, GTP provides the strongest evidence for smoking-related organ injury.

From a significance perspective, these distribution plots support the statistical *t*-test results, boxplot analysis, correlation heatmap, feature importance, and SHAP explainability findings. Variables such as GTP, ALT, hemoglobin, and systolic blood pressure show visible distributional differences between smokers and non-smokers, indicating statistically meaningful variation rather than random fluctuations. The consistent elevation of liver biomarkers particularly strengthens the conclusion that smoking has the greatest impact on liver function.

Based on the aggregated organ impact scores derived from statistical analysis, feature importance, and SHAP explainability, liver-related features consistently achieved higher combined scores compared to kidney and metabolic features. This provides a quantitative basis for the inferred organ damage pattern.

Comparison with state-of-the-art (SOTA) methods

Several recent studies have applied machine learning techniques for smoking prediction and health risk assessment. Prior works have primarily focused on predicting smoking status, identifying disease risks such as cardiovascular disorders, cancer, or respiratory diseases, and analyzing smoking behavior using clinical or textual data. However, most previous studies focused on prediction performance or disease-specific outcomes and provided limited organ-level interpretability [7,10,25,28].

Compared to existing approaches, the present study introduces a multi-dimensional framework that integrates prediction, statistical validation, and XAI. While previous studies emphasize prediction accuracy, this work extends the analysis by incorporating SHAP-based explainability to identify the contribution of individual biomarkers.

How to cite this article:

Furthermore, unlike traditional approaches that focus on a single disease or outcome, this study provides a holistic multi-organ perspective by simultaneously analyzing liver, kidney, and metabolic systems. The proposed organ-level interpretation offers clinically meaningful insights into how smoking affects different physiological systems [7,10,25,28].

Although deep learning models have been used in some studies for smoking-related prediction, they often lack interpretability, which limits their applicability in healthcare settings. In contrast, the use of interpretable ML models combined with SHAP enhances transparency and supports clinical decision-making.

Overall, the proposed framework contributes to the state-of-the-art by combining predictive modeling with explainability and multi-organ analysis, providing a more comprehensive and interpretable understanding of smoking-related health effects.

Limitations

Although this study provides comprehensive insights using statistical analysis and explainable machine learning techniques, it has certain limitations. The organ-level damage ranking proposed in this work is inferred from feature-level importance, statistical testing, and SHAP explainability rather than direct experimental validation. Specifically, feature-group ablation and LOGO analyses were not performed to quantify the contribution of each organ system. Therefore, the proposed ranking (Liver > Kidney > Metabolic System) should be interpreted as indicative rather than definitive. Future work should incorporate organ-level validation experiments and longitudinal clinical data to strengthen causal interpretation and improve clinical applicability.

The aggregated Organ Impact Scores indicated that liver-related features achieved the highest score, followed by kidney and metabolic features, supporting the inferred ranking of Liver > Kidney > Metabolic System.

In addition, potential confounding factors such as age, sex, alcohol consumption, lifestyle characteristics, and underlying comorbidities may influence biomarker levels and were not comprehensively controlled in the present dataset.

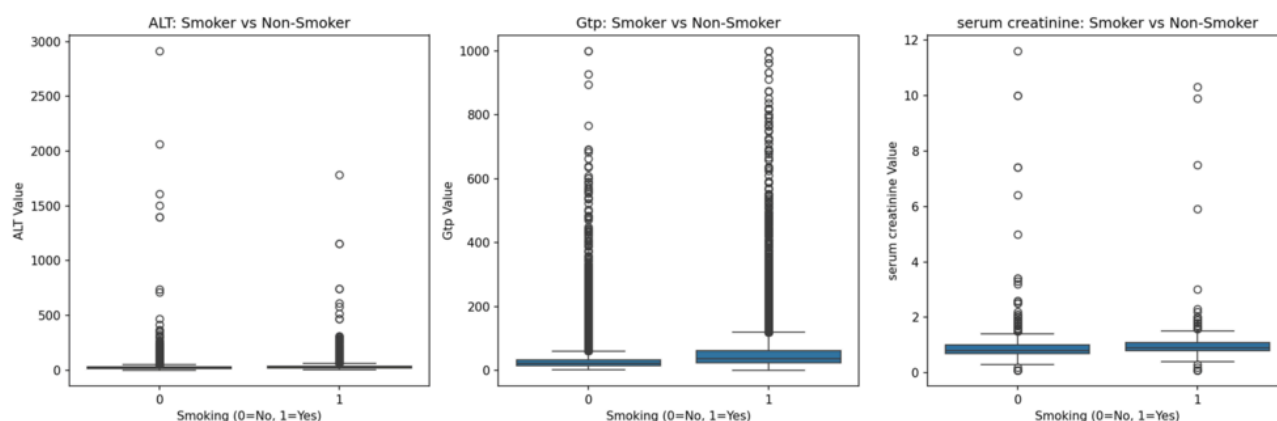


FIGURE 2: Boxplot comparison of liver, kidney, and metabolic biomarkers between smokers and non-smokers.

ALT, Alanine Aminotransferase

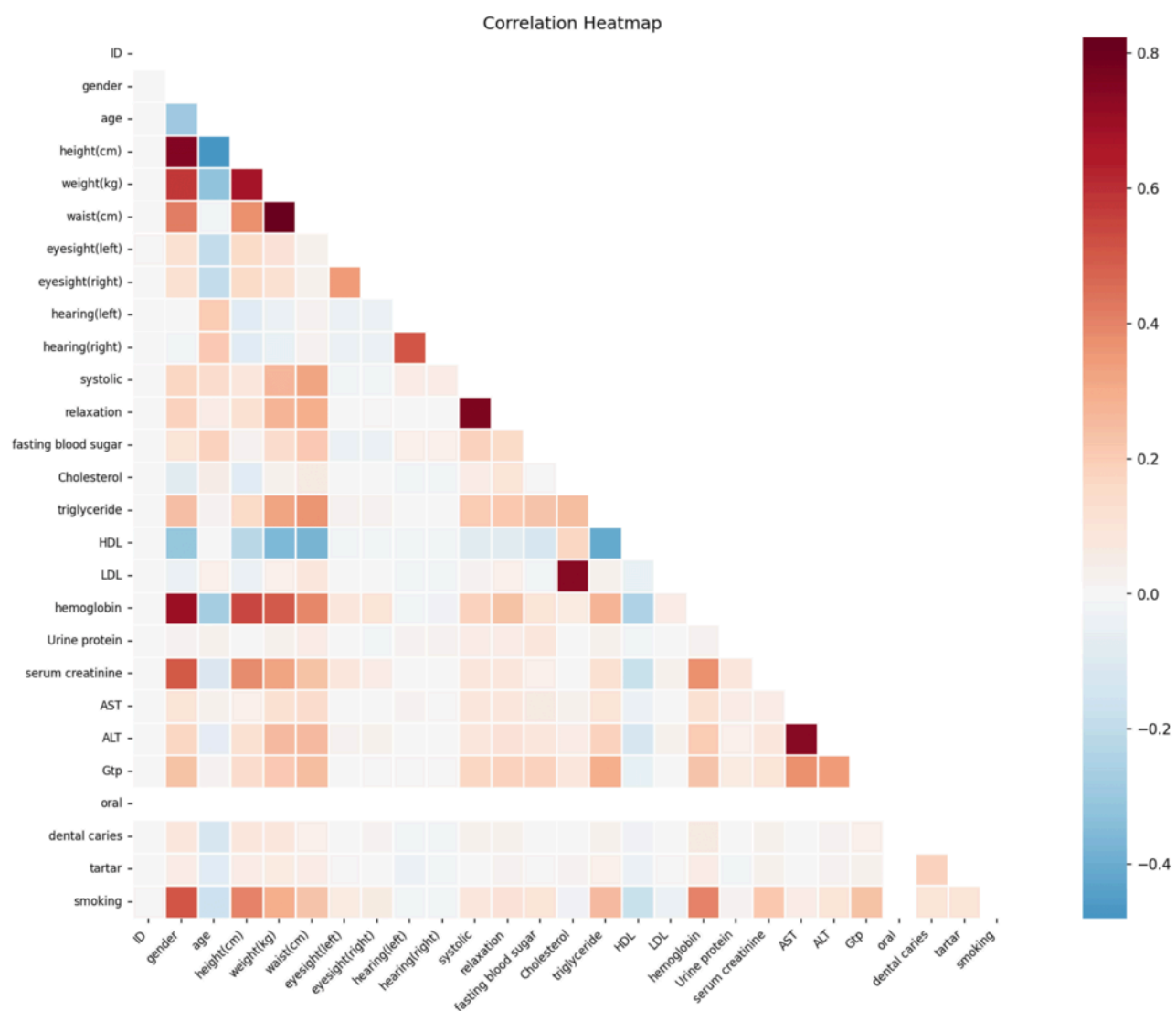


FIGURE 3: Correlation heatmap of clinical biomarkers and smoking status.

HDL, High-density Lipoprotein; LDL, Low-density Lipoprotein; ALT, Alanine Aminotransferase; AST, Aspartate Aminotransferase; GTP, Gamma-Glutamyl Transferase

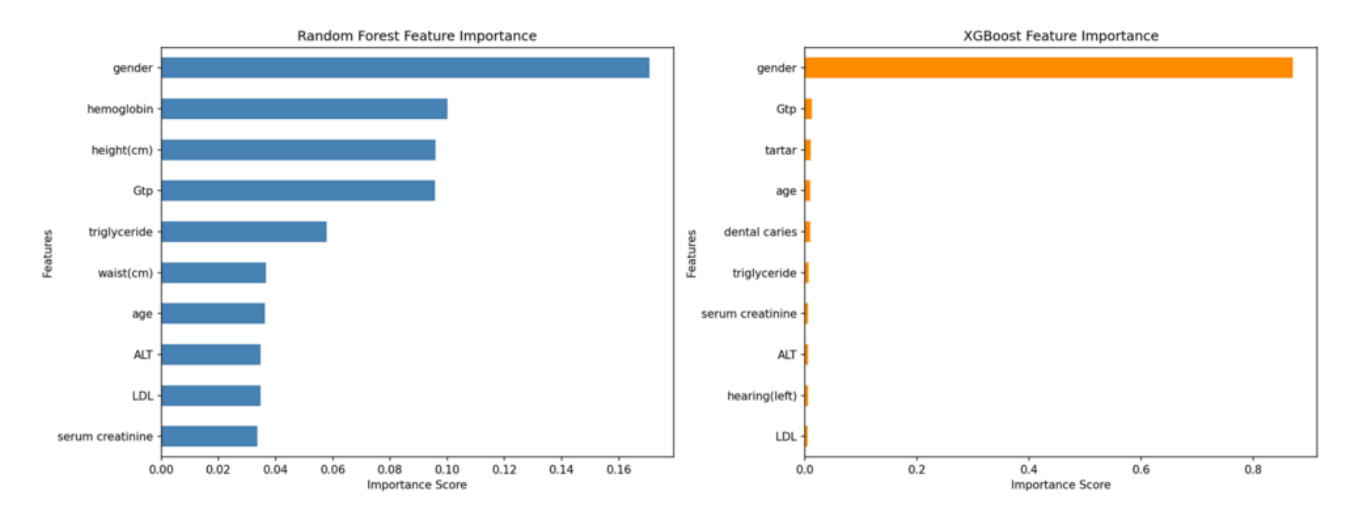


FIGURE 4: Feature importance comparison of Random Forest and XGBoost models for smoking prediction.

ALT, Alanine Aminotransferase; LDL, Low-density Lipoprotein; GTP, Gamma-glutamyl Transferase

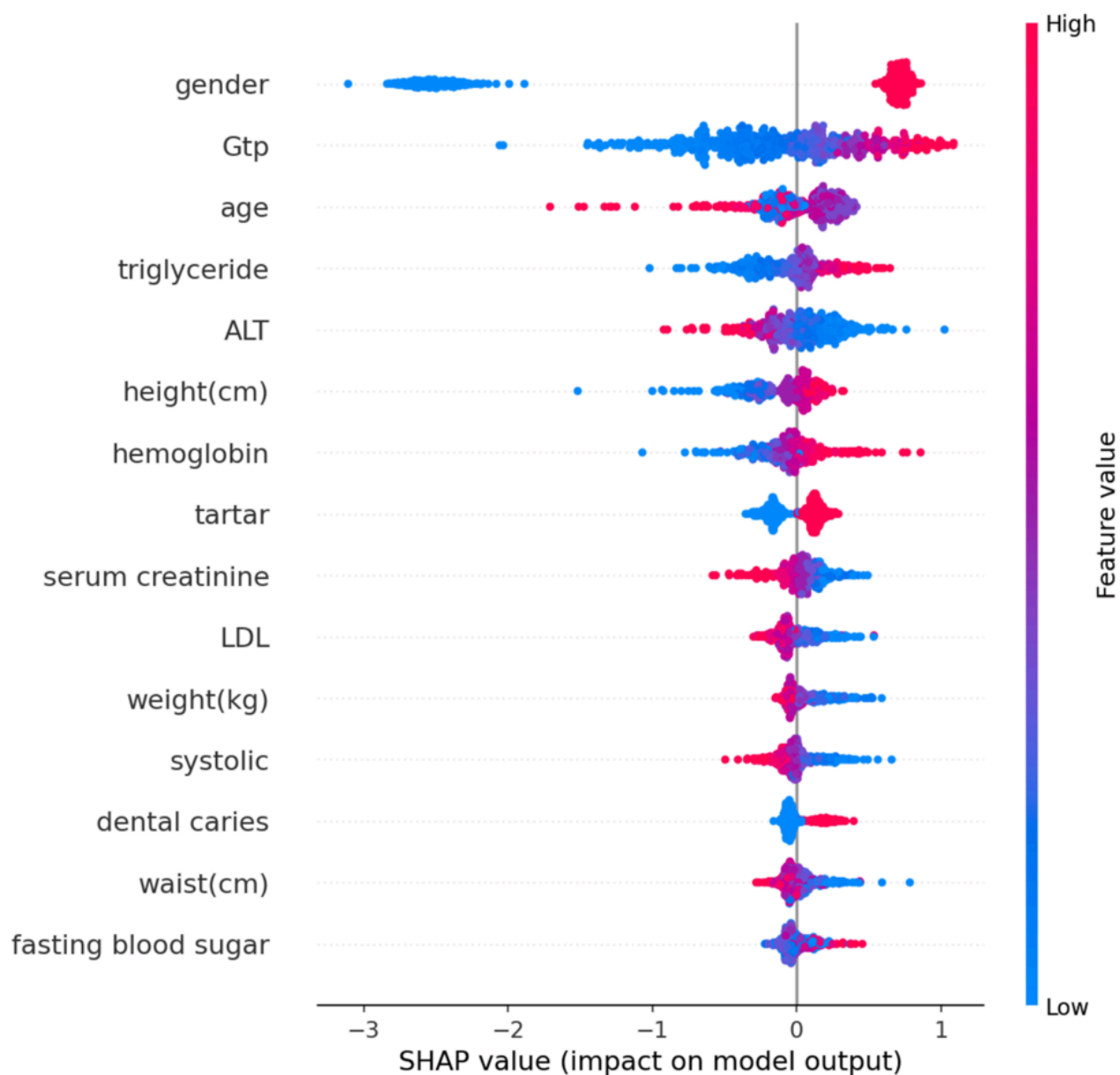


FIGURE 5: SHAP summary plot for explainable analysis of smoking-related biomarkers.
SHAP, SHapley Additive exPlanations; LDL, Low-density Lipoprotein; GTP, Gamma-glutamyl Transferase; ALT, Alanine Aminotransferase

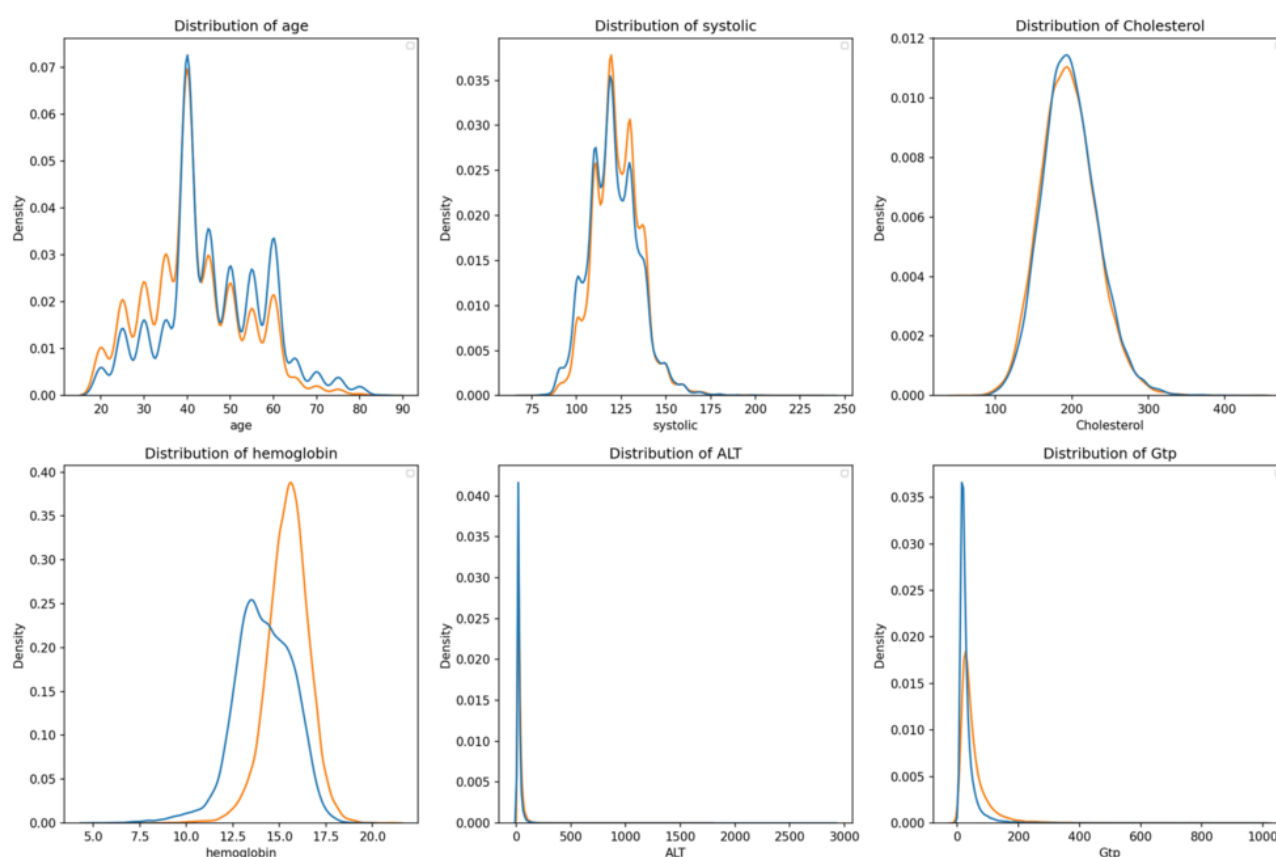


FIGURE 6: Distribution plots of clinical and biochemical variables between smokers and non-smokers.

ALT, Alanine Aminotransferase; GTP, Gamma-glutamyl Transferase

Conclusions

This study presented an XAI-based framework to investigate the multi-organ effects of smoking using the Kaggle Body Signal of Smoking dataset comprising 55,692 records and 27 clinical features. Random Forest, XGBoost, and Logistic Regression models were developed and evaluated using statistical validation and SHAP-based explainability. The results consistently identified GTP, ALT, serum creatinine, hemoglobin, and triglycerides as key biomarkers associated with smoking. Combined evidence from statistical testing, feature importance analysis, and SHAP interpretation suggested a potential organ impact pattern of Liver > Kidney > Metabolic System, indicating stronger smoking-related effects on liver function compared with kidney and metabolic health.

The primary contribution of this work is the integration of ML, statistical analysis, and XAI into a transparent framework for smoking-related health assessment and biomarker interpretation. The proposed approach may support early risk identification and explainable clinical decision-making. Future studies should incorporate organ-level validation methods such as feature-group ablation and LOGO analysis, together with longitudinal clinical datasets, to further validate the proposed organ ranking.

Additional Information

Author Contributions

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

How to cite this article:

V S (June 16, 2026) Multi-Organ Impact of Smoking Using Explainable Artificial Intelligence: Prediction, Validation, and Organ Impact Pattern. *Cureus J Comput Sci* 3 : es44389-026-00121-y. DOI <https://doi.org/10.7759/s44389-026-00121-y>

Concept and design: SOMASHEKAR V

Acquisition, analysis, or interpretation of data: SOMASHEKAR V

Drafting of the manuscript: SOMASHEKAR V

Critical review of the manuscript for important intellectual content: SOMASHEKAR V

Disclosures

Human subjects: All authors have confirmed that this study did not involve human participants or tissue. **Animal subjects:** All authors have confirmed that this study did not involve animal subjects or tissue. **Conflicts of interest:** In compliance with the ICMJE uniform disclosure form, all authors declare the following: **Payment/services info:** All authors have declared that no financial support was received from any organization for the submitted work. **Financial relationships:** All authors have declared that they have no financial relationships at present or within the previous three years with any organizations that might have an interest in the submitted work. **Other relationships:** All authors have declared that there are no other relationships or activities that could appear to have influenced the submitted work.

Data Availability Statements

The datasets (and/or code) supporting this study are available from the corresponding author upon reasonable request.

References

1. Olawade DB, Aienobe-Asekharen CA: [Artificial intelligence in tobacco control: a systematic scoping review of applications, challenges, and ethical implications](#). International Journal of Medical Informatics. 2025, 202:1-18. [10.1016/j.ijmedinf.2025.105987](#)
2. Jonas MA, Oates JA, Ockene JK, Hennekens CH: [Statement on smoking and cardiovascular disease for health care professionals](#). Circulation. 1992, 86:1664-1669. [10.1161/01.cir.86.5.1664](#)
3. Gometz ED: [Health effects of smoking and the benefits of quitting](#). The AMA Journal of Ethics. 2011, 13:31-35. [10.1001/virtualmentor.2011.13.1.cprl1-1101](#)
4. Alarabi AB, Alshbool FZ, Khasawneh FT: [Impact of smoking on the complement system: a narrative review](#). Frontiers in Immunology. 2025, 16:1-13. [10.3389/fimmu.2025.1619835](#)
5. Parmar MP, Kaur M, Bhavanam S, et al.: [A systematic review of the effects of smoking on the cardiovascular system and general health](#). Cureus Journal of Medical Science. 2023, 15:e38073. [10.7759/cureus.38073](#)
6. Gupta S, Kumar V, Gupta P: [A comprehensive study on the harmful effects of the smoking on human beings](#). Challenges in Information, Communication and Computing Technology. Sharmila V, Kannadhasan S, Kannan AR, Sivakumar P, Vennila V (ed): CRC Press, London; 2024. 577-582.
7. Chakma V, Nerab MDJH, Rouf A, Sayed A, Saim HMD, Khan MN: [Machine learning models for predicting smoking-related health decline and disease risk](#). arXiv. 2025, 1-30. [10.48550/arXiv.2511.14682](#)
8. Gallucci G, Tartarone A, Lerosé R, Lalinga AV, Capobianco AM: [Cardiovascular risk of smoking and benefits of smoking cessation](#). Journal of Thoracic Disease. 2020, 12:3866-3876. [10.21037/jtd.2020.02.47](#)
9. Chang R, Jiao H, Nie W, et al.: [Organ-Agents: virtual human physiology simulator](#). arXiv. 2025, 1-40. [10.48550/arXiv.2508.14357](#)
10. Aishwarya S, Siddalingaswamy PC, Chadaga K: [Explainable artificial intelligence driven insights into smoking prediction using machine learning and clinical parameters](#). Scientific Reports. 2025, 15:1-23. [10.1038/s41598-025-09409-w](#)
11. Zhang T, Chung T, Dey A, Bae SW: [Exploring algorithmic explainability: generating explainable ai insights for personalized clinical decision support focused on cannabis intoxication in young adults](#). International Conference on Activity and Behavior Computing. 2024, 1-15. [10.1109/ABC61795.2024.10652070](#)
12. Ayadi H, Elbéji A, Despotovic V, Fagherazzi G: [Digital vocal biomarker of smoking status using ecological audio recordings: results from the Colive Voice Study](#). Digital Biomarkers. 2024, 8:159-170. [10.1159/000540327](#)

How to cite this article:

V S (June 16, 2026) Multi-Organ Impact of Smoking Using Explainable Artificial Intelligence: Prediction, Validation, and Organ Impact Pattern. Cureus J Comput Sci 3 : es44389-026-00121-y. DOI <https://doi.org/10.7759/s44389-026-00121-y>

13. Caccamisi A, Jørgensen L, Dalanian H, Rosenlund M: [Natural language processing and machine learning to enable automatic extraction and classification of patients' smoking status from electronic medical records](#). Upsala Journal of Medical Sciences. 2020, 125:316-324. [10.1080/03009734.2020.1792010](#)
14. Evans A, Hughes R, Nolan L, Little K, Newbury-Davies L, Davies AR: [Profiles of tobacco smokers and ex-smokers in a large-scale random sample survey across Wales: an unsupervised machine-learning cluster analysis](#). The Lancet. 2023, 402:S7. [10.1016/s0140-6736\(23\)02070-6](#)
15. Harris DR, Henderson DW, Corbeau A: [Improving the utility of tobacco-related problem list entries using natural language processing](#). AMIA Annual Symposium Proceedings. 2021, 2020:534-543.
16. Hegde H, Shimpi N, Glurich I, Acharya A: [Tobacco use status from clinical notes using natural language processing and rule based algorithm](#). Technology and Health Care. 2018, 26:445-456. [10.3233/thc-171127](#)
17. Mamoshina P, Kochetov K, Cortese F, et al.: [Blood biochemistry analysis to detect smoking status and quantify accelerated aging in smokers](#). Scientific Reports. 2019, 9:1-10. [10.1038/s41598-018-35704-w](#)
18. Palmer EL, Hassanpour S, Higgins J, Doherty JA, Onega T: [Building a tobacco user registry by extracting multiple smoking behaviors from clinical notes](#). BMC Medical Informatics and Decision Making. 2019, 19:1-10. [10.1186/s12911-019-0863-3](#)
19. Patel J, Siddiqui Z, Krishnan A, Thyvalikakath T: [Identifying patients' smoking status from electronic dental records data](#). IOS Press, 2017. [10.3233/978-1-61499-830-3-1281](#)
20. Amaral JLM, Lopes AJ, Jansen JM, Faria ACD, Melo PL: [An improved method of early diagnosis of smoking-induced respiratory changes using machine learning algorithms](#). Computer Methods and Programs in Biomedicine. 2013, 112:441-454. [10.1016/j.cmpb.2013.08.004](#)
21. Fiste O, Gkiozos I, Charpidou A, Syrigos NK: [Artificial intelligence-based treatment decisions: a new era for NSCLC](#). Cancers. 2024, 16:1-15. [10.3390/cancers16040831](#)
22. Karlsson A, Ellonen A, Irjala H, et al.: [Impact of deep learning-determined smoking status on mortality of cancer patients: never too late to quit](#). ESMO Open. 2021, 6:1-8. [10.1016/j.esmoop.2021.100175](#)
23. Ma Q, Shen Y, Guo W, Feng K, Huang T, Cai Y: [Machine learning reveals impacts of smoking on gene profiles of different cell types in lung](#). Life. 2024, 14:1-26. [10.3390/life14040502](#)
24. Muro S, Ishida M, Horie Y, et al.: [Machine learning methods for the diagnosis of chronic obstructive pulmonary disease in healthy subjects: retrospective observational cohort study](#). JMIR Medical Informatics. 2021, 9:1-13. [10.2196/24796](#)
25. Wang X, Qiao Y, Cui Y, et al.: [An explainable artificial intelligence framework for risk prediction of COPD in smokers](#). BMC International Health and Human Rights. 2023, 23:1-18. [10.1186/s12889-023-17011-w](#)
26. Belsare P, Senyurek VY, Imtiaz MH, Tiffany ST, Sazonov E: [DeepPuff: utilizing deep learning for smoking behavior identification in free-living environment](#). Annual International Conference of the IEEE Engineering in Medicine and Biology Society. 2023, 1-5. [10.1109/embc40787.2023.10340528](#)
27. Barbieri S, Mehta S, Wu B, Bharat C, Poppe K, Jorm L, Jackson R: [Predicting cardiovascular risk from national administrative databases using a combined survival analysis and deep learning approach](#). International Journal of Epidemiology. 2021, 51:931-944. [10.1093/ije/dyab258](#)
28. Khimani A, Hornback A, Jain N, Avula P, Jaishankar A, Wang MD: [Predicting cardiovascular disease risk in tobacco users using machine learning algorithms](#). Annual International Conference of the IEEE Engineering in Medicine and Biology Society. 2024, 1-5. [10.1109/embc53108.2024.10782885](#)
29. Meeradevi T, Sasikala S, Murali L, Manikandan N, Ramaswamy K: [Lung cancer detection with machine learning classifiers with multi-attribute decision-making system and deep learning model](#). Scientific Reports. 2025, 15:1-19. [10.1038/s41598-025-88188-w](#)
30. Wang KM, Chen KH, Hernanda CA, Tseng SH, Wang KJ: [How is the lung cancer incidence rate associated with environmental risks? Machine-learning-based modeling and benchmarking](#). International Journal of Environmental Research and Public Health. 2022, 19:1-19. [10.3390/ijerph19148445](#)
31. Zhan Y, Weng M, Guo Y, et al.: [Identification and validation of the nicotine metabolism-related signature of bladder cancer by bioinformatics and machine learning](#). Frontiers in Immunology. 2024, 15:1-15. [10.3389/fimmu.2024.1465638](#)
32. Zhu H, Chu B, Zhang C, Liu F, Jiang L, He Y: [Hyperspectral imaging for presymptomatic detection of tobacco disease with successive projections algorithm and machine-learning classifiers](#). Scientific Reports. 2017, 7:1-12. [10.1038/s41598-017-04501-2](#)

How to cite this article:

33. Hébert ET, Suchting R, Ra CK, Alexander AC, Kendzor DE, Vidrine DJ, Businelle MS: [Predicting the first smoking lapse during a quit attempt: a machine learning approach](#). Drug and Alcohol Dependence. 2020, 218:108340. [10.1016/j.drugalcdep.2020.108340](#)
34. Kim N, McCarthy DE, Loh WY, Cook JW, Piper ME, Schlam TR, Baker TB: [Predictors of adherence to nicotine replacement therapy: machine learning evidence that perceived need predicts medication use](#). Drug and Alcohol Dependence. 2019, 205:107668. [10.1016/j.drugalcdep.2019.107668](#)
35. Singh T, Truong M, Roberts K, Myneni S: [Sequencing conversational turns in peer interactions: an integrated approach for evidence-based conversational agent for just-in-time nicotine cravings intervention](#). Digital health. 2024, 10:1-12. [10.1177/20552076241228430](#)
36. Suchting R, Hébert ET, Ma P, Kendzor DE, Businelle MS: [Using elastic net penalized Cox proportional hazards regression to identify predictors of imminent smoking lapse](#). Nicotine & Tobacco Research. 2017, 21:173-179. [10.1093/ntr/ntx201](#)
37. Franzwa F, Harper LA, Anderson KG: [Examination of social smoking classifications using a machine learning approach](#). Addictive Behaviors. 2021, 126:107175. [10.1016/j.addbeh.2021.107175](#)
38. Engelhard MM, Oliver JA, Henao R, et al.: [Identifying smoking environments from images of daily life with deep learning](#). JAMA Network Open. 2019, 2:e197939. [10.1001/jamanetworkopen.2019.7939](#)
39. Çetintaş E, Luo Y, Nguyen C, Guo Y, Li L, Zhu Y, Ozcan A: [Characterization of exhaled e-cigarette aerosols in a vape shop using a field-portable holographic on-chip microscope](#). Scientific Reports. 2022, 12:1-8. [10.1038/s41598-022-07150-2](#)
40. Parks J, McLean KE, McCandless L, et al.: [Assessing secondhand and thirdhand tobacco smoke exposure in Canadian infants using questionnaires, biomarkers, and machine learning](#). Journal of Exposure Science & Environmental Epidemiology. 2021, 32:112-123. [10.1038/s41370-021-00350-4](#)
41. Body signal of smoking. (2022). <https://www.kaggle.com/datasets/kukuroo3/body-signal-of-smoking>.

How to cite this article: