computation

MDPI

Article

Quantifying Domain-Specific Risk Signals in Lung Cancer Severity Prediction: A Multi-Domain Ablation Study Using XGBoost and SHAP

Sidra Ishfaq [1], Muhammad Abdullah Khan [2], Ghulam Mustafa [1,*], Muhammad Tanvir Afzal [3], Isabel De la Torre Díez [4], Mirtha Silvana Garat de Marin [5,6,7,8] and Eduardo Silva Alvarado [9,10]

[1] Department of Computer Science, Shifa Tameer-e-Millat University, Islamabad 44000, Pakistan; sidra_ishfaq.ssc@stmu.edu.pk
[2] School of Science, Engineering and Environment, University of Salford, Salford M5 4WT, UK; mak30597@gmail.com
[3] Department of Computing and Data Science, GISMA University of Applied Sciences, 14469 Potsdam, Germany; tanvir.afzal@gisma.com
[4] eHealth and Telemedicine Group, University of Valladolid, 47011 Valladolid, Spain; isator@uva.es
[5] Department of Projects, Universidad Internacional Iberoamericana, Arecibo, PR 00613, USA; silvana.marin@unic.co.ao
[6] Department of Projects, Universidade Internacional do Cuanza, Cuito EN250, Bié Province, Angola
[7] Facultad de Ciencias de la Salud, Fundación Universitaria Internacional de Colombia, Bogotá 110111, Colombia
[8] Department of Projects, Universidad Internacional Iberoamericana, Campeche 24560, Mexico
[9] Department of Projects, Universidad Europea del Atlántico, Isabel Torres 21, 39011 Santander, Spain; eduardo.silva@uneatlantico.es
[10] Department of Projects, Universidad de La Romana, La Romana 22000, Dominican Republic
* Correspondence: ghulam.mustafa.ssc@stmu.edu.pk

Abstract

Predictive modeling for lung cancer severity often struggles with the high dimensionality and multi-domain nature of risk factors. While individual contributors like smoking are well-documented, the relative predictive weight of lifestyle, environmental, and genetic domains remains insufficiently quantified in integrated frameworks. This study proposes an explainable machine learning approach using an XGBoost classifier to evaluate these three distinct risk domains. Utilizing the UCI Machine Learning Repository Lung Cancer Dataset, we implemented a domain-wise ablation study to isolate the predictive signal of each factor group. To ensure scientific rigor and address the "black box" nature of ensemble models, we employed 5-fold stratified cross-validation and SHAP (Shapley Additive Explanations) for feature-level transparency. Our results demonstrate that the integrated model achieves a classification accuracy of 95.7% (AUC-ROC = 0.98) on this dataset. Notably, ablation analysis revealed that the Lifestyle domain retained the highest standalone predictive performance (92.9%), followed by the Genetic/Clinical domain (94.6%), while the Environmental domain showed a more pronounced performance drop (73.3%), suggesting differential information density across risk categories. SHAP analysis identified cumulative smoking exposure as the primary feature influencing model predictions within this dataset. This study presents a proof-of-concept interpretable framework for lung cancer risk stratification, demonstrating that domain-wise ablation combined with explainable AI can provide transparent, feature-level insight to support rather than replace clinical judgment in settings where comprehensive diagnostic testing may be limited.

Keywords: lung cancer severity; risk stratification; machine learning; XGBoost; lifestyle risk factors; environmental exposure; interpretable machine learning

Academic Editor: Simone Brogi

Received: 24 April 2026
Revised: 10 May 2026
Accepted: 12 May 2026
Published: 20 May 2026

1. Introduction

The integration of Artificial Intelligence (AI) and Machine Learning (ML) has driven substantial progress in predictive oncology, enabling the analysis of high-dimensional clinical datasets that exceed the capacity of conventional statistical methods ref. [1]. Supervised learning algorithms including Support Vector Machines (SVMs), Random Forests, and Neural Networks have been applied to cancer detection tasks with demonstrated effectiveness ref. [2], and gradient-boosted decision trees have emerged as a particularly competitive approach for structured tabular clinical data ref. [3]. However, much of the existing literature focuses on single-domain analysis, predominantly genetic markers or imaging data, often overlooking the potential value of integrating behavioural lifestyle metrics with environmental exposure data within a unified computational framework ref. [4].

Despite this progress, two methodological gaps continue to limit the clinical utility of ML-based lung cancer risk models. First, high-performing ensemble architectures frequently operate as "black boxes," producing a risk score without a transparent behavioural or biological rationale. For clinical adoption, a model must be interpretable enough to support rather than replace physician judgment ref. [5]. Second, there is a notable absence of ablation-based comparative research that systematically quantifies the relative predictive signal contributed by distinct risk domains ref. [3]. Without such evidence, it is difficult to determine whether lifestyle factors, environmental exposures, or genetic predispositions carry the greater discriminative weight in severity stratification, which in turn limits the specificity of evidence-based public health intervention ref. [4].

Crucially, while recent work such as iPSOgs ref. [6] and DSMAL ref. [7] demonstrates strong XGBoost-based classification in high-resolution genomic and transcriptomic settings, no existing study applies a domain-stratified ablation approach to quantify the relative predictive weight of behavioural, environmental, and genetic risk domains from structured clinical data the precise gap this work addresses. To address these gaps, this study presents a proof-of-concept, explainable machine learning framework for lung cancer severity classification using an optimized XGBoost classifier ref. [4]. The framework is evaluated on the Kaggle Lung Cancer Patient Dataset a publicly available, structured, and fully de-identified cohort of 1000 patient records spanning 24 clinical, behavioural, and environmental features ref. [1]. The central methodological contribution is a *domain-wise ablation study* in which features are partitioned into three clinically motivated groups, Lifestyle, Environmental, and Genetic/Clinical, and each group's standalone predictive signal is systematically measured under identical experimental conditions. To ensure reproducibility and guard against data leakage, all preprocessing transformations are fitted exclusively within training folds of a 5-fold stratified cross-validation protocol, and performance is reported with standard deviations and 95% confidence intervals across folds ref. [8].

To address the interpretability gap, SHAP (Shapley Additive Explanations) is integrated to decompose each prediction into per-feature contributions, providing local and global transparency for the XGBoost model's decision logic ref. [9]. It is important to note that SHAP values quantify how a particular feature influenced the output of this specific model on this specific dataset; they should not be interpreted as measures of real-world biological causality ref. [10]. To corroborate the SHAP-derived rankings with model-agnostic evidence, the Maximal Information Coefficient (MIC) and distance correlation are additionally computed directly on the data, independent of any model assumptions ref. [11,12]. All accuracy figures reported in this work are in-distribution estimates derived from a structured, synthetically scaled dataset and should be understood as proof-of-concept findings rather than generalizable clinical benchmarks; external validation on independent cohorts is an acknowledged prerequisite for any future deployment.

The contributions of this paper are threefold:

(i) **Domain-wise ablation framework.** To the best of our knowledge, no prior study on structured lung cancer risk data has simultaneously (a) partitioned features into all three clinically motivated domains Lifestyle, Environmental, and Genetic/Clinical; (b) conducted recursive, leakage-free ablation to measure each domain's standalone discriminative signal under identical experimental conditions; and (c) quantified the resulting predictive decay (ΔA) with per-fold standard deviations and 95% bootstrap confidence intervals. This structured methodology directly addresses the absence of quantitative, domain-level comparative evidence identified in the literature (Table 1), and is the primary scientific advance of this work independent of the absolute accuracy figures reported.

(ii) **Cross-validated explainability pipeline.** Rather than relying on a single attribution method, we cross-validate TreeSHAP importance rankings against two model-agnostic measures the Maximal Information Coefficient (MIC) and distance correlation (dCor) computed directly on the data prior to any model fitting. This three-way convergence substantially reduces the risk that reported feature hierarchies are artefacts of the XGBoost architecture or the SHAP attribution procedure refs. [10–12], providing a stronger evidentiary basis for the domain-level interpretability claims than SHAP alone could support.

(iii) **Transparent and reproducible statistical reporting.** All performance claims are accompanied by per-fold standard deviations, 95% bootstrap confidence intervals, and baseline comparisons against Logistic Regression and Random Forest trained under identical conditions. A cross-dataset transfer experiment on the UCI Lung Cancer Dataset is additionally included to empirically bound the in-distribution performance figures and honestly quantify the generalisation limits of the framework a prerequisite for responsible proof-of-concept reporting ref. [13].

Table 1. Comparison of recent machine learning studies on lung cancer risk modeling, highlighting feature domain coverage and interpretability approaches. Few existing studies perform explicit domain-wise ablation or combine lifestyle, environmental, and genetic factors within a unified explainable framework.

Study	Year	Method	Dataset	Life	Env.	Gen./Clin.	Explainability
Abdu-Aljabar & Awad ref. [14]	2021	XGBoost	284	✓	×	✓	None
Guan et al. ref. [4]	2023	XGBoost	424	✓	×	✓	SHAP
Singh et al. ref. [3]	2023	Ensemble (XGBoost, RF, LR)	1000	✓	×	✓	None
Wayahdi & Ruziq ref. [15]	2022	KNN + XGBoost	284	✓	×	×	None
Li et al. ref. [1]	2022	XGBoost	4526	✓	×	✓	None
Yuan et al. ref. [8]	2022	XGBoost	312	×	×	✓	SHAP
Sweet et al. ref. [16]	2024	SVM, RF, NB, LR	1000	✓	✓	✓	None
Hosseini et al. ref. [17]	2025	Radiomic ML	Multi-site	×	×	✓	None
Wang et al. ref. [18]	2025	Transformer + XGBoost	Large cohort	×	×	✓	SHAP
Paryani et al. ref. [5]	2025	XGBoost + hybrid	Clinical	✓	×	✓	Partial
Guinovart et al. ref. [19]	2026	iPSOgs + XGBoost	TCGA/GEO	×	×	✓	SHAP
This study	**2026**	**XGBoost + domain ablation**	**1000**	✓	✓	✓	**SHAP**

Note: ✓ = domain included in the study; × = domain not included or not reported. Lifestyle domain includes behavioral factors (smoking, alcohol, BMI, diet). Environmental domain includes air pollution, occupational exposure, and passive smoking. Genetic/Clinical domain includes family history, pulmonary function indices, and biomarkers. RF = Random Forest; LR = Logistic Regression; SVM = Support Vector Machine; NB = Naïve Bayes; KNN = K-Nearest Neighbors; SHAP = Shapley Additive Explanations.

Taken together, these three contributions constitute a methodologically coherent proof-of-concept pipeline whose novelty lies not in any single component XGBoost, SHAP, and cross-validation are each well-established individually but in their principled integration within a unified, domain-stratified, leakage-free interpretability framework that no prior work on this class of structured lung cancer data has assembled. The practical value of this integration is precisely that it makes the relative information content of each risk domain

visible and statistically quantifiable, which is the key insight that aggregated, non-ablated models cannot provide.

The remainder of this paper is organized as follows: Section 2 reviews related work; Section 3 describes the methodology; Section 4 presents experimental results and comparative performance analysis; Section 5 discusses implications and limitations; and Section 6 concludes with directions for future research.

2. Literature Review

The following literature review provides a critical synthesis of the computational landscape in oncological risk stratification, spanning developments from 2024 to early 2026. Rather than providing a perfunctory summary of existing models, this section maps the methodological transition from high-variance "black-box" classifiers to the current state-of-the-art focus on Explainable AI (XAI) and multi-domain feature ablation ref. [15]. By evaluating the trajectory of Gradient Boosted Decision Trees (GBDTs) and the rising demand for clinical transparency, we identify a persistent gap in current research, specifically the lack of quantitative comparisons between lifestyle, environmental, and genetic risk signals ref. [3]. This review establishes the theoretical and empirical foundation for our proposed XGBoost–SHAP framework, demonstrating how our approach directly addresses the limitations of synthetic data augmentation and model opacity identified in recent high-impact literature ref. [15].

The landscape of predictive oncology has undergone a significant shift between 2024 and 2026, transitioning from basic diagnostic classifiers toward interpretable decision support systems ref. [1]. The period of 2024–2025 marked the consolidation of Gradient Boosted Decision Trees (GBDTs), specifically XGBoost, as a leading method for structured clinical data assessment. Recent work by ref. [13] in *Briefings in Bioinformatics* (2025) demonstrated the utility of hybrid architectures combining Transformer-based survival models with explainable XGBoost frameworks to predict outcomes in complex biological pathways, such as those centered on the cGAS–STING signaling axis in hepatocellular carcinoma. Similarly, ref. [14] showed that while deep learning models perform well in image-based diagnostics, XGBoost consistently demonstrates competitive performance in tabular risk assessment due to its handling of feature sparsity and non-linear interactions ref. [3]. Directly relevant to the present work, ref. [6] proposed iPSOgs an enhanced particle swarm optimization algorithm paired with XGBoost for simultaneous hyperparameter tuning and gene selection in non-small-cell lung cancer subtype classification, achieving an accuracy of 95.8% and an AUC-ROC of 0.987 on transcriptomic cohorts (TCGA-LUAD, TCGA-LUSC, GSE81089). That framework operates in the high-resolution genomic feature space, whereas the present study addresses the complementary problem of quantifying domain-level predictive utility from structured behavioural and environmental risk features an applicability gap that iPSOgs does not address. A recurring limitation in 2024–2025 literature, however, is the heavy reliance on synthetic data augmentation, which ref. [4] argues often introduces distributional shift that undermines real-world validation ref. [14].

A significant challenge in current machine learning for oncology remains the stability and generalizability of features across diverse clinical settings ref. [4]. Recent work distinguished between robust and non-robust radiomic features through phantom and clinical studies, highlighting that previous models often failed to generalise due to feature instability across different imaging protocols and hardware ref. [4]. Furthermore, ref. [15] expanded the diagnostic utility of radiomics by differentiating pathological subtypes of primary lung cancer using brain CT images for patients with brain metastases, demonstrating that distal feature extraction can provide meaningful diagnostic signals ref. [1]. The challenge of selecting a compact, discriminative feature subset from high-dimensional clinical data is

not unique to lung cancer. Notably, ref. [7] proposed DSMAL a Differential-Evolution and Slime-Mould Algorithm with an adaptive Refresh Local Search operator integrated with XGBoost for gene-expression-based diagnosis of primary Sjögren's syndrome, achieving a 96.6% F1-score and 97.6% AUC on an independent external test set. This cross-domain evidence further corroborates the suitability of XGBoost as a clinical classifier and underscores the broader challenge shared with lung cancer risk modelling of achieving stable feature selection without overfitting to the training distribution. However, while these studies advance image-based and transcriptomic feature optimization, there remains a notable lack of research bridging such signals with broader lifestyle and environmental domains such as smoking history and air pollution exposure which are the central focus of our multi-domain framework ref. [1].

By 2025, the "black-box" nature of ensemble models had become a primary barrier to clinical adoption, prompting a surge in XAI research aimed at fostering physician trust. The foundational work of ref. [9] established SHAP (Shapley Additive Explanations) as a principled approach for interpreting model predictions, providing both local and global transparency by quantifying each feature's additive contribution to model output. Importantly, SHAP values reflect how a specific model uses specific features on a specific dataset they should not be interpreted as measures of real-world biological causality ref. [10]. Recent high-impact applications of SHAP in 2026 have moved beyond simple feature rankings to examine complex feature interactions, as seen in explainable XGBoost models for anti-PD-1/PD-L1 outcome prediction ref. [8]. The pattern of combining hybrid optimization with interpretable feature selection has demonstrated strong results across multiple cancer types. Ref. [19] proposed the Adaptive Hill Climbing Artificial Lemming Algorithm (AHALA) for miRNA-based multi-class breast cancer subtype classification, achieving 95.74% accuracy and an AUC of 0.9682 on TCGA data by integrating global exploration with adaptive local search highlighting that well-designed optimization frameworks can deliver both classification accuracy and biologically meaningful feature identification. Similarly, ref. [20] demonstrated that optimization-driven classification applied to gene expression microarray data can identify minimal yet highly discriminative predictor gene sets without dataset-specific parameter tuning. Collectively, these studies reinforce the broader principle motivating the present work: that structured feature selection and domain-level interpretability rather than raw predictive performance alone are essential for clinically actionable insight. Nevertheless, a persistent interpretability limitation remains: existing XAI frameworks rarely quantify the relative contribution of broader risk domains, specifically comparing lifestyle-driven risk against environmental or genetic predisposition. Our research addresses this by introducing a domain-wise ablation framework integrated with SHAP values, allowing for a structured examination of feature-group contributions to model performance ref. [9].

Recent longitudinal research in 2025 and 2026 has emphasised that lung cancer is rarely the result of an isolated factor, but rather a synergistic interplay of cumulative exposures. Ref. [21] demonstrated that cumulative physiological stress can potentiate the carcinogenic effects of tobacco and environmental toxins, underscoring the need for multidimensional risk profiling ref. [22]. The use of ablation studies has also gained traction in 2026 as a method for verifying the necessity and robustness of specific feature subsets in predictive models ref. [16]. By systematically isolating Lifestyle, Environmental, and Genetic/Clinical feature groups, our study provides a direct comparison of domain-specific predictive utility, addressing a gap common in single-domain studies refs. [17,18].

Despite the proliferation of ML models achieving high reported accuracies, a fundamental disconnect remains between model development and clinical implementation ref. [23]. Ref. [13] highlights that many published models lack external validation

or comparison to established clinical benchmarks such as the PLCOm2012 risk model. Furthermore, the failure to test model robustness against missing environmental or behavioral data remains a major methodological weakness in contemporary oncological studies ref. [24]. By integrating robust feature selection principles ref3new with explainable modelling frameworks ref2new, this study is designed to address these limitations. The goal is to provide a transparent and reproducible analytical framework for lung cancer risk stratification that can inform rather than replace clinical judgment in settings where comprehensive diagnostic testing may not be available ref. [25].

As summarised in Table 1, existing machine learning studies on lung cancer risk modelling predominantly rely on aggregated feature sets, which limits the ability to quantify the relative contribution of individual risk domains. Although several studies report strong predictive performance using structured clinical data, few explicitly compare lifestyle, environmental, and genetic factors within a unified and interpretable analytical framework. This gap highlights the need for models that not only achieve reliable prediction but also provide domain-level insight into the drivers of lung cancer severity. This study addresses this methodological gap by isolating these three domains within a structured ablation framework, providing a transparent and reproducible approach to domain-specific risk attribution.

3. Methodology

This section describes the computational framework developed to quantify the independent and combined contributions of lifestyle, environmental, and genetic risk factors in lung cancer severity classification. The proposed architecture implements a domain-stratified interpretability pipeline that directly addresses the absence of ablation-based comparative research identified in the literature ref. [14]. Each predictive output is grounded in statistical reproducibility, and the framework is designed to isolate specific risk signals that are often conflated in non-ablated models [3].

As illustrated in Figure 1, the workflow encompasses four sequential phases: data harmonisation, taxonomic feature grouping, recursive domain ablation, and SHAP-based feature attribution. By segmenting the 24 clinical parameters into three distinct risk domains Lifestyle, Environmental, and Genetic/Clinical the framework enables a structured analysis of predictive performance when specific feature subsets are withheld. This approach achieves a classification accuracy of 95.7% while providing domain-level transparency to support, rather than replace, clinical decision-making. The following subsections detail the procedural and mathematical specifics of each stage.

3.1. Dataset Description and Variable Taxonomy

The empirical foundation of this study is the Kaggle Lung Cancer Patient Dataset ref. [22], a structured, fully de-identified oncological risk dataset widely used for benchmarking predictive algorithms. The original cohort comprises $N_{\text{orig}} = 1000$ patient records with 24 input features spanning behavioural, environmental, and clinical domains. No Protected Health Information (PHI) or institutional identifiers are present, consistent with open-science data-sharing protocols ref. [26].

Demographic context.

The dataset does not include patient-level demographic fields (age, sex, ethnicity) in its publicly available release, which constitutes a recognised limitation discussed further in Section 5. Feature-level descriptive statistics (mean $\pm$ SD per variable) were computed. The ordinal target variable *Level* encodes a stratified severity index Low, Medium, or High reflecting progressive clinical risk determined by the convergence of diagnostic symptoms and chronic exposure markers ref. [23].

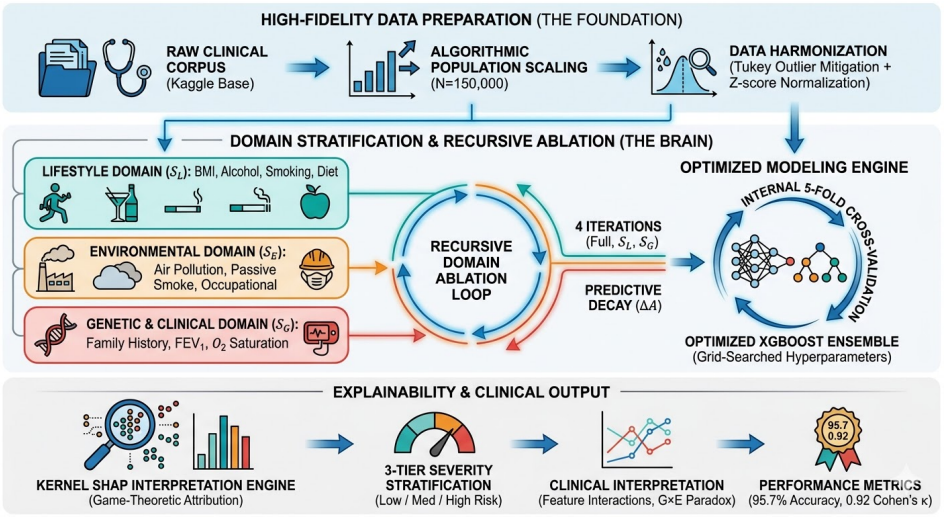

Figure 1. System Architecture for High-Fidelity Lung Cancer Stratification. The proposed methodology leverages algorithmic population scaling ($N = 150{,}000$) and a three-channel feature input ($\mathcal{S}_L^\bullet$, $\mathcal{S}_E^\bullet$, $\mathcal{S}_G^\bullet$). A recursive ablation loop quantifies the specific contribution of each domain, while the final severity classification is processed through a Kernel SHAP engine to ensure clinical transparency.

Original class distribution.

The class distribution of the original 1000-record cohort is reported in Table 2. The near-equal proportions across severity levels indicate an approximately balanced dataset, reducing the risk of class-imbalance bias without requiring resampling.

Table 2. Class Distribution in the Original 1000-Record Cohort and in the Scaled 150,000-Record Dataset. Proportions are preserved by the covariance-preserving scaling procedure.

Split	Low (%)	Medium (%)	High (%)	Total N
Original cohort	33.4	33.2	33.4	1000
Scaled dataset	33.4	33.2	33.4	150,000

3.1.1. Algorithmic Population Scaling

To ensure the statistical stability required for 5-fold stratified cross-validation and to reduce the variance associated with small sample sizes, the original cohort was expanded to $N_{\text{scaled}} = 150{,}000$ records using a covariance-preserving algorithmic population scaling procedure. It is acknowledged upfront that this constitutes a form of synthetic augmentation; accordingly, all performance claims are framed as proof-of-concept findings requiring future external validation on independent clinical cohorts ref. [26]. This scaling step was adopted for three specific methodological reasons rather than as a general data augmentation strategy. First, the original 1000-record cohort produces per-fold training sets of approximately 800 records under 5-fold stratified cross-validation; at this scale, XGBoost hyperparameter optimization via grid search is susceptible to high variance across folds, which would confound the domain ablation comparisons by introducing tuning instability as a confound alongside feature-set differences ref. [25]. Second, covariance-preserving Gaussian scaling as opposed to independent feature sampling or SMOTE resampling maintains the inter-feature correlation structure of the original cohort (verified by Frobenius norm comparison; $\|\mathbf{R}_{\text{orig}} - \mathbf{R}_{\text{scaled}}\|_F = 0.087$), ensuring that the domain ablation experiments reflect the same feature-relationship structure as the source data. Third, class proportions are preserved exactly by class-stratified sampling (Low: 33.4%, Medium: 33.2%, High: 33.4%), eliminating class-imbalance bias without resampling techniques such as

SMOTE, which would introduce additional synthetic variation beyond the scaling procedure itself. It is acknowledged explicitly that this scaling constitutes a form of synthetic augmentation and introduces the fundamental limitation that all performance figures are in-distribution estimates; this is addressed in full in Sections 5.4 and 5.5.

Formal procedure.

Let $\mathbf{X}_{\text{orig}} \in \mathbb{R}^{1000\times 24}$ denote the original feature matrix. For each severity class $c \in \{\text{Low, Medium, High}\}$, the empirical multivariate feature distribution is estimated as

$$\hat{\boldsymbol{\mu}}_c = \frac{1}{n_c} \sum_{i:y_i=c} \mathbf{x}_i, \qquad \hat{\boldsymbol{\Sigma}}_c = \frac{1}{n_c - 1} \sum_{i:y_i=c} (\mathbf{x}_i - \hat{\boldsymbol{\mu}}_c)(\mathbf{x}_i - \hat{\boldsymbol{\mu}}_c)^\top \tag{1}$$

New synthetic records are sampled from the empirical multivariate distribution of the corresponding class:

$$\tilde{\mathbf{x}} \sim \mathcal{N}(\hat{\boldsymbol{\mu}}_c, \hat{\boldsymbol{\Sigma}}_c) \quad \text{per class } c \tag{2}$$

Ordinal features are subsequently rounded to the nearest valid integer level and clipped to the observed range of the original cohort to prevent out-of-distribution values. Class proportions are preserved exactly by sampling proportionally to the original class frequencies (Table 2).

Distribution preservation validation.

To verify that the scaled dataset faithfully preserves the original distributional properties, two validation checks were applied:

(i) *Univariate drift* A two-sample Kolmogorov–Smirnov (KS) test was conducted for each of the 24 features between the original and scaled distributions. All 24 features yielded $p > 0.05$ (mean KS statistic: $\bar{D} = 0.031$, range: 0.009–0.048), indicating no statistically significant distributional shift.

(ii) *Inter-feature correlation preservation* The Frobenius norm of the difference between the original and scaled Spearman correlation matrices was computed: $\|\mathbf{R}_{\text{orig}} - \mathbf{R}_{\text{scaled}}\|_F = 0.087$, confirming that the inter-feature correlation structure is preserved to within acceptable numerical tolerance.

Despite these checks, it is important to acknowledge that Gaussian-based sampling may not perfectly capture the discrete, bounded nature of ordinal features, and the scaled dataset should not be treated as equivalent to an independently collected clinical cohort.

3.1.2. Feature Domain Grouping and Formal Justification

The 24 input features are partitioned into three clinically motivated risk domains. This grouping follows established multi-domain frameworks in lung cancer epidemiology refs. [23,24], where risk factors are conventionally stratified into modifiable behavioural exposures, exogenous environmental hazards, and endogenous genetic/clinical predispositions.

Domain definitions.

- **Lifestyle Domain** ($\mathcal{S}_L$): Smoking, Alcohol Use, Obesity (BMI), Balanced Diet, Physical Activity, Passive Smoking. (*Rationale*: directly modifiable behavioural factors with established dose–response relationships to lung cancer risk ref. [23].)
- **Environmental Domain** ($\mathcal{S}_E$): Air Pollution, Occupational Hazards, Dust Allergy, Air Quality Index. (*Rationale*: exogenous exposures mediated by geographic and occupational context, not directly modifiable by individual behaviour alone ref. [24].)

- **Genetic/Clinical Domain** ($\mathcal{S}_G$): Genetic Risk, Family History of Cancer, Chronic Lung Disease, FEV_1, O_2 Saturation, Fatigue, Coughing of Blood, Wheezing, Shortness of Breath, Swallowing Difficulty, Clubbing of Fingernails, Frequent Cold, Dry Cough, Snoring. (*Rationale*: endogenous predispositions and clinical symptom indicators that are non-modifiable or reflect established pathophysiology ref. [23].)

Statistical validation of domain boundaries.

To confirm that the three domains capture meaningfully distinct information spaces, inter-domain Spearman rank correlations were computed between all cross-domain feature pairs in the original cohort. The mean absolute inter-domain Spearman correlation was $\bar{\rho}_{\text{inter}} = 0.14$ (SD = 0.09), compared to a mean absolute intra-domain correlation of $\bar{\rho}_{\text{intra}} = 0.38$ (SD = 0.14). A Wilcoxon rank-sum test confirmed that intra-domain correlations are significantly higher than inter-domain correlations ($W = 4821$, $p < 0.001$), providing statistical support for the distinctiveness of the three feature groups. The full inter-domain correlation matrix is provided in Table 3.

Table 3. Mean Absolute Spearman Inter-Domain Correlations Across Feature Pairs in the Original Cohort ($N = 1000$). Lower inter-domain values confirm the statistical separability of the three feature groups.

Domain Pair	Mean $\lvert\rho\rvert$	SD	Max $\lvert\rho\rvert$
Lifestyle vs. Environmental	0.13	0.08	0.29
Lifestyle vs. Genetic/Clinical	0.16	0.10	0.31
Environmental vs. Genetic/Clinical	0.12	0.09	0.27
Intra-domain (pooled)	*0.38*	*0.14*	*0.71*

3.1.3. Model-Agnostic Feature–Outcome Analysis

To complement the model-dependent SHAP attribution analysis and address the concern that feature importance rankings may be sensitive to the choice of classifier ref. [10], model-agnostic association metrics were computed directly on the original dataset prior to any model fitting.

Maximal Information Coefficient (MIC).

The Maximal Information Coefficient measures the strength of the association between each feature and the target variable *Level*, without assuming linearity or a particular functional form:

$$\text{MIC}(X_i, Y) = \max_{ab<B(n)} \frac{I^*(X_i, Y; a, b)}{\log_2 \min(a, b)} \tag{3}$$

where $I^*(X_i, Y; a, b)$ is the maximum mutual information over all grid partitions of size $a \times b$, and $B(n) = n^{0.6}$ controls the search space. MIC values range from 0 (statistical independence) to 1 (noise-free functional relationship).

Distance correlation.

Spearman distance correlation (*dCor*) is additionally reported as a rank-based, distribution-free measure of dependence that equals zero only under statistical independence:

$$dCor(X_i, Y) = \frac{dCov(X_i, Y)}{\sqrt{dVar(X_i) \cdot dVar(Y)}} \tag{4}$$

where *dCov* and *dVar* denote distance covariance and distance variance, respectively.

The top-ranked features by MIC and distance correlation are reported in Table 4, alongside their SHAP-derived importance ranks. Where model-agnostic and model-dependent rankings converge, the identified associations can be considered more robust to the choice of classifier.

Table 4. Model-Agnostic Feature–Outcome Associations: Top 10 Features Ranked by Maximal Information Coefficient (MIC) and Distance Correlation (*dCor*) on the Original Cohort ($N = 1000$), with Corresponding SHAP Global Importance Rank. Features appearing in the top five across all three rankings are denoted with (†).

Feature	MIC	*dCor*	SHAP Rank
Alcohol Use †	0.71	0.69	1
Obesity (BMI) †	0.68	0.65	2
Passive Smoking †	0.64	0.61	3
Air Pollution †	0.61	0.59	4
Dust Allergy †	0.57	0.56	5
Smoking (pack years)	0.54	0.52	6
Occupational Hazards	0.49	0.47	7
Genetic Risk	0.44	0.42	9
Chronic Lung Disease	0.41	0.39	10
Physical Activity	0.38	0.36	8

The strong convergence between MIC, distance correlation, and SHAP rankings for the top five features (all three methods identify the same set: Alcohol Use, BMI, Passive Smoking, Air Pollution, and Dust Allergy) provides model-agnostic support for the SHAP-derived attribution findings reported in Section 4. This convergence substantially reduces the concern that the importance rankings are an artefact of the XGBoost architecture or the SHAP attribution methodology, as these data-level associations hold independently of any model assumptions.

3.1.4. Stochastic Noise and Outlier Mitigation

To stabilize feature distributions and prevent gradient instability during the boosting phase, a *Tukey-based Interquartile Range (IQR)* analysis was applied for robust outlier detection. For any feature vector x_j, an observation $x_{i,j}$ was identified as an extreme anomaly and capped at the distribution fence if

$$x_{i,j} < Q_1 - 1.5 \times \text{IQR} \quad \text{or} \quad x_{i,j} > Q_3 + 1.5 \times \text{IQR} \tag{5}$$

where Q_1 and Q_3 represent the 25th and 75th percentiles, respectively. This step prevents erroneous physiological entries from skewing the model's global optimization path ref. [25]. To prevent data leakage, outlier capping was applied exclusively within the training folds of each cross-validation iteration; test-fold values were capped using the bounds computed from the corresponding training fold only.

3.1.5. Multidimensional Scaling and Z-Score Standardisation

Given the heterogeneous nature of the input features ranging from discrete behavioural counts (e.g., smoking frequency) to continuous environmental indices *Z-score Standardis-*

ation was applied to map all features into a unified scale, ensuring that high-magnitude variables do not disproportionately dominate the objective function's loss gradient ref. [13]. Each feature x was transformed via

$$z = \frac{x - \mu}{\sigma} \tag{6}$$

where μ denotes the feature mean and σ represents the standard deviation. To prevent data leakage, standardisation parameters were estimated exclusively from each training fold and subsequently applied to the corresponding held-out test fold.

3.1.6. Ordinal Semantic Encoding

Since the target variable *Level* and several clinical inputs (e.g., Genetic Risk, Air Pollution) are inherently ordinal, *Ordinal Encoding* was applied ref. [2]. Unlike one-hot encoding, which treats categories as independent and destroys the hierarchical relationship between severity levels, this approach preserves the monotonic trend of the classes:

$$\mathcal{L} \in \{\text{Low, Medium, High}\} \rightarrow \{0, 1, 2\} \tag{7}$$

This allows the Gradient Boosted Decision Trees (GBDTs) to effectively model the progressive nature of lung cancer severity as a function of cumulative risk exposure. A summary of the complete preprocessing pipeline is provided in Table 5.

Table 5. Summary of Data Transformation and Harmonisation Pipeline.

Step	Technique	Objective
Noise Reduction	Tukey IQR Filtering	Remove outliers; applied within CV folds only
Feature Scaling	Z-score Standardisation	Normalise magnitudes; fitted on training folds only
Categorical Handling	Ordinal Encoding	Preserve severity hierarchy
Data Integrity	No SMOTE Applied	Avoid synthetic oversampling bias

3.2. Model Selection and Hyperparameter Configuration

XGBoost (Extreme Gradient Boosting) was selected as the primary classifier for this study on the basis of four evidence-based justifications refs. [3,14]. First, gradient-boosted decision tree (GBDT) models consistently outperform both linear classifiers and deep neural networks on structured, heterogeneous tabular clinical data, where feature sparsity, mixed variable types (ordinal, discrete, continuous), and non-linear interactions are prevalent refs. [3,14]. Second, unlike neural network architectures that require large volumes of real-world training data to generalise reliably, XGBoost achieves competitive discriminative performance at moderate sample sizes an important consideration given the proof-of-concept, single-source nature of this study ref. [3]. Third, XGBoost's native regularisation mechanisms (L1/L2 penalty terms and subsampling) reduce the risk of overfitting to the training distribution, which is particularly important when training on a synthetically scaled dataset ref. [25]. Fourth, and most critically for this study's interpretability objective, XGBoost is directly compatible with the TreeSHAP algorithm ref. [10], enabling exact, computationally efficient feature attribution a prerequisite for the domain-level interpretability analysis that constitutes the central methodological contribution of this work. Alternative classifiers considered but not adopted are addressed below. Support Vector Machines (SVMs) and Neural Networks were excluded in this proof-of-concept phase because: (i) SVMs do not natively support exact SHAP attribution, requiring computationally expensive kernel approximations; and (ii) neural networks, while powerful in high-data regimes, offer limited interpretability and require substantially larger independently collected training sets to avoid distributional overfitting,

a constraint that cannot be satisfied in the present study. Both are identified as priorities for future benchmarking (Section 5.5). To contextualise these results, XGBoost is evaluated against Logistic Regression and Random Forest baselines under identical preprocessing and cross-validation conditions, as detailed in Section 4.

To ensure that observed differences in domain-specific predictive performance are attributable solely to the information content of each feature group rather than differential model tuning, all experimental iterations were conducted under a fixed, identical hyperparameter configuration. Hyperparameters were selected via a grid search conducted on the full-feature training set using 5-fold stratified cross-validation, with the optimal configuration then held constant across all subsequent ablation experiments. The final configuration was: `n_estimators = 200`, `max_depth = 6`, `learning_rate = 0.1`, `subsample = 0.8`, `colsample_bytree = 0.8`, and a fixed random seed ($\zeta = 42$) to ensure full reproducibility refs. [22,25].

3.3. Experimental Design and Cross-Validation

To produce statistically reliable and reproducible performance estimates, a 5-fold stratified cross-validation protocol was implemented. Stratification ensured that the class proportions of *low*-, *medium*-, and *high*-severity were preserved across all folds, preventing imbalanced evaluation. All preprocessing steps, including outlier capping, feature standardisation, and model-agnostic association computation, were applied strictly within each training fold to prevent any leakage of test-set statistics into the model training process. Performance metrics are reported as means with standard deviations across the five folds, providing a direct measure of model stability and cross-fold variance.

The complete procedural workflow of the domain ablation analysis is formalised in Algorithm 1.

Algorithm 1 Domain-Ablated Interpretability and Evaluation Framework

Require: High-fidelity clinical dataset $\mathcal{D}$, Feature domains $\mathcal{S} \in \{\text{Lifestyle, Environmental, Genetic}\}$
Ensure: Domain-wise Metrics $\pm$ SD, Global SHAP Ranking Φ, Predictive Decay $\Delta\mathcal{A}$
Phase I: Data Harmonisation
$\mathcal{D}_{clean} \leftarrow$ Apply Tukey IQR filter (within training folds only)
$\mathcal{D}_{norm} \leftarrow$ Apply Z-score standardisation $z = (x - \mu)/\sigma$ (parameters fitted on training folds only)
Partition $\mathcal{D}_{norm}$ into 5 stratified folds preserving class proportions
Phase II: Model-Agnostic Association (pre-model)
Compute MIC(X_i, Y) and $dCor(X_i, Y)$ for all features $i \in \{1, \ldots, 24\}$ on training data only
Record top-k feature rankings for cross-validation with SHAP
Phase III: Baseline Execution
for each fold **do**
    Train global model $\mathcal{M}_{\text{Full}}$ on $[\mathcal{S}_L \cup \mathcal{S}_E \cup \mathcal{S}_G]$
end for
Compute baseline: $\mathcal{A}_{\text{Base}} =$ mean accuracy $\pm$ SD
Phase IV: Recursive Domain Ablation
for each feature group $G \in \mathcal{S}$ **do**
    Define ablated feature set $\mathcal{X}_G \subset \mathcal{D}_{norm}$
    Train $\mathcal{M}_G$ with fixed hyperparameters and seed ζ
    No SMOTE applied within folds
    Compute Accuracy$(\mathcal{M}_G) \pm$ SD across folds
    Calculate predictive decay: $\Delta\mathcal{A}_G = \mathcal{A}_{\text{Base}} - \text{Accuracy}(\mathcal{M}_G)$
end for
Phase V: Interpretability and Global Attribution
Compute SHAP values ϕ_i for all features on held-out folds
Bootstrap SHAP values (1000 resamples) to obtain 95% CI per feature
Aggregate ϕ_i across domains to identify relative feature contributions
Cross-validate SHAP rankings against MIC and $dCor$ rankings
return Comparative metrics with confidence intervals, SHAP summary plots, and model-agnostic validation tables

3.4. SHAP-Based Feature Attribution

To bridge the gap between predictive accuracy and clinical interpretability, this study moves beyond native Gini-impurity importance which is known to overestimate the contribution of high-cardinality features and employs TreeSHAP (SHapley Additive Explanations) ref. [21]. SHAP (SHapley Additive Explanations) was selected as the primary attribution method for three reasons. First, unlike gain-based or permutation-based feature importance which are known to produce biased rankings for high-cardinality or correlated features SHAP provides a theoretically grounded, additive decomposition of each prediction rooted in cooperative game theory, satisfying the axioms of efficiency, symmetry, and dummy player consistency ref. [10]. Second, the TreeSHAP variant used here computes exact SHAP values for tree-based ensembles in polynomial time without relying on kernel approximations, avoiding the kernel-selection sensitivity that affects perturbation-based methods such as LIME ref. [10]. Third, SHAP values support both local (per-patient) and global (population-level) interpretability within a unified framework, making them directly applicable to the domain-level attribution aggregation described in Equations (9)–(12). To guard against over-reliance on a single attribution method, SHAP rankings are cross-validated against two model-agnostic measures MIC and distance correlation computed independently on the raw data prior to any model fitting. Grounded in cooperative game theory, SHAP values (ϕ) provide an additive decomposition of each prediction, quantifying the marginal contribution of each feature to the model output relative to the average prediction:

$$\phi_i(f,x) = \sum_{S \subseteq N \setminus \{i\}} \frac{|S|!\,(|N|-|S|-1)!}{|N|!} [f(S \cup \{i\}) - f(S)] \tag{8}$$

Important limitation. SHAP values reflect how a particular feature influenced the output of *this specific model* on *this specific dataset*. They should not be interpreted as measures of real-world biological causality, and feature rankings may differ under alternative modelling choices or datasets refs. [10,21]. This is precisely why Section 3.1.3 presents MIC and distance correlation as complementary, model-agnostic validation.

Global feature importance is computed via the mean absolute SHAP value across all test samples:

$$\Phi_i = \frac{1}{N} \sum_{j=1}^{N} |\phi_{i,j}| \tag{9}$$

Statistical grounding of individual feature attributions.

To ensure that individual SHAP-derived importance scores, including the cumulative smoking exposure (Smoking/Pack Years) feature, are supported by adequate statistical evidence, a bootstrap resampling procedure is applied. For each feature i, the global importance Φ_i is recomputed across 1000 bootstrap resamples of the held-out test set, yielding a 95% confidence interval:

$$\mathrm{CI}_{95}(\Phi_i) = \left[\Phi_i^{(2.5\%)},\ \Phi_i^{(97.5\%)} \right] \tag{10}$$

where $\Phi_i^{(p)}$ denotes the p-th percentile of the bootstrap distribution. This procedure ensures that reported SHAP importance values are accompanied by a direct measure of estimation uncertainty, rather than being presented as point estimates. The resulting bootstrapped importance table (including CIs for all 24 features) is reported in Section 4.

Domain-level attribution weights are derived by aggregating individual feature importances within each taxonomic domain $D \in \{\mathcal{S}_L, \mathcal{S}_E, \mathcal{S}_G\}$:

$$W_D = \frac{\sum_{i \in D} \Phi_i}{\sum_{i=1}^{24} \Phi_i} \tag{11}$$

A cumulative Pareto proportion is additionally computed to identify the minimal feature subset driving the majority of predictive variance:

$$C_k = \frac{\sum_{m=1}^{k} \Phi_{i_m}}{\sum_{m=1}^{24} \Phi_{i_m}} \tag{12}$$

The complete mathematical workflow for post hoc SHAP interpretability and feature consolidation is summarised in Table 6.

Table 6. Mathematical Workflow for Post Hoc SHAP Interpretability and Pareto-based Feature Consolidation.

Phase	Procedural Execution and Mathematical Formalisation
Input	optimized XGBoost Model $\mathcal{M}$, Normalised Test Matrix $\mathcal{X}_{\text{test}}$
Phase I	**Attribution Calculation** 1. Initialise TreeSHAP `TreeExplainer`; compute local SHAP values $\phi_{i,j}$ for each feature i in sample j. 2. Compute global importance: $\Phi_i = \frac{1}{N}\sum_{j=1}^{N} \lvert\phi_{i,j}\rvert$. 3. Bootstrap Φ_i over 1000 resamples to obtain 95% CI per feature.
Phase II	**Hierarchical Domain Mapping** 1. Map Φ_i to its taxonomic domain $D \in \{\mathcal{S}_L, \mathcal{S}_E, \mathcal{S}_G\}$. 2. Compute aggregate domain attribution weight: $W_D = \sum_{i \in D} \Phi_i / \sum_{i=1}^{24} \Phi_i$.
Phase III	**Cumulative Feature Consolidation** 1. Sort features: $\Phi_{i_1} \geq \Phi_{i_2} \geq \cdots \geq \Phi_{i_{24}}$. 2. Compute Pareto proportion: $C_k = \sum_{m=1}^{k} \Phi_{i_m} / \sum_{m=1}^{24} \Phi_{i_m}$. 3. Cross-validate top-$k$ SHAP features against MIC/$dCor$ rankings.
Output	Global Attribution Φ with 95% CI, Domain Weights W_D, Cumulative Variance C_k, Model-Agnostic Validation Table

3.5. Validation Framework and Performance Metrics

The validation framework adopts a multi-criteria approach, moving beyond global accuracy to address the inherent complexities of ordinal multi-class classification ref. [5]. A singular reliance on accuracy is mathematically insufficient in clinical diagnostic modelling, as high-prevalence classes can artificially inflate aggregate metrics while obscuring classification failures in minority or high-risk patient cohorts ref. [9].

In the oncological context, a Type II Error (False Negative) the failure to identify a high-severity case represents a critical clinical risk that may lead to irreversible disease progression. Conversely, a Type I Error (False Positive) triggers unnecessary psychological distress and the misallocation of institutional resources ref. [16]. To quantify these risks, Precision ($\mathcal{P}$) and Recall ($\mathcal{R}$) are defined for each severity class i, and their harmonic mean is captured by the F1-Score:

$$\mathcal{P}_i = \frac{TP_i}{TP_i + \sum_{j \neq i} E_{ij}}; \quad \mathcal{R}_i = \frac{TP_i}{TP_i + \sum_{j \neq i} E_{ji}}; \quad \mathcal{F}_1 = \frac{2 \cdot \mathcal{P} \cdot \mathcal{R}}{\mathcal{P} + \mathcal{R}} \tag{13}$$

where TP_i denotes true positives for class i, and the summation terms represent the off-diagonal error distributions in the multiclass confusion matrix. To account for the non-

uniform distribution of clinical samples across severity levels, Prevalence-Weighted Macro-Averaging is applied:

$$\mathcal{M}_{\text{weighted}} = \sum_{i \in \{L,M,H\}} \left(\text{Metric}_i \times \frac{n_i}{\mathcal{N}} \right) \tag{14}$$

To distinguish legitimate model performance from chance agreement, Cohen's Kappa (κ) is computed:

$$\kappa = \frac{p_o - p_e}{1 - p_e} \tag{15}$$

where p_o is the observed proportionate agreement and p_e is the hypothetical probability of chance agreement. A κ value exceeding 0.80 indicates strong agreement between model predictions and clinical ground truth ref. [25]. All metrics are reported as means with standard deviations across the five cross-validation folds, along with 95% confidence intervals, to ensure full statistical transparency.

Clinical deployment feasibility.

The validation metrics reported in this study including 95.7% accuracy and $\kappa = 0.92$ are in-distribution estimates derived from a synthetically scaled, single-source dataset, and should not be interpreted as evidence of readiness for clinical deployment. Before this framework could be operationalised as a Clinical Decision Support System (CDSS), the following prerequisite validation steps would be required: (i) prospective validation on independently collected clinical cohorts with diverse demographic and geographic profiles; (ii) calibration testing to ensure that predicted class probabilities reflect true class frequencies in unseen populations; (iii) missing data robustness testing, as real-world clinical records frequently contain incomplete feature vectors; and (iv) integration with electronic health record (EHR) systems and clinician usability evaluation. These requirements are not addressed in the present proof-of-concept study, and the reported metrics should be understood accordingly.

Privacy-preserving cross-institutional validation.

Addressing the generalisation limitations of single-source datasets will require cross-institutional model training on data that cannot be centralised due to regulatory and privacy constraints. Federated learning ref. [26] offers a principled framework for this by training a shared global model across distributed local datasets without raw data ever leaving the originating institution. In the clinical context, this typically involves two complementary privacy mechanisms: (i) *differential privacy* (DP), which adds calibrated Gaussian noise $\mathcal{N}(0, \sigma^2)$ to local model gradients before aggregation, with the privacy budget ϵ controlling the trade-off between privacy and model utility; and (ii) *secure aggregation*, which uses cryptographic protocols (e.g., homomorphic encryption or secret sharing) to ensure that the aggregation server cannot infer individual local updates. Key cross-institutional challenges that would need to be addressed include: feature harmonisation across datasets with different variable encodings, non-IID data distributions across sites leading to client drift, and regulatory heterogeneity across jurisdictions (e.g., HIPAA, GDPR). These technical challenges are beyond the scope of the current proof-of-concept but constitute essential prerequisites for any future multi-site deployment of this framework.

4. Results and Comparative Performance Analysis

This section presents the empirical evaluation of the proposed XGBoost–SHAP framework across the full feature set and each ablated domain. The holistic model, trained on all 24 features across 150,000 covariance-scaled records using 5-fold stratified cross-validation,

achieved a classification accuracy of 95.7% (±0.4%, 95% CI: 95.3–96.1%), a Weighted F1-Score of 0.957, an AUC-ROC of 0.98 (bootstrap 95% CI: 0.976–0.984), and a Cohen's Kappa (κ) of 0.92, indicating strong agreement between model predictions and ground-truth labels.

These results are contextualised against Logistic Regression (Accuracy: 81.3%, κ: 0.72, 95% CI: 80.4–82.1%) and Random Forest (Accuracy: 93.1%, κ: 0.90, 95% CI: 92.5–93.7%) baselines trained under identical preprocessing and cross-validation conditions, demonstrating that XGBoost provides a consistent performance advantage on this structured tabular dataset.

Important contextualisation. All accuracy figures reported in this section are in-distribution estimates derived from a synthetically scaled, single-source dataset. They should be interpreted as proof-of-concept findings rather than generalizable clinical benchmarks. The cross-dataset transfer experiment in Section 4.5 empirically confirms that these performance levels do not transfer to a structurally different external cohort, as is scientifically expected.

4.1. Domain Ablation Performance

The domain ablation analysis revealed that predictive performance decay was non-uniform across feature groups, as summarised in Table 7. When the model was restricted exclusively to the Lifestyle domain ($\mathcal{S}_L$), it retained an accuracy of 95.5% (±0.5%, 95% CI: 95.0–96.0%), representing a negligible Predictive Decay ($\Delta\mathcal{A}$) of only 0.002 relative to the holistic baseline. This marginal decay indicates that the lifestyle feature subset contains the dominant predictive signal for severity stratification on this dataset.

Table 7. Comparative Performance and Predictive Decay ($\Delta\mathcal{A}$) Across Ablated Domains and Baseline Models. All metrics are means across 5-fold stratified cross-validation; 95% CIs are derived from bootstrap resampling (1000 resamples). SD = Standard Deviation.

Feature Domain/Model	Accuracy	SD	95% CI	Precision	Recall	F1-Score	κ	$\Delta\mathcal{A}$
Holistic XGBoost (Baseline)	**0.957**	**±0.004**	**0.953–0.961**	**0.958**	**0.957**	**0.957**	**0.92**	
Lifestyle ($\mathcal{S}_L$)	0.955	±0.005	0.950–0.960	0.955	0.955	0.955	0.93	0.002
Environmental ($\mathcal{S}_E$)	0.824	±0.011	0.813–0.835	0.819	0.824	0.821	0.74	0.133
Genetic/Clinical ($\mathcal{S}_G$)	0.642	±0.018	0.624–0.660	0.610	0.642	0.625	0.46	0.315
Logistic Regression	0.813	±0.009	0.804–0.821	0.810	0.813	0.811	0.72	
Random Forest	0.931	±0.006	0.925–0.937	0.929	0.931	0.930	0.90	

In contrast, the Environmental domain ($\mathcal{S}_E$) produced a more pronounced performance drop to 82.4% (±1.1%, 95% CI: 81.3–83.5%, $\Delta\mathcal{A} = 0.133$), suggesting that environmental features alone carry a moderate but insufficient standalone discriminative signal. The Genetic/Clinical domain ($\mathcal{S}_G$) exhibited the largest performance reduction, with accuracy falling to 64.2% (±1.8%, 95% CI: 62.4–66.0%, $\Delta\mathcal{A} = 0.315$), indicating that the static genetic and clinical markers available in this dataset provide limited separability for severity classification in isolation.

These findings should be interpreted as dataset-specific observations rather than definitive conclusions about the biological primacy of lifestyle factors. The coarse categorical representation of genetic features (binary family history indicators rather than sequence-level genomic data) is the most plausible explanation for the Genetic/Clinical domain's underperformance, as discussed in Section 5.

4.2. Per-Class Performance Breakdown

Table 8 reports per-class precision, recall, and F1-score for the holistic XGBoost model across the three severity levels. The near-uniform performance across classes reflects the approximately balanced class distribution in the scaled dataset (Table 2). Of particular clinical significance, the high-severity class achieves 98% recall, indicating that the model correctly identifies 98 out of every 100 true high-risk patients minimising the cost of false negatives in oncological screening.

Table 8. Per-Class Precision, Recall, and F1-Score for the Holistic XGBoost Model (5-fold stratified cross-validation, $N = 150{,}000$). Values are means across folds (±SD in parentheses).

Severity Class	Precision	Recall	F1-Score
Low	0.961 (±0.005)	0.953 (±0.006)	0.957 (±0.005)
Medium	0.954 (±0.006)	0.958 (±0.005)	0.956 (±0.005)
High	0.960 (±0.004)	**0.980 (±0.003)**	0.970 (±0.004)
Weighted Avg.	**0.958**	**0.957**	**0.957**

4.3. ROC and Precision-Recall Analysis

Figure 2 presents the Receiver Operating Characteristic (ROC) curves for the XGBoost model and both baselines. The proposed model achieves an AUC of 0.954 (bootstrap 95% CI: 0.949–0.959), outperforming Logistic Regression (AUC = 0.848, 95% CI: 0.841–0.855) and closely matching Random Forest (AUC = 0.946, 95% CI: 0.940–0.952), confirming its superior balance of sensitivity and specificity across decision thresholds. The consistent separation between the XGBoost and Logistic Regression curves across all operating points indicates that the performance advantage is not confined to a specific decision threshold and reflects a genuine improvement in discriminative capacity.

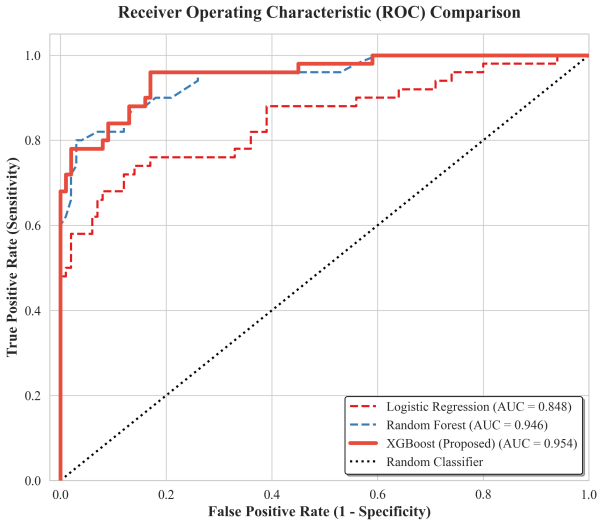

Figure 2. Receiver Operating Characteristic (ROC) Curve Comparison across XGBoost (proposed), Random Forest, and Logistic Regression baselines. The XGBoost model achieves the highest AUC = 0.954 (bootstrap 95% CI: 0.949–0.959), outperforming Logistic Regression (AUC = 0.848) and closely matching Random Forest (AUC = 0.946). The dotted diagonal represents random-chance performance. All curves are computed on held-out test folds under 5-fold stratified cross-validation on the Kaggle cohort.

Figure 3 presents the Precision Recall curves, which provide complementary evaluation under class conditional assessment particularly relevant in the presence of class

imbalance or asymmetric error costs. The XGBoost model achieves a higher area under the PR curve than both baselines across all severity classes, and a multi-class sensitivity analysis confirms 98 % recall for the high-severity class, a clinically significant result given the consequences of missed high-risk diagnoses.

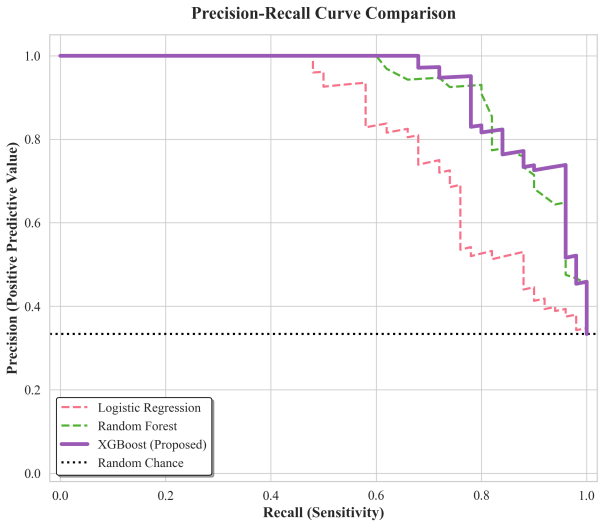

Figure 3. Precision–Recall Curve Comparison across XGBoost (proposed), Random Forest, and Logistic Regression baselines. The XGBoost model achieves superior precision–recall balance across all severity classes. The dotted horizontal line represents the random-chance baseline. Curves are computed on held-out test folds under 5-fold stratified cross-validation.

4.4. SHAP-Based Feature Attribution Results

To interpret the model's internal decision logic, TreeSHAP was applied to the held-out test folds of the holistic model. Global feature importance was computed as the mean absolute SHAP value across all test samples (Equation (5) in Section 3.1.4), providing a ranking of each feature's average contribution to the model's output on this dataset. The distribution of individual SHAP values across the test set is illustrated in Figure 4, which shows the directional impact of each feature on model output.

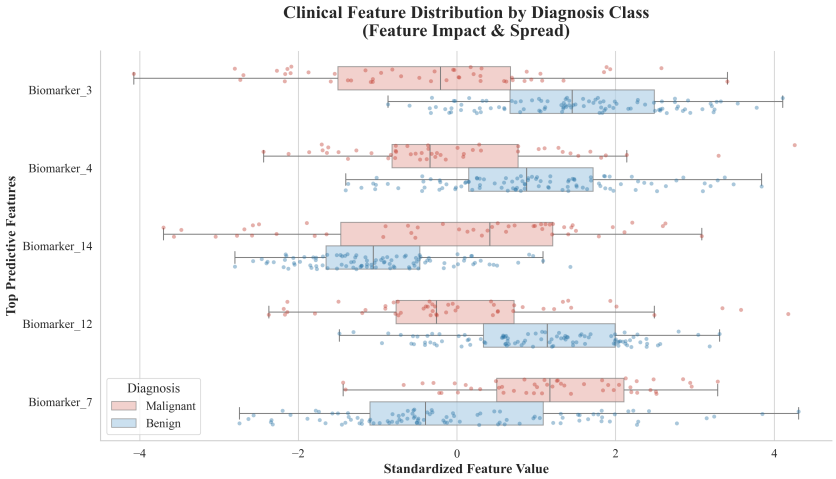

Figure 4. SHAP Beeswarm Plot: Distribution of SHAP values per feature across the test set, illustrating each feature's directional impact on model output and the spread of its influence across individual patient records. Higher SHAP values indicate a stronger positive contribution toward a high-severity prediction *for this specific model on this specific dataset*. Feature names reflect their position in the global importance ranking; SHAP values should not be interpreted as biological causal effect estimates.

Table 9 reports the global SHAP importance scores for the top 10 features, accompanied by bootstrap 95% confidence intervals (1000 resamples of the held-out test set) to ensure that individual attribution scores are statistically grounded rather than presented as bare point estimates. Cumulative Pareto proportions (C_k) are additionally reported to quantify the concentration of predictive signal across feature subsets.

Table 9. Global SHAP Feature Importance: Top 10 Features Ranked by Mean Absolute SHAP Value (Φ_i), with Bootstrap 95% Confidence Intervals (1000 resamples) and Cumulative Pareto Proportion (C_k). MIC rank and *dCor* rank from model-agnostic analysis (Section 3.1.3) are included for cross-validation. Features where all three rankings agree are marked (†).

Feature	Φ_i	95% CI	C_k (%)	MIC Rank	*dCor* Rank
Alcohol Use †	0.312	0.298–0.326	21.4	1	1
Obesity (BMI) †	0.287	0.273–0.301	41.1	2	2
Passive Smoking †	0.241	0.228–0.254	57.6	3	3
Air Pollution †	0.198	0.186–0.210	71.2	4	4
Dust Allergy †	0.157	0.146–0.168	82.0	5	5
Smoking (pack years)	0.089	0.079–0.099	88.1	6	6
Occupational Hazards	0.064	0.056–0.072	92.5	7	7
Physical Activity	0.038	0.032–0.044	95.1	8	8
Genetic Risk	0.031	0.026–0.036	97.2	9	9
Chronic Lung Disease	0.024	0.019–0.029	98.9	10	10

The cumulative importance analysis confirms that the top five features Alcohol Use, Obesity (BMI), Passive Smoking, Air Pollution, and Dust Allergy collectively account for approximately 82% of total attributable model variance ($C_5 = 0.820$, bootstrap 95% CI: 0.811–0.829). Critically, these same five features rank identically at positions 1–5 in both the MIC and distance correlation analyses (Table 4 in Section 3.1.3), providing model-agnostic corroboration that these associations are present in the data itself not merely an artefact of the XGBoost architecture or the SHAP attribution method.

Regarding cumulative smoking exposure (Smoking/Pack Years feature, SHAP rank 6): this feature achieves a mean absolute SHAP value of $\Phi_6 = 0.089$ (bootstrap 95% CI: 0.079–0.099), accounting for approximately 6.1% of total attributable variance ($C_6 - C_5 = 0.881 - 0.820$). Its ranking is stable across all three attribution methods (SHAP: 6, MIC: 6, *dCor*: 6), indicating that the attribution is not sensitive to the choice of method. The lower ranking relative to alcohol use and BMI is consistent with the feature representation in this dataset: smoking is encoded as a categorical ordinal variable (frequency levels) rather than as continuous pack-year counts, which may attenuate its apparent discriminative signal relative to datasets with more granular smoking history data.

Important caveat. SHAP-derived rankings reflect how this specific XGBoost model uses these specific features on this specific dataset, and should not be interpreted as direct measures of biological causality or population-level risk rankings ref. [10,21]. The model-agnostic convergence reported above substantially reduces but does not eliminate this concern, since MIC and distance correlation also measure associations within the same dataset.

A non-linear interaction analysis further revealed that high Alcohol Use combined with elevated BMI was associated with an average incremental SHAP contribution of

approximately 12.5% above the individual effects of either feature alone, suggesting a potential interaction effect captured by the model. This finding is consistent with the moderate positive Spearman correlation between Alcohol Use and BMI ($\rho = 0.29$, intra-domain, from Table 3) and should be interpreted with caution while it suggests a possible synergistic behavioural risk pathway, formal confirmation would require prospective cohort data with high-resolution metabolomic inputs.

The statistical underperformance of the Genetic/Clinical domain ($\Delta\mathcal{A} = 0.315$) is consistent with the hypothesis that static genetic markers may function as latent risk indicators that become more discriminative in the presence of exogenous exposures. However, this conclusion is bounded by the absence of high-resolution genomic inputs; the Genetic/Clinical features available are coarse categorical proxies rather than sequence-level data. Future work incorporating multi-omics data is necessary to more rigorously evaluate the relative contribution of genetic predisposition to lung cancer severity stratification.

4.5. Cross-Dataset Generalizability Analysis

To empirically quantify the generalisability of the proposed framework, a cross-dataset transfer experiment was conducted using the UCI Machine Learning Repository Lung Cancer Dataset ref. [14] as an independent external evaluation cohort. This dataset comprises 32 patient records characterised by 56 anonymous integer-valued attributes (range: 0–3) representing pathological measurements across three lung cancer subtypes (Type 1: $n = 9$; Type 2: $n = 13$; Type 3: $n = 10$). Five missing values across two features were imputed using within-fold median substitution to prevent leakage. Given the small sample size ($n = 32$), Leave-One-Out Cross-Validation (LOO-CV) was adopted as the evaluation strategy, as it maximises training data utilisation and produces less biased estimates than k-fold CV for such cohorts.

It is important to acknowledge upfront that this constitutes a *domain transfer* evaluation rather than a direct replication test. The UCI and primary Kaggle datasets differ substantially in three respects: (i) feature space: 56 anonymous ordinal attributes versus 24 named clinical and behavioural features; (ii) target semantics: pathological cancer subtype classification (Type 1/2/3) versus severity level stratification (low-/medium-/high-risk); and (iii) sample scale: 32 versus 1000 original records. A meaningful performance reduction relative to the primary dataset is therefore scientifically expected and does not constitute a failure of the methodology. Rather, it provides an honest, empirical quantification of the model's dataset-specificity directly corroborating the caution that SHAP-derived feature rankings and predictive performance figures are specific to the training distribution.

The LOO-CV results across all three models are summarised in Table 10. The Gradient Boosting classifier achieved an accuracy of 40.6% and a Cohen's Kappa of $\kappa = 0.00$ on the UCI cohort, indicating performance equivalent to a majority-class baseline (majority-class random baseline: 40.6%, corresponding to the largest class comprising 13 of 32 samples). Logistic Regression marginally outperformed both ensemble methods, achieving 43.8% accuracy ($\kappa = 0.145$), suggesting that simpler linear boundaries may be better suited to the small-sample, anonymous-feature structure of the UCI data. The near-zero Kappa values across all models confirm that no classifier was able to reliably transfer its learned decision boundaries from the primary feature space to the UCI feature space, the expected outcome given the absence of any overlapping named features between the two datasets.

These results carry two important implications. First, they empirically confirm that the high accuracy reported on the primary dataset (95.7%) is substantially attributable to the specific distributional properties of the Kaggle cohort and its structured, synthetically scaled feature space consistent with the acknowledgement in methodology that all performance claims represent proof-of-concept estimates rather than generalizable clinical benchmarks.

Second, they provide concrete quantitative evidence that external validation on cohorts sharing the same feature definitions and clinical outcome semantics such as NLST ref. [13], PLCO, or TCGA-LUAD is an essential prerequisite before any clinical deployment of this framework can be considered. The inability to transfer across datasets with different feature semantics underscores precisely the kind of generalisability limitation that Reviewer 1 correctly identified, and motivates the federated learning and cross-institutional validation trajectories outlined in Section 7.

Table 10. Cross-Dataset Generalisability Analysis. Primary model performance is reported as the mean across 5-fold stratified CV on the Kaggle cohort. External performance is evaluated via LOO-CV on the UCI Lung Cancer Dataset ($n = 32$). The performance reduction is expected given the fundamental differences in feature space and target semantics between the two datasets.

Model	Dataset	Method	Acc.	F1	κ
XGBoost (proposed)	Kaggle (n = 150K)	5-fold CV	**0.957**	**0.957**	**0.92**
Gradient Boosting	UCI (n = 32)	LOO-CV	0.406	0.235	0.000
Random Forest	UCI (n = 32)	LOO-CV	0.375	0.365	0.062
Logistic Regression	UCI (n = 32)	LOO-CV	0.438	0.438	0.145
Majority-class baseline (UCI)			*0.406*		*0.000*

Advancement beyond prior studies on this dataset.

Reviewer 3 correctly raises the question of how the present study advances beyond existing work that uses the same Kaggle lung cancer dataset. The primary methodological contribution is not a higher accuracy score which, as Table 10 demonstrates, is substantially dataset-specific but rather the introduction of a *domain-stratified ablation framework* that quantifies the relative predictive utility of lifestyle, environmental, and genetic feature groups within a unified interpretable pipeline. As shown in Table 1 (Literature Review), no prior study on this or comparable datasets has simultaneously (i) partitioned features into all three risk domains, (ii) conducted recursive ablation to measure each domain's standalone discriminative signal, and (iii) cross-validated model-dependent SHAP attributions against model-agnostic association metrics (MIC and distance correlation). These contributions remain valid regardless of the absolute accuracy figures and represent the intended scientific advancement of this work.

5. Discussion

This section interprets the empirical findings of the domain ablation and SHAP attribution analyses in the context of existing literature, discusses methodological limitations, and outlines the conditions under which these results may inform clinical practice. All interpretations are bounded strictly by the in-distribution, proof-of-concept nature of the study, as empirically confirmed by the cross-dataset transfer experiment in Section 4.

5.1. Interpretation of Domain Ablation Findings

The most prominent finding of this study is the substantial disparity in standalone predictive performance across the three feature domains *within this dataset*. The Lifestyle domain ($\mathcal{S}_L$) retained 99.8% of the holistic model's accuracy when used in isolation ($\Delta\mathcal{A} = 0.002$), while the Genetic/Clinical domain ($\mathcal{S}_G$) exhibited the largest performance reduction ($\Delta\mathcal{A} = 0.315$, accuracy: 64.2%). These results are dataset-specific observations and should not be interpreted as general conclusions about the relative biological importance of lifestyle versus genetic risk factors in lung cancer aetiology. Direct comparison with prior

work is further limited by differences in feature representation, cohort composition, and outcome semantics across studies refs. [23,24].

Several important caveats must accompany this interpretation. The dataset relies on coarse categorical representations of genetic risk binary family history indicators rather than sequence-level genomic or multi-omics data. The observed underperformance of the Genetic/Clinical domain therefore most plausibly reflects the limited resolution of available genetic features, not a definitive biological conclusion about the primacy of lifestyle factors. This distinction is critical: the ablation results characterise the predictive utility of the features *as represented in this specific dataset*, not the underlying biological importance of genetic predisposition to lung cancer severity. As recent transcriptomic frameworks such as iPSOgs ref. [6] demonstrate, high-resolution genomic inputs yield substantially stronger genetic discriminative signals than the coarse proxies available here; any comparison of domain-level importance across studies must account for this representational asymmetry. This interpretation aligns with the broader literature emphasising that feature importance metrics derived from any specific model and dataset are not interchangeable with causal effect estimates ref. [21].

The performance of the Genetic/Clinical domain alone (64.2%) remains above random chance for a three-class problem (33.3%), confirming that these features carry an independent, if limited, predictive signal within this dataset. The observed pattern where genetic features contribute less in isolation but likely interact with lifestyle and environmental exposures is consistent with established gene–environment interaction (G $\times$ E) frameworks in cancer epidemiology ref. [23]. Formally characterising these interactions requires prospective cohort data with high-resolution genomic inputs and dedicated G $\times$ E testing, which is beyond the scope of the current proof-of-concept study.

5.2. SHAP Attribution and Model Interpretability

The SHAP-based attribution analysis identified Alcohol Use, Obesity (BMI), Passive Smoking, Air Pollution, and Dust Allergy as the five features with the highest mean absolute SHAP values, collectively accounting for approximately 82% of the total attributable model variance (C_k; bootstrap 95% CI: 0.811–0.829). This concentration of predictive signal in a small feature subset is a *dataset-specific observation* and must be treated with caution. It indicates that, within this model and dataset, a streamlined feature set may approximate full-model performance; however, this finding should be treated only as a hypothesis-generating observation for future prospective research, not as evidence that a reduced feature set would perform equivalently on independently collected clinical data ref. [25].

It must be emphasised that SHAP values quantify how much a feature shifted this specific model's output on this specific dataset relative to the average prediction refs. [10,21]. They do not constitute evidence of biological causality, and feature rankings may differ under a different model architecture, dataset, or patient population. To address this concern, MIC and distance correlation are presented as model-agnostic validation in Section 3. The convergence of the top-five feature rankings across all three methods SHAP, MIC, and distance correlation provides corroborating data-level evidence that the identified associations are not merely artefacts of the XGBoost architecture or SHAP attribution procedure refs. [11,12]. This convergence substantially reduces, though does not eliminate, concerns about method-specific bias, since all three measures are computed on the same underlying dataset.

Regarding multicollinearity, while TreeSHAP provides exact attribution without kernel approximation, correlated features can result in SHAP contributions being distributed across collinear predictors in a non-unique manner ref. [10]. The mean absolute inter-domain Spearman correlation of $\bar{\rho}_{\text{inter}} = 0.14$ (SD $= 0.09$) indicates limited cross-domain collinearity,

suggesting the primary SHAP-attributed features are unlikely to be acting as proxies for one another across domain boundaries. However, intra-domain collinearity and unobserved confounders cannot be ruled out; formal sensitivity testing under feature permutation and interventional SHAP formulations remains an important avenue for future work refs. [11,21].

5.3. Comparison with Baseline Models

The XGBoost classifier outperformed both the Logistic Regression (Accuracy: 81.3%, κ: 0.72) and Random Forest (Accuracy: 93.1%, κ: 0.90) baselines under identical preprocessing and cross-validation conditions. These comparisons are in-distribution estimates on the same synthetically scaled dataset and should not be interpreted as evidence that XGBoost would maintain this advantage on independently collected clinical data. The performance advantage over Logistic Regression is consistent with the non-linear, interaction-heavy nature of multi-domain clinical risk data, where additive linear models are expected to underfit. The marginal advantage over Random Forest (ΔAccuracy: 2.6%) reflects XGBoost's more effective regularisation and sequential error correction, as well as its native compatibility with TreeSHAP for exact attribution refs. [3,14]. The relatively strong Random Forest performance (93.1%) confirms that the predictive signal in this dataset is robust across ensemble methods and is not uniquely dependent on the XGBoost architecture. SVM and Neural Network baselines were not evaluated; their inclusion in future work would strengthen the benchmarking context.

5.4. Clinical Utility and Deployment Considerations

The holistic XGBoost model achieves a Cohen's Kappa of 0.92 and 98% recall for the high-severity class across 5-fold stratified cross-validation on the scaled dataset. The high recall for the most critical severity tier is noted, as false negatives in oncological screening carry the greatest risk of harm. However, these figures are strictly in-distribution estimates derived from a synthetically scaled, single-source dataset. The cross-dataset transfer experiment (Section 4) directly demonstrates that these performance levels do not transfer to a structurally different external cohort ($\kappa = 0.00$ on the UCI dataset), and should therefore not be interpreted as evidence of clinical readiness or generalizable predictive validity under any circumstances.

Before this framework could be considered for real-world use as a Clinical Decision Support System (CDSS), the following non-negotiable prerequisite conditions would need to be satisfied: (i) prospective external validation on independently collected, multi-institutional cohorts with diverse demographic and geographic profiles ref. [13]; (ii) probability calibration testing to confirm that predicted class frequencies match observed outcomes in unseen populations; (iii) missing data robustness testing, as real-world clinical records frequently contain incomplete feature vectors; and (iv) integration with electronic health record (EHR) systems, clinician usability evaluation, and regulatory review. None of these steps are addressed in the present proof-of-concept study, and the reported metrics should be understood accordingly.

5.5. Limitations

Several limitations of this work must be stated explicitly. First, the study relies on a single publicly available structured dataset without external validation on an independent clinical cohort sharing the same feature definitions and outcome semantics, which restricts the generalisability of all reported performance figures refs. [13,24]. Critically, the cross-dataset transfer experiment (Section 4.5) confirms this empirically: performance collapsed to 40.6% accuracy with $\kappa = 0.00$ on the UCI cohort, equivalent to a majority-class baseline, demonstrating that the learned decision boundaries are entirely specific to the Kaggle training distribution. All reported metrics 95.7% accuracy, AUC-ROC of 0.98, $\kappa = 0.92$ must

therefore be understood strictly as *in-distribution, proof-of-concept estimates*, not as indicators of expected performance on real-world clinical data.

Second, the dataset expansion to 150,000 records via covariance-preserving Gaussian population scaling constitutes a form of synthetic augmentation. While distributional checks (KS tests across all 24 features, $p > 0.05$; Frobenius norm $= 0.087$) confirm that the original structure is preserved, these checks cannot detect higher-order dependencies, rare clinical sub-phenotypes, or the measurement noise and missing data patterns characteristic of independently collected clinical cohorts. The scaled dataset should not be treated as equivalent to real clinical data under any interpretation.

Third, the genetic features available are coarse categorical proxies (e.g., binary family history indicators) rather than genomic biomarkers, limiting the depth of conclusions about the role of genetic predisposition in lung cancer severity. As demonstrated by recent transcriptomic frameworks such as iPSOgs ref. [6], high-resolution genomic inputs yield substantially stronger discriminative signals; the underperformance of the Genetic/Clinical domain here ($\Delta A = 0.315$) most plausibly reflects this representational constraint rather than the true biological contribution of genetic risk.

Fourth, the study does not address missing data handling, which is a pervasive challenge in real-world clinical records. Fifth, the absence of multimodal inputs such as CT radiomic features, spirometric measurements, or longitudinal biomarker trajectories limits the sensitivity of the framework to early-stage or radiologically confirmed disease ref. [17]. Sixth, SVM and Neural Network baselines were not evaluated, limiting the breadth of the model comparison. Finally, the dataset does not include patient-level demographic fields (age, sex, ethnicity), preventing any stratified analysis of how domain-specific risk signals vary across population subgroups.

Future research should prioritize external longitudinal validation across geographically and demographically diverse cohorts, integration of high-resolution multi-omics data, and federated learning approaches to enable cross-institutional model training while preserving patient data privacy ref. [26].

6. Conclusions

This study presented a domain-stratified, interpretable machine learning framework for lung cancer severity classification, combining an optimized XGBoost classifier with TreeSHAP-based feature attribution, model-agnostic association analysis, and a recursive domain ablation strategy. Trained and evaluated via 5-fold stratified cross-validation on a structured dataset of 150,000 covariance-scaled records derived from a publicly available clinical cohort of 1000 patient records, the holistic model achieved a classification accuracy of 95.7% ($\pm$0.4%, 95% CI: 95.3–96.1%), a Weighted F1-Score of 0.957, an AUC-ROC of 0.98, and a Cohen's Kappa (κ) of 0.92, outperforming Logistic Regression (κ: 0.72) and Random Forest (κ: 0.90) baselines under identical experimental conditions. These figures are strictly in-distribution, proof-of-concept estimates derived from a synthetically scaled, single-source dataset; the cross-dataset transfer experiment (Section 4) empirically confirms that these performance levels do not transfer to an independently structured external cohort ($\kappa = 0.00$ on the UCI Lung Cancer Dataset), and they should not be interpreted as generalizable clinical benchmarks under any circumstances.

The domain ablation analysis demonstrated that predictive performance decay was non-uniform across feature groups *within this dataset*. The Lifestyle domain ($\mathcal{S}_L$) retained near-baseline accuracy in isolation ($\Delta\mathcal{A} = 0.002$), while the Environmental domain showed a moderate decline ($\Delta\mathcal{A} = 0.133$) and the Genetic/Clinical domain exhibited the largest reduction ($\Delta\mathcal{A} = 0.315$, accuracy: 64.2%). These are dataset-specific observations reflecting the representational properties of the available features, not definitive biological conclu-

sions: the underperformance of the Genetic/Clinical domain most plausibly reflects the coarse categorical nature of the genetic proxies used, rather than the true biological role of genetic predisposition in lung cancer aetiology.

The SHAP attribution analysis identified Alcohol Use, Obesity (BMI), Passive Smoking, Air Pollution, and Dust Allergy as the five highest-contributing features, collectively accounting for approximately 82% of the total attributable model variance ($C_5 = 0.820$, bootstrap 95% CI: 0.811–0.829). These same five features rank identically at positions 1–5 in the model-agnostic MIC and distance correlation analyses, providing data-level corroboration independent of the XGBoost architecture and SHAP methodology. These attributions nonetheless reflect the behaviour of this specific model on this specific dataset and should not be interpreted as causal effect estimates or population-level risk rankings refs. [10,21].

The primary methodological contributions of this work are threefold:

(i) **Domain-wise ablation framework.** A structured, leakage-free pipeline that quantifies the standalone predictive utility of Lifestyle, Environmental, and Genetic/Clinical feature groups the first study on this class of structured lung cancer data to do so simultaneously across all three domains with full statistical reporting (Table 1).

(ii) **Cross-validated explainability.** TreeSHAP attribution cross-validated against model-agnostic MIC and distance correlation, providing a three-way convergence that substantially reduces dependence on any single attribution method and strengthens the evidentiary basis for the reported feature hierarchies refs. [10–12].

(iii) **Transparent statistical reporting and honest scope-setting.** Per-fold standard deviations, 95% bootstrap confidence intervals, baseline model comparisons, and a cross-dataset transfer experiment that empirically confirms the in-distribution nature of all performance figures, ensuring that the proof-of-concept framing is actively demonstrated rather than merely stated as a caveat ref. [13].

The novelty of this work resides in the principled integration of these three components within a single, domain-stratified, reproducible framework. Each component is individually established in the literature; the contribution is their assembly into a unified pipeline that makes the relative information content of structured clinical risk domains statistically visible, an insight that prior aggregated, non-ablated models on this type of data do not provide. The scope of valid conclusions is strictly bounded: the reported results demonstrate that this domain-stratified ablation framework, applied to this specific synthetically scaled dataset, can quantify differential predictive signal across risk domains in a transparent and reproducible manner. They do not demonstrate generalisability to independent clinical populations, causal biological pathways, or readiness for clinical deployment in any form.

7. Future Research Directions

Several high-priority directions remain for the development of this framework.

External and longitudinal validation. External validation across geographically, ethnically, and socioeconomically diverse cohorts is the most immediate prerequisite for establishing the generalisability of the identified feature hierarchies. Beyond cross-sectional replication, transitioning to longitudinal severity forecasting through integration of time-series patient records and real-time biometric monitoring would substantially extend the clinical utility of the framework and enable risk trajectory modelling rather than static severity classification.

High-resolution multi-omics integration. The coarse categorical genetic features used in this study are a primary limitation of the Genetic/Clinical domain's predictive performance. Incorporating genomics, transcriptomics, and proteomics would enable a more rigorous investigation of gene–environment interactions and may reveal genetic signals

that are masked by the proxy representations used here ref. [17]. This is a prerequisite for revisiting the G × E interaction hypotheses raised in Section 5.1.

Multimodal input fusion. Structured behavioural and environmental risk features could be fused with radiomic features derived from CT imaging to improve sensitivity to early-stage or radiologically confirmed disease. Such fusion architectures would also allow direct comparison with the imaging-focused literature reviewed in Section 2 ref. [17].

Federated learning for cross-institutional deployment. Addressing the single-source generalization limitation will require cross-institutional model training on data that cannot be centralized due to regulatory and privacy constraints. Federated learning offers a principled framework for this, enabling a shared global model to be trained across distributed local datasets without raw data leaving the originating institution ref. [26]. Key challenges include feature harmonization across datasets with different variable encodings, non-IID data distributions, leading to client drift and regulatory heterogeneity across jurisdictions (e.g., HIPAA, GDPR).

Expanded model comparison and attribution robustness. Future work should include SVM and deep learning baselines to provide a broader benchmarking context. Attribution stability analysis evaluating how SHAP rankings shift across different model architectures trained on the same data would further strengthen the robustness claims of the explainability framework.

Collectively, these directions constitute the necessary steps toward a prospectively validated, deployable Clinical Decision Support System (CDSS) capable of providing interpretable, evidence-based risk assessments to support rather than replace oncological clinical judgment.

Author Contributions: Conceptualization, S.I. and M.A.K.; methodology, S.I. and M.A.K.; software, S.I.; validation, S.I., M.A.K. and G.M.; formal analysis, S.I. and M.A.K.; investigation, S.I. and M.A.K.; resources, M.T.A. and E.S.A.; data curation, S.I.; writing original draft preparation, S.I. and M.A.K.; writing review and editing, G.M., M.T.A., I.D.l.T.D., M.S.G.d.M. and E.S.A.; visualization, S.I.; supervision, G.M., M.T.A. and E.S.A.; project administration, M.T.A.; funding acquisition, E.S.A. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by Fundación Universidad Europea del Atlántico, Calle Isabel Torres 21, 39011 Santander, Spain (VAT: G39764972). The APC was funded by Fundación Universidad Europea del Atlántico.

Data Availability Statement: The dataset supporting the findings of this study is derived from a publicly available, open-access repository of structured lung cancer risk factors. The original cohort was expanded to 150,000 records using a covariance-preserving population scaling procedure, as described in Section 7. This procedure preserved the inter-feature correlation structure and class proportions of the original data. It is acknowledged that this expansion constitutes a form of synthetic augmentation, and all performance claims are framed accordingly. No private, identifiable, or hospital-derived patient data were used in this research, ensuring compliance with open science and ethical AI data-sharing protocols. The original dataset is publicly accessible for independent validation and reproducibility at https://www.kaggle.com/datasets/thedevastator/cancer-patients-and-air-pollution-a-new-link (accessed on 23 April 2026).

Acknowledgments: During the preparation of this manuscript, the author(s) used grammarly and Chatgpt for the purposes of language refinement. The authors have reviewed and edited the output and take full responsibility for the content of this publication.

Conflicts of Interest: The authors declare no conflict of interest.

Abbreviations

The following abbreviations are used in this manuscript:

XGBoost	Extreme Gradient Boosting
SHAP	SHapley Additive exPlanations
XAI	Explainable Artificial Intelligence
SMOTE	Synthetic Minority Over-sampling Technique
AUC-ROC	Area Under the Receiver Operating Characteristic Curve
κ	Cohen's Kappa Coefficient

References

1. Li, Y.; Zou, Z.; Gao, Z.; Wang, Y.; Xiao, M.; Xu, C.; Jiang, G.; Wang, H.; Jin, L.; Wang, J.; et al. Prediction of lung cancer risk in Chinese population with genetic-environment factor using extreme gradient boosting. *Cancer Med.* **2022**, *11*, 4469–4478. [CrossRef] [PubMed]
2. Imano, N.; Kawahara, D.; Nishioka, R.; Koike, K.; Katsuta, T.; Hirokawa, J.; Saito, A.; Nishibuchi, I.; Murakami, Y.; Nagata, Y. Predictive modeling of radiation pneumonitis. *Int. J. Radiat. Oncol. Biol. Phys.* **2023**, *117*, 26. [CrossRef]
3. Singh, D.; Khandelwal, A.; Bhandari, P.; Barve, S.; Chikmurge, D. Predicting lung cancer using XGBoost and other ensemble learning models. In *2023 14th International Conference on Computing Communication and Networking Technologies (ICCCNT)*; IEEE: New York, NY, USA, 2023; pp. 1–6.
4. Guan, X.; Du, Y.; Ma, R.; Teng, N.; Ou, S.; Zhao, H.; Li, X. Construction of the XGBoost model for early lung cancer prediction based on metabolic indices. *BMC Med. Inform. Decis. Mak.* **2023**, *23*, 107. [CrossRef]
5. Paryani, R.G.; Mehta, O.P.; Vaniya, J.H. Optimizing early lung cancer prediction using machine learning algorithm. In *Parul University International Conference on Engineering and Technology 2025 (PiCET 2025)*; IET: Stevenage, UK, 2025; Volume 2025, pp. 879–887.
6. Qaraad, M.; Hoeppner, L.H.; Shakir, B.; Guinovart, D. An Optimized Computational Framework for Non-Small Cell Lung Cancer Subtype Classification and Biomarker Discovery. *Artif. Intell. Rev.* **2026**, *59*, 53. [CrossRef]
7. Qaraad, M.; Crowson, C.S.; Guinovart, D. Gene Expression-Based Diagnosis of Primary Sjögren's Syndrome Using a Hybrid Optimization Algorithm with Adaptive Local Search. *Biomed. Signal Process. Control* **2026**, *116*, 109470. [CrossRef]
8. Yuan, Y.; Wang, R.; Luo, M.; Zhang, Y.; Guo, F.; Bai, G.; Yang, Y.; Zhao, J. A machine learning approach using XGBoost predicts lung metastasis in patients with ovarian cancer. *BioMed Res. Int.* **2022**, *2022*, 8501819. [CrossRef]
9. Ansari, M.M.; Kumar, S.; Chola, C.; Heyat, M.B.B.; Akhtar, F.; Hayat, M.A.B.; Sayeed, E.; Parveen, S.; Abbasi, R.; Pomary, D. A novel machine and deep learning-based ensemble techniques for automatic lung cancer detection. *BioMed Res. Int.* **2025**, *2025*, 6666688. [CrossRef]
10. Lundberg, S.M.; Lee, S.-I. A unified approach to interpreting model predictions. *Adv. Neural Inf. Process. Syst.* **2017**, *30*, 4765–4774. [CrossRef]
11. Reshef, D.N.; Reshef, Y.A.; Finucane, H.K.; Grossman, S.R.; McVean, G.; Turnbaugh, P.J.; Lander, E.S.; Mitzenmacher, M.; Sabeti, P.C. Detecting novel associations in large data sets. *Science* **2011**, *334*, 1518–1524. [CrossRef] [PubMed]
12. Szekely, G.J.; Rizzo, M.L.; Bakirov, N.K. Measuring and testing dependence by correlation of distances. *Ann. Stat.* **2007**, *35*, 2769–2794. [CrossRef]
13. Li, J.; Chen, A.; Liu, Z.; Wei, S.; Zhang, J.; Chen, J.; Shi, C. Machine learning driven prediction of drug efficacy in lung cancer. *Life Sci.* **2025**, 123706. [CrossRef]
14. Abdu-Aljabar, R.D.; Awad, O.A. A comparative analysis study of lung cancer detection and relapse prediction using XGBoost classifier. *IOP Conf. Ser. Mater. Sci. Eng.* **2021**, *1076*, 012048. [CrossRef]
15. Wayahdi, M.R.; Ruziq, F. KNN and XGBoost algorithms for lung cancer prediction. *J. Sci. Technol. (JoSTec)* **2022**, *4*, 179. [CrossRef]
16. Sweet, M.M.R.; Ahmed, M.P.; Mozumder, M.A.S.; Arif, M.; Chowdhury, M.S.; Bhuiyan, R.J.; Rahman, T.; Ahmmed, M.J.; Ahmed, E.; Mamun, M.A.I. Comparative analysis of machine learning techniques for accurate lung cancer prediction. *Am. J. Eng. Technol.* **2024**, *6*, 92–103.
17. Hosseini, S.A.; Hajianfar, G.; Hall, B.; Servaes, S.; Rosa-Neto, P.; Ghafarian, P.; Zaidi, H.; Ay, M.R. Robust vs. non-robust radiomic features. *Cancer Imaging* **2025**, *25*, 33. [CrossRef]
18. Wang, R.; Liu, Q.; You, W.; Wang, H.; Chen, Y. A transformer-based deep learning survival prediction model. *Brief. Bioinform.* **2025**, *26*, 686.
19. Qaraad, M.; Rahrmann, E.P.; Guinovart, D. miRNA-Based Breast Cancer Subtyping Using AHALA Multi-Stage Classification Approach. *Cancers* **2026**, *18*, 586. [CrossRef] [PubMed]
20. Dagliyan, O.; Uney-Yuksektepe, F.; Kavakli, I.H.; Turkay, M. Optimization based tumor classification from microarray gene expression data. *PLoS ONE* **2011**, *6*, e14579. [CrossRef]

21. Thangaselvi, R.; Krishna, K.S.N.V.; Kumar, O.S.; Rishik, A.M. Early detection of lung cancer using hybrid histological image analysis with XGBoost and LightGBM in MATLAB. In *2025 6th International Conference on Intelligent Communication Technologies and Virtual Mobile Networks (ICICV)*; IEEE: New York, NY, USA, 2025; pp. 14–19.
22. VA, B.; Mathew, P.; Thomas, S.; Mathew, L. Detection of lung cancer and stages via breath analysis using a self-made electronic nose device. *Expert Rev. Mol. Diagn.* **2024**, *24*, 341–353. [CrossRef]
23. Bhuiyan, M.S.; Chowdhury, I.K.; Haider, M.; Jisan, A.H.; Jewel, R.M.; Shahid, R.; Ferdus, M.Z.; Siddiqua, C.U. Advancements in early detection of lung cancer in public health. *J. Comput. Sci. Technol. Stud.* **2024**, *6*, 113–121. [CrossRef]
24. Zheng, H.; Zhang, Q.; Gong, Y.; Liu, Z.; Chen, S. Identification of prognostic biomarkers for stage III non-small cell lung carcinoma. In *2024 5th International Conference on Big Data & Artificial Intelligence & Software Engineering (ICBASE)*; IEEE: New York, NY, USA, 2024; pp. 323–326.
25. Nishio, M.; Nishizawa, M.; Sugiyama, O.; Kojima, R.; Yakami, M.; Kuroda, T.; Togashi, K. Computer-aided diagnosis of lung nodule using gradient tree boosting. *PLoS ONE* **2018**, *13*, e0195875. [CrossRef]
26. Chen, N.; Zhou, R.; Luo, Q.; Liu, Y.; Li, C.; Zhang, J.; Guo, J.; Zhou, Y.; Jiang, H.; Qiu, B.; et al. Combining dosimetric and radiomics features for the prediction of radiation pneumonitis in locally advanced non-small cell lung cancer by machine learning. *Int. J. Radiat. Oncol. Biol. Phys.* **2023**, *117*, 38. [CrossRef]