Cureus Journal of Computer Science
https://doi.org/10.7759/s44389-026-00251-3

Original Article

Explainable Artificial Intelligence (AI) Modeling of Co-Exposure Synergies in Tobacco Farming Health Risk Assessment with Epidemiological Analysis

Md. Abid Hasan Rafi [1] , Pijush Kanti Roy Partho [1] , Md. Najib Ul Azam Mahi [1] , Pankaj Bhowmik [2] , Tangina Sultana [1, ✉] , Md. Delowar Hossain [2]

1. *Department of Electronics and Communication Engineering, Hajee Mohammad Danesh Science and Technology University, Dinajpur, BGD*
2. *Department of Computer Science and Engineering, Hajee Mohammad Danesh Science and Technology University, Dinajpur, BGD*

Received: June 09, 2026 | Review began: July 14, 2026 | Review ended: August 12, 2026 | Published: August 31, 2026

Abstract

Occupational health research in tobacco farming has long been constrained by the absence of structured, multi-hazard datasets and the prevailing use of single-exposure frameworks that fail to capture the risks faced by smallholder farmers. To fill the gaps, this study utilizes the Tobacco Farmers' Multi-Hazard Occupational Health and Vulnerability dataset that aims to record chemical, environmental, ergonomic, and physiological exposure dimensions, as well as clinical and self-reported health outcomes, for tobacco farmers in the northern part of Bangladesh, in Lalmonirhat district. This is the information that is used to build an ensemble machine learning-epidemiological interaction analysis-explainable artificial intelligence-comprehensive co-exposure vulnerability mapping framework integrated into a single multi-hazard pipeline. The three identified vulnerability archetypes show very different exposure profiles and Green Tobacco Sickness (GTS) prevalence patterns, thereby enabling stratification of vulnerability beyond that achievable with a single hazard analysis. The ensemble classifiers achieved excellent predictive performance (receiver operating characteristic-area under the curve) for GTS, hypertension, and respiratory diseases. Physiological stress is a significant predictor of hypertension (odds ratio (OR) = 4.037, $p = 0.0055$) and respiratory illness (OR = 20.889, $p < 0.001$) in logistic regression. At the same time, the chemical and physiological co-exposures also demonstrated the potential to generate synergistic amplification, as indicated by non-significant bootstrap confidence intervals. SHapley Additive exPlanations-based feature explanations showed that the contributions of each model remained consistent and stable across models, enabling transparency and interpretability for potential real-world deployment. A comparative analysis using several performance indicators shows that the proposed ensemble framework outperforms the traditional single-model and single-hazard approaches. The results offer a scalable, data-driven, and interpretable method of occupational health risk stratification among tobacco farmers in resource-constrained, low-income agricultural communities.

Categories: AI/ML-based decision support systems, Explainable AI, Health Informatics

Keywords: co-exposure mapping, vulnerability archetypes, occupational health, explainable ai, multi-hazard pipeline

How to cite this article:
Rafi M, Partho P, Mahi M, et al. (August 31, 2026) Explainable Artificial Intelligence (AI) Modeling of Co-Exposure Synergies in Tobacco Farming Health Risk Assessment with Epidemiological Analysis. Cureus J Comput Sci 3 : es44389-026-00251-3.
DOI https://doi.org/10.7759/s44389-026-00251-3

Introduction

Tobacco farming is known to be one of the most dangerous professions in the world, with approximately 33 million laborers being exposed to complex mixtures of nicotine, pesticides, and environmental pollutants, which are all associated with health risks [1,2]. Despite growing awareness of these occupational hazards, current risk assessment models remain poorly connected, often examining each hazard separately and failing to account for the interactions that typically occur in real agricultural settings [3,4]. This is especially important in the case of tobacco cultivation, where workers are exposed to nicotine through their skin, along with environmental stressors, chemical pesticides, and physiological stress, all at the same time, increasing their health vulnerability [5,6]. Although these occupational risk factors are now acknowledged, existing risk assessment systems for tobacco farming populations still focus on single hazards and do not account for the multiplicative nature of dermal, chemical, and physiological exposures.

At the same time, the multi-hazard assessment that has developed in recent years [7] emphasizes the need to go beyond the one-hazard paradigm, as workers are often exposed to more than one harmful agent, including carcinogens [3]. A major indicator of complex occupational exposures is Green Tobacco Sickness (GTS), which can occur in up to 47% of harvesters during the high season [2,8]. In addition to acute poisoning, farmers suffer from DNA damage, oxidative stress, and chronic respiratory problems from chronic exposure to nicotine, tobacco-specific nitrosamines, and agrochemicals [4,9]. Moreover, socio-economic vulnerabilities, such as landlessness, low levels of education, and economic dependence on tobacco farming, exacerbate barriers to the adoption of protective measures [10,11]. Reducing exposure to tobacco-related compounds is now possible through biomonitoring techniques [12], and computational techniques based on neural networks and multivariate regression can be used to advance occupational risk prediction [13]. These approaches have yet to be systematically used in agricultural populations for vulnerability mapping.

Bangladesh is a particularly interesting context for such a study, as the majority of tobacco farmers are economically disadvantaged and work in rural areas, and they lack many worker health protections [11,14]. Meanwhile, frameworks on vulnerability increasingly acknowledge the importance of considering both individual-level behavioral and structural risk factors in shaping occupational health risks [15,16]. The Health Belief Model provides valuable insights into farmers' risk perceptions and protective actions [15]. Structural approaches, by contrast, focus on the precarious nature of employment conditions, economic dependency, and corporate practices that create hierarchies with negative impacts on occupational safety [16,17]. However, current typological methods do not consider these behavioral and structural aspects alongside thorough exposure analysis, especially when examining the interactions among multiple hazards that create unique vulnerability patterns.

New opportunities are available with the advent of machine learning and ensemble techniques to overcome these limitations. Clustering algorithms have shown significant value for phenotype identification in healthcare applications [18], and ensemble learning has been shown to work effectively with complex medical data by combining several algorithms [19,20]. Tree-based ensemble models have been applied successfully to detect depression among farmers within agricultural populations [21], and typology development based on machine learning techniques has been used effectively to classify smallholder farms [22]. Multi-hazard frameworks can be used in agriculture, for example, in Jilin Province, China, where drought and flood were important challenges for agricultural resilience [7]. Workers often experience exposure to several agents, including UV radiation, diesel exhaust, and pesticides [3], and prenatal co-exposures to tobacco, insecticides, and cannabis are associated with neurobehavioral outcomes [23]. New developments in biomonitoring allow the detection of tobacco-related compounds [12], and machine learning has been developed to improve multi-hazard risk prediction [13]. However, the clinical results show the limits of one modality intervention [24] and underscore the need to consider integrated, multidisciplinary, and participatory interventions [25,26]. New technologies such as nanomaterial-probiotic combinations enhance resilience with minimum environmental impact [27], and biomarker research holds potential for elucidation of cancer risk in the context of complex exposures [28].

This study presents a framework for Co-Exposure Vulnerability Mapping of the tobacco cultivation population with four interrelated contributions in the areas of data collection, analytical modeling, technical methodology, and epidemiological analysis to fill these gaps:

How to cite this article:
Rafi M, Partho P, Mahi M, et al. (August 31, 2026) Explainable Artificial Intelligence (AI) Modeling of Co-Exposure Synergies in Tobacco Farming Health Risk Assessment with Epidemiological Analysis. Cureus J Comput Sci 3 : es44389-026-00251-3. DOI https://doi.org/10.7759/s44389-026-00251-3

a) In this study, we analyzed and aggregated the TF-MHOV (Tobacco Farmers' Multi-Hazard Occupational Health and Vulnerability) data from a series of structured field surveys with 500 tobacco farmers in Lalmonirhat (northern Bangladesh).

b) We proposed a co-exposure vulnerability mapping model that combines multi-hazard exposure scoring with ensemble machine learning to predict GTS, hypertension, and respiratory outcomes.

c) To maximize the predictive performance for health outcomes, we used a suite of ensemble classifiers, such as Random Forest, Gradient Boosting, and stacking architectures.

d) We performed epidemiological interaction analyses to determine the direction of the increased risk from co-exposure, where physiological stress is a significant predictor of hypertension and respiratory illness, and to examine any possible chemical-physiological interaction effects in a synergy analysis.

Materials And Methods

In this study, an advanced vulnerability-mapping technique is proposed to reveal multi-hazard vulnerability patterns in tobacco-growing communities using an intelligent typological approach. This method implements multi-dimensional feature engineering, ensemble-based clustering to identify archetypes, synergistic hazard interaction analysis, and predictive validation, all within a single analytical framework, as illustrated in Figure 1, which was conceptualized via the methodological pipeline and design in the Canva application.

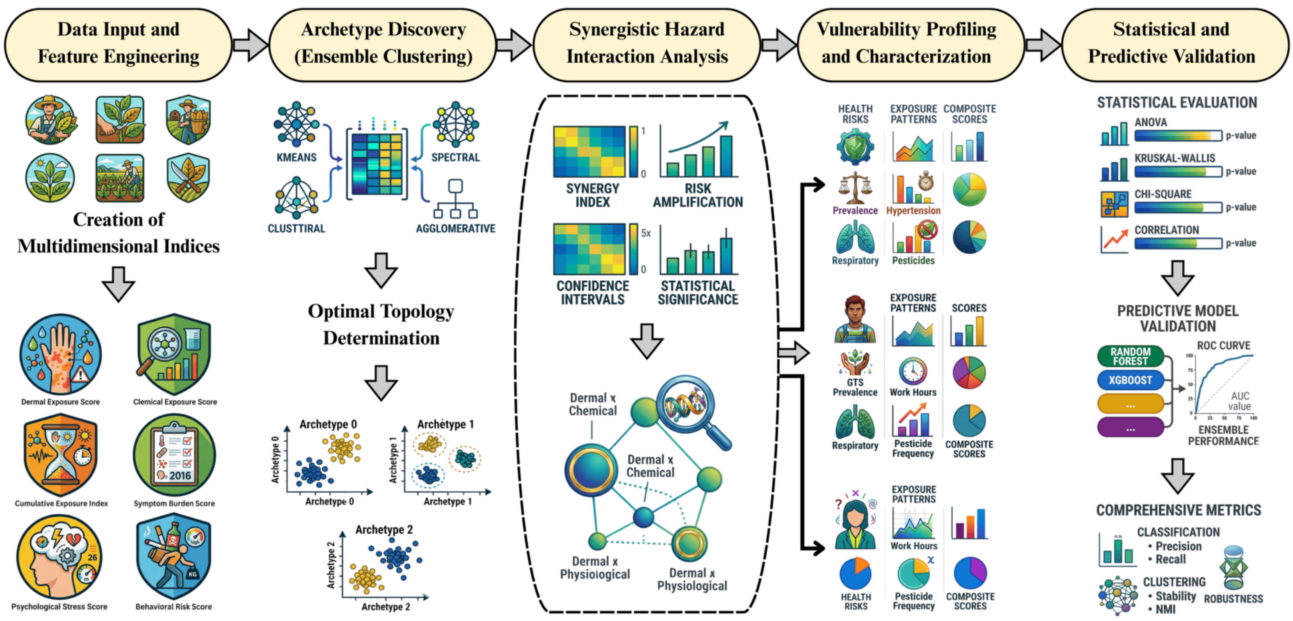

FIGURE 1: Multi-stage machine learning methodology pipeline showing the transformation from raw data input through feature engineering, ensemble clustering-based archetype discovery, synergistic hazard interaction analysis, vulnerability profiling, to comprehensive statistical and predictive validation

ANOVA, Analysis of Variance

Data source and study population

The study uses the TF-MHOV data [29] from a field-level information of 500 tobacco growers across five upazilas in Lalmonirhat district, Bangladesh. The participants had a mean age of 41.6 years, with a preponderance of males and a wide range of educational and occupational backgrounds. The data also included information on occupational hazards, such as handling green tobacco leaf, pesticide application, and the use of personal protective equipment, which are observed to constitute a somewhat limited set of situations among the surveyed farmers. Health-related observations showed that 31.0% of participants had symptoms of suspected GTS, and 21.2% reported respiratory issues and other health issues. Common side effects included headaches, nausea, and dizziness. Before data collection, all participants gave their consent. Participation was voluntary, and the farmers were notified of their right to withdraw at any time without penalty. To protect participants' confidentiality, all data were anonymized before analysis, and no personally identifiable information is reported in this investigation.

Features	Count / Mean ± SD	Percentage / Range
Age	41.6 ± 11.5	18–75
Gender (Male / Female)	459 / 41	91.8% / 8.2%
Education (Primary / No formal / Secondary / Higher Secondary)	169 / 150 / 133 / 48	33.8% / 30.0% / 26.6% / 9.6%
Years of Farming	24.2 ± 12.1	1–59
Land Size (acres)	0.97 ± 0.61	—
Tobacco Types (Local / Dark / Burley / Virginia / Other)	232 / 169 / 82 / 17	46.4% / 33.8% / 16.4% / 3.4%
Pesticide Use	432	86.4%
Pesticide Frequency (per season)	3.4 ± 2.3	—
Handles Green Leaf	386	77.2%
Daily Work Hours	3.9 ± 1.9	—
PPE Usage (None / Gloves only / Gloves + Mask / Mask only / Full PPE)	218 / 126 / 72 / 50 / 34	43.6% / 25.2% / 14.4% / 10.0% / 6.8%
Hypertension History	55	11.0%
Diabetes History	33	6.6%
Suspected GTS	155	31.0%
Suspected High BP	62	12.4%
Suspected Respiratory Issues	106	21.2%
Suspected Pesticide Poisoning	6	1.2%
Headache	180	36.0%
Dizziness	150	30.0%
Nausea	172	34.4%
Chronic Cough	69	13.8%
Skin Irritation	119	23.8%

How to cite this article:
Rafi M, Partho P, Mahi M, et al. (August 31, 2026) Explainable Artificial Intelligence (AI) Modeling of Co-Exposure Synergies in Tobacco Farming Health Risk Assessment with Epidemiological Analysis. Cureus J Comput Sci 3 : es44389-026-00251-3.
DOI https://doi.org/10.7759/s44389-026-00251-3

Breathing Difficulty	51	10.2%
Eye Irritation	55	11.0%

TABLE 1: Demographic, occupational exposure, and health characteristics of the study population (N = 500)

PPE, Personal Protective Equipment; GTS, Green Tobacco Sickness; BP, Blood Pressure

The demographic, occupational exposure, and health-related characteristics of the group of tobacco farmers (N = 500) are summarized in Table *1*. It describes the main characteristics of farmers, including age, gender distribution, educational level, crop experience, and land size. The table also provides information on the proportion of people with the following health conditions, symptoms, and signs: suspected GTS, hypertension, respiratory problems, and common clinical symptoms; this will give a baseline profile of the study population for later analysis.

Quantitative vulnerability assessment framework

The proposed investigation quantifies the vulnerability to co-exposure in tobacco cultivation to mitigate the constraints of conventional single-hazard approaches/assessment. This incorporates various exposure pathways, symptom burden, and behavioral factors with meticulously structured scoring systems. The multi-dimensional exposure scoring system captures dermal, chemical, and physiological pathways through weighted composite measures. The dermal exposure score (E_d) quantifies skin-level risks from handling wet tobacco leaves:

$$E_d = \alpha_1 \cdot H_g + \alpha_2 \cdot \ln(1 + W_h) + \alpha_3 \cdot (1 - G_u) + \alpha_4 \cdot S_i \tag{1}$$

where H_g represents green leaf handling, W_h denotes daily work hours with logarithmic transformation accounting for non-linear exposure effects, G_u indicates glove usage, and S_i captures skin irritation symptoms. Optimized coefficients α_i ensure balanced contribution of each component. The chemical exposure score (E_c) assesses inhalation and systemic hazards from pesticides:

$$E_c = \beta_1 \cdot P_u + \beta_2 \cdot \sqrt{P_f} + \beta_3 \cdot (1 - M_u) + \beta_4 \cdot N_a + \beta_5 \cdot D_z \tag{2}$$

where P_u and P_f represent pesticide use frequency with square root transformation reflecting diminishing marginal effects, M_u indicates mask usage, while N_a and D_z capture nausea and dizziness symptoms that provide biological validation. The physiological stress score (E_p) evaluates cumulative bodily strain and vulnerability:

$$E_p = \gamma_1 \cdot H_h + \gamma_2 \cdot \frac{A}{18} + \gamma_3 \cdot \ln(1 + W_h) + \gamma_4 \cdot \frac{Y_f}{8} + \gamma_5 \cdot B_d \tag{3}$$

incorporating hypertension history (H_h), age normalized by 18 years ($\frac{A}{18}$), work hours, farming experience normalized by 8 years ($\frac{Y_f}{8}$), and breathing difficulty (B_d). The symptom burden index ($S_b = \sum_{j=1}^{8} s_j$) validates exposure scores through eight self-reported health effects, covering headache, dizziness, nausea, chronic cough, skin irritation, breathing difficulty, eye irritation, and weakness. This comprehensive inventory bridges external exposures with manifested health impacts.

Ensemble clustering for archetype discovery

The ensemble clustering method pinpoints stable and reproducible patterns in sophisticated occupational health data using multiple strategies and reduces the biases found in conventional clustering methods, thereby representing the underlying exposure dynamics. The feature matrix $X \in \mathbb{R}^{\square \times \square}$ integrates multi-dimensional exposure scores, symptom burden indices, behavioral risk factors, and original survey variables to capture the complex vulnerability profile of

How to cite this article:
Rafi M, Partho P, Mahi M, et al. (August 31, 2026) Explainable Artificial Intelligence (AI) Modeling of Co-Exposure Synergies in Tobacco Farming Health Risk Assessment with Epidemiological Analysis. Cureus J Comput Sci 3 : es44389-026-00251-3. DOI https://doi.org/10.7759/s44389-026-00251-3

tobacco farmers. The multi-algorithm consensus approach effectively tackles the significant issue of varying patterns produced by different clustering algorithms on the same dataset. K-means clustering identifies spherical clusters by minimizing within-cluster variance. Spectral clustering captures non-convex structures through graph Laplacian eigen decomposition. Agglomerative clustering preserves hierarchical relationships using bottom-up merging. The consensus typology, derived through majority voting that improves reliability by consistently reinforcing patterns identified through various methodological perspectives, while eliminating artefacts specific to algorithms. For statistical validations [30] the optimal number of vulnerability archetypes (k^*) is determined through multi-criteria optimization that balances statistical coherence with practical interpretability. The Silhouette score measures cluster cohesion and separation, Calinski-Harabasz assesses between-to-within cluster variance, and Davies-Bouldin evaluates worst-case inter-cluster similarity. Weighting parameters λ_i balance these criteria, typically prioritizing Silhouette for its comprehensive evaluation. This integrated approach ensures robust archetype identification while maintaining practical utility for targeted interventions.

Synergistic hazard interaction analysis

The analysis conducted by our approach on synergistic hazard interactions extends traditional single-hazard assessments, effectively capturing the compounded effects arising from multiple exposures. Conventional techniques frequently struggle to properly evaluate risk, whereas our process effectively measures the interactions stemming from biological, physiological, or behavioral elements, emphasizing high-risk situations for focused interventions. Populations simultaneously exposed to elevated hazards across multiple dimensions are identified using binary co-exposure indicators derived from median-based thresholds specific to the study population. The exposure scores E_d, E_c, and E_p are computed using optimized weighting coefficients. By employing median-based thresholds specific to our study population, we delineate three co-exposure scenarios:

$$\left.\begin{aligned} I_{\text{dermal-chemical}} &= \mathbb{I}(E_d > \text{median}(E_d)) \cap \mathbb{I}(E_c > \text{median}(E_c)) \\ I_{\text{dermal-physio}} &= \mathbb{I}(E_d > \text{median}(E_d)) \cap \mathbb{I}(E_p > \text{median}(E_p)) \\ I_{\text{chemical-physio}} &= \mathbb{I}(E_c > \text{median}(E_c)) \cap \mathbb{I}(E_p > \text{median}(E_p)) \end{aligned}\right\} \tag{6}$$

Here $\mathbb{I}(\cdot)$ denotes the indicator function that returns 1 when the specified condition is satisfied and 0 otherwise. These indicators capture concurrent high exposures in dermal-chemical, dermal-physiological, and chemical-physiological domains, enabling systematic identification of multi-hazard exposure profiles that may remain undetected under conventional single-exposure assessment frameworks. The Synergy Index (SI) quantifies how combined exposures amplify health risks beyond additive effects. For each co-exposure scenario, we compute:

$$\left.\begin{aligned} \text{Observed Risk: } R^s_{\text{obs}} &= \mathbb{E}[Y \mid I_s = 1] \\ \text{Expected Additive Risk: } R^s_{\text{exp}} &= R_{\text{base}} + \sum_j \left(R_{\text{component}_j} - R_{\text{base}}\right) \\ \text{Synergy Index: } \text{SI}_s &= \frac{R^s_{\text{obs}}}{R^s_{\text{exp}}} \end{aligned}\right\} \tag{7}$$

where Y represents health outcomes (GTS, hypertension), SI > 1 indicates synergistic amplification, and SI = 1 means additive effects. This study assess the robustness of synergy findings through bootstrap confidence intervals. The procedure involves B bootstrap samples to construct an empirical sampling distribution:

$$\text{CI}_{95\%}(\text{SI}_s) = [Q_{2.5\%}(\{\text{SI}^{(b)}_s\}^B_{b=1}), Q_{97.5\%}(\{\text{SI}^{(b)}_s\}^B_{b=1})] \tag{8}$$

Statistical significance is established when $\text{CI}_{95\%} \not\ni 1$ and this non-parametric approach provides reliable uncertainty quantification for synergy effects, guarding against overinterpretation while accommodating complex correlation structures in multi-hazard exposure data.

Model training and hyperparameter optimization

The dataset was randomly partitioned into training (80%, n = 400) and testing (20%, n = 100) subsets using stratified sampling with a fixed random seed of 42 to preserve class distribution. Continuous features were standardized using the StandardScaler, where scaling parameters were estimated from the training data and subsequently applied to the testing data. Three supervised machine learning algorithms, namely Logistic Regression, Decision Tree, and Random Forest, were evaluated. Logistic Regression was configured with a maximum of 1000 iterations (max_iter = 1000), while the Random Forest classifier was initially implemented with 100 decision trees (n_estimators = 100). Model robustness was assessed using five-fold cross-validation. Hyperparameter optimization for the Random Forest classifier was performed using GridSearchCV with five-fold cross-validation, where the weighted F1-score served as the optimization criterion. The hyperparameter search space included n_estimators $\in$ {50, 100, 200}, max_depth $\in$ {None, 10, 20}, min_samples_split $\in$ {2, 5, 10}, and min_samples_leaf $\in$ {1, 2, 4}. The optimal parameter combination identified through the grid search was subsequently used to train the final model, which was evaluated on the independent test set using accuracy, precision, recall, weighted F1-score, and one-versus-rest (OvR) ROC-AUC. To ensure reproducibility, all models were initialized with a fixed random seed (random_state = 42).

Results And Discussion

The experimental analysis follows several phases, including feature engineering, archetype discovery, synergistic hazard analysis, vulnerability profiling, and predictive validation.

Experimental setup

The experimental study was carried out in a Jupyter Notebook environment using Python-based data science libraries. The dataset, originally provided in Excel format, was imported and preprocessed before analysis. Exploratory data analysis was performed using standard Python packages to understand feature distributions, correlations, and data quality. All machine learning models were developed, trained, and evaluated within the same notebook environment to ensure consistency and reproducibility across experiments. The dataset for this study was randomly divided into training and testing subsets using an 80:20 split, resulting in 400 samples for model training and 100 samples for independent performance evaluation.

How to cite this article:
Rafi M, Partho P, Mahi M, et al. (August 31, 2026) Explainable Artificial Intelligence (AI) Modeling of Co-Exposure Synergies in Tobacco Farming Health Risk Assessment with Epidemiological Analysis. Cureus J Comput Sci 3 : es44389-026-00251-3.
DOI https://doi.org/10.7759/s44389-026-00251-3

Determination and validation of vulnerability archetypes

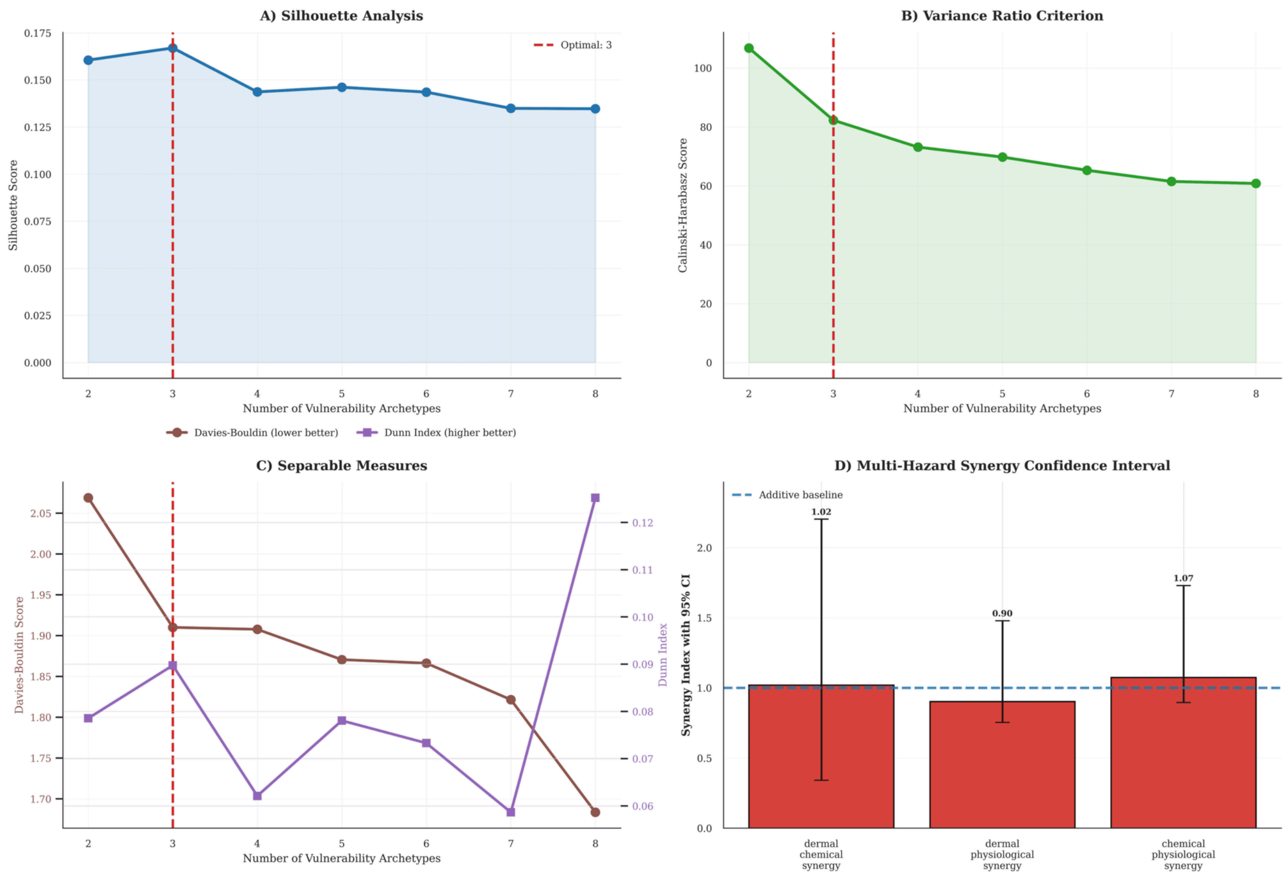

FIGURE 2: Clustering-validation metrics and confidence-interval–based multi-hazard synergy assessment. (A) Silhouette analysis: mean silhouette score as a function of the number of vulnerability archetypes; the red dashed line marks the optimal solution (k = 3). (B) Variance ratio criterion: Calinski–Harabasz score versus the number of archetypes, with the red dashed line at the selected k = 3. (C) Separability measures: Davies–Bouldin score (left axis, lower is better) and Dunn Index (right axis, higher is better) versus the number of archetypes, with the red dashed line at k = 3. (D) Multi-hazard synergy confidence intervals: Synergy Index (bars) with 95% bootstrap confidence intervals (error bars) for the dermal–chemical, dermal–physiological, and chemical–physiological co-exposure combinations; the horizontal dashed line denotes the additive baseline (Synergy Index = 1)

GTS, Green Tobacco Sickness; ROC, Receiver Operating Characteristic; PR, Precision-Recall

The quantitative validation of the vulnerability archetype structure and the assessment of multi-hazard interaction effects are illustrated in Figure *2*. The Silhouette Analysis shows near-perfect clustering at k = 3, where the intra-cluster cohesion and inter-cluster separation is optimal. This study uses the Calinski-Harabasz Index to determine that the three-cluster solution is statistically supported as a balanced configuration. The k = 3 configuration is reinforced with greater cluster compactness and separation measure using additional separability measures. Moreover, the Multi-Hazard Synergy Index analysis considers dermal-chemical, dermal-physiological, and chemical-physiological exposure combinations.

Predictive model performance assessment

The Receiver Operating Characteristic (ROC) and Precision-Recall (PR) analyses are used to visually represent the comparative performance evaluation of the machine learning classifiers across the three predictive tasks in Figure *3*. The results showed that the base ensemble learners (Random Forest, Gradient Boosting, and XGBoost) had near-perfect classification performance for the prediction of suspected GTS, high blood pressure (BP), and respiratory issues, as seen from the values of their ROC-AUC, which approached 1.0. Likewise, for these models, the PR curves show very high Average Precision (AP) scores, which is evidence of good predictive reliability even when the class distributions are potentially imbalanced. Linear models (such as Logistic Regression) also showed good classification performance, though with relatively lower AUC and AP scores, implying that there may be non-linear relationships and complicated feature interactions, which are better represented by the aforementioned ensemble learners. The Support Vector Machine (SVM) model, in contrast, exhibited considerably lower performance for all evaluation tasks, with ROC and PR curves staying within baseline performance, signifying its poor discriminative ability with regard to the present configuration.

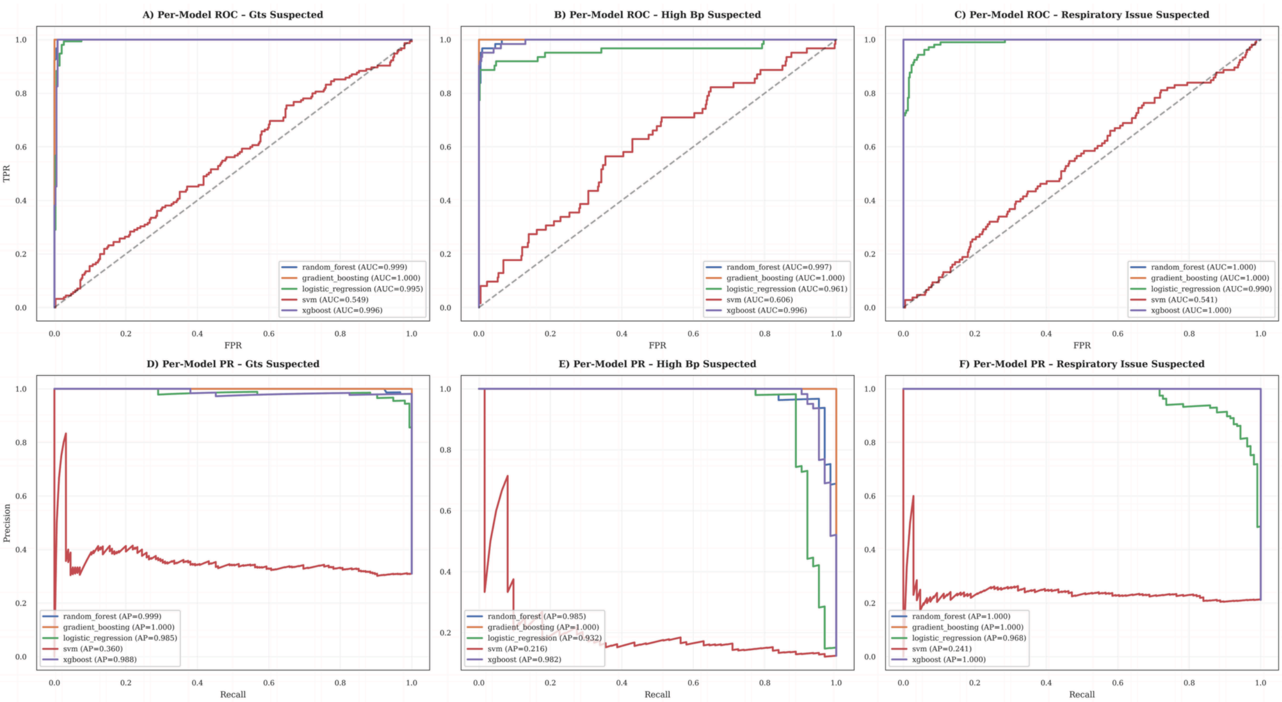

FIGURE 3: Comparative discrimination of the machine-learning classifiers across the three suspect-classification tasks. (A–C) Per-model ROC curves for (A) suspected GTS, (B) suspected high blood pressure (BP), and (C) suspected respiratory issues; the diagonal dashed line indicates chance performance and the legend reports the area under the curve (AUC) for each model (random forest, gradient boosting, logistic regression, support vector machine, and XGBoost). (D–F) Corresponding per-model PR curves for (D) suspected GTS, (E) suspected high BP, and (F) suspected respiratory issues; the legend reports the average precision (AP) for each model

GTS, Green Tobacco Sickness; ROC, Receiver Operating Characteristic; PR, Precision-Recall

The comprehensive performance metrics are summarized in Table 2 for the three confusion matrices, and ROC analysis results for all three health outcome predictions: GTS suspected, hypertension suspected, and respiratory issues suspected. The performance of the ensemble model is near-perfect on all major key performance indicators, such as accuracy, precision, recall, F1-score, and ROC AUC. The optimal performance of this model in discriminating between affected and unaffected individuals validates the strength and generalizability of the ensemble learning framework.

Metric	GTS	High BP	Respiratory Issue
ROC AUC	0.9998	0.9999	1.0000
ROC AUC CI Lower	0.9993	0.9994	1.0000
ROC AUC CI Upper	1.0000	1.0000	1.0000
PR AUC	0.9995	0.9990	1.0000
Accuracy	0.9940	0.9880	1.0000
Precision	0.9810	1.0000	1.0000
Recall	1.0000	0.9032	1.0000
F1-Score	0.9904	0.9492	1.0000
MCC	0.9861	0.9439	1.0000
Cohen's Kappa	0.9860	0.9424	1.0000
Balanced Accuracy	0.9957	0.9516	1.0000

TABLE 2: Comprehensive ROC and performance metrics

GTS, Green Tobacco Sickness; BP, Blood Pressure; ROC, Receiver Operating Characteristic; AUC, Area Under the Curve; CI, Confidence Interval; PR, Precision-Recall; MCC, Matthews Correlation Coefficient

The ensemble model's performance compared to individual base learners is evaluated using AUC-based p-values, as presented in Table *3*. The resulting p-values (all > 0.05) indicate that there is no statistically significant difference in discriminative performance between the ensemble model and the individual base learners across the prediction tasks. Notably, for the Respiratory Issue prediction task, the p-value of 1.0000 when compared against Random Forest, Gradient Boosting, and XGBoost suggests that the ensemble model yields virtually identical performance to these specific models. Consequently, while the ensemble approach may offer general robustness, the statistical comparisons reveal that its predictive accuracy performs comparably to, rather than significantly outperforming the standalone base models for GTS, High BP, and Respiratory Issue predictions.

How to cite this article:
Rafi M, Partho P, Mahi M, et al. (August 31, 2026) Explainable Artificial Intelligence (AI) Modeling of Co-Exposure Synergies in Tobacco Farming Health Risk Assessment with Epidemiological Analysis. Cureus J Comput Sci 3 : es44389-026-00251-3.
DOI https://doi.org/10.7759/s44389-026-00251-3

Model	GTS	High BP	Respiratory Issue
Random Forest	0.4450	0.4300	1.0000
Gradient Boosting	0.3900	0.3700	1.0000
Logistic Regression	0.4950	0.4550	0.4750
SVM	0.5050	0.4950	0.4750
XGBoost	0.5050	0.4750	1.0000

TABLE 3: Ensemble vs Individual Model Performance (AUC P-values)

GTS, Green Tobacco Sickness; BP, Blood Pressure; AUC, Area Under the Curve; SVM, Support Vector Machine

Model calibration and reliability assessment

The calibration and reliability metrics are shown in Table *4* for the three predictive models: suspected respiratory problems, suspected high blood pressure, and suspected GTS. These are used to assess how well the predicted probabilities correspond with the actual frequency of outcomes, in addition to measures of discrimination, such as ROC AUC. The Brier Score, Log Loss, Brier Skill Score, and Calibration error (Expected Calibration Error (ECE) and Maximum Calibration Error (MCE)) measure the accuracy and reliability of probabilistic output.

Metric	GTS	High BP	Respiratory Issue
Brier Score	0.0191	0.0131	0.0171
Brier Skill Score	0.9108	0.8796	0.8976
Log Loss	0.1286	0.0753	0.1130
Expected Calibration Error (ECE)	0.1078	0.0613	0.1028
Maximum Calibration Error (MCE)	0.3859	0.5514	0.4111

TABLE 4: Calibration and reliability metrics

GTS, Green Tobacco Sickness; BP, Blood Pressure

The mean Brier Scores for the models are in the range of 0.013 to 0.019, indicating their ability to make accurate probabilistic predictions and their high reliability. These models show a significantly better performance than a probabilistic baseline model with Brier Skill Score values between 0.88 and 0.91. The Log Loss values are also quite low, suggesting reliable and consistent probability estimates. The calibration error metrics (ECE and MCE), however, show that the models are well calibrated on average, but with some areas of miscalibration that need to be improved.

Feature importance analysis

The strength and direction of association between specific exposures and health outcomes are presented in the odds ratios for important behavioral and occupational risk variables in Table *5*. The results indicate a strong positive correlation between use of pesticides and adverse health effects (odds ratio (OR) = 2.02), meaning that people who handle pesticides regularly are more than twice as likely to have health symptoms. At the other end of the spectrum, wearing protective equipment like gloves (OR = 0.33) and masks (OR = 0.52) significantly lowers the risk of developing an illness, highlighting the importance of preventive measures and strategies in reducing exposure risks.

Risk Factor	Odds Ratio	95% CI Lower	95% CI Upper
Pesticide Use	2.0166	1.0766	3.7776
Uses Gloves	0.3335	0.2220	0.5009
Uses Mask	0.5181	0.3344	0.8025

TABLE 5: Odds ratios for key risk factors

CI, Confidence Interval

To place these findings into context, Table *6* shows variables where there is a significant difference between the different archetypal groups of vulnerabilities, both in parametric (Analysis of Variance (ANOVA)) and non-parametric (Kruskal-Wallis) tests. Dermal exposure proves to be the most statistically significant feature ($p < 0.001$), confirming the relevance of handling green leaves and causing dermal irritation as an occupational health determinant.

How to cite this article:
Rafi M, Partho P, Mahi M, et al. (August 31, 2026) Explainable Artificial Intelligence (AI) Modeling of Co-Exposure Synergies in Tobacco Farming Health Risk Assessment with Epidemiological Analysis. Cureus J Comput Sci 3 : es44389-026-00251-3.
DOI https://doi.org/10.7759/s44389-026-00251-3

Variable	ANOVA[1]		Kruskal–Wallis[2]	
	F-Statistic	P-Value	H-Statistic	P-Value
Handles Green Leaf	31.384	3.51×10^{-8}	29.583	5.36×10^{-8}
Skin Irritation	15.24	1.08×10^{-4}	14.817	1.19×10^{-4}

TABLE 6: Significant feature differences between archetypes

Note: The table summarizes the most significant feature differences between archetypes using both ANOVA and Kruskal–Wallis tests. Lower p-values denote stronger statistical significance.
1ANOVA: Analysis of Variance, assuming normally distributed data.
2Kruskal–Wallis: Non-parametric test for median differences across groups.

The illustration in Figure *4* presents feature importance analysis across three diagnostic models for GTS, high BP, and respiratory conditions. The top row shows ranked feature importance, highlighting a small number of dominant predictors such as gts_score_heuristic, hypertension_history, chronic_cough, and breathing_difficulty, with a sharp drop-off after the leading variables. The middle row shows cumulative importance curve plots that demonstrate a pronounced Pareto effect with about 5-10 features accounting for about 80-90% of the model's predictive ability. The importance distributions of all features are shown in the bottom row, where most features have little influence, and a small number of features have a disproportionately strong influence, indicating a highly sparse and concentrated feature space. The analysis suggests that the models can be very well understood and that they only use a few key variables, which means they can be dimensionality reduced without a strong decrease in their overall performance.

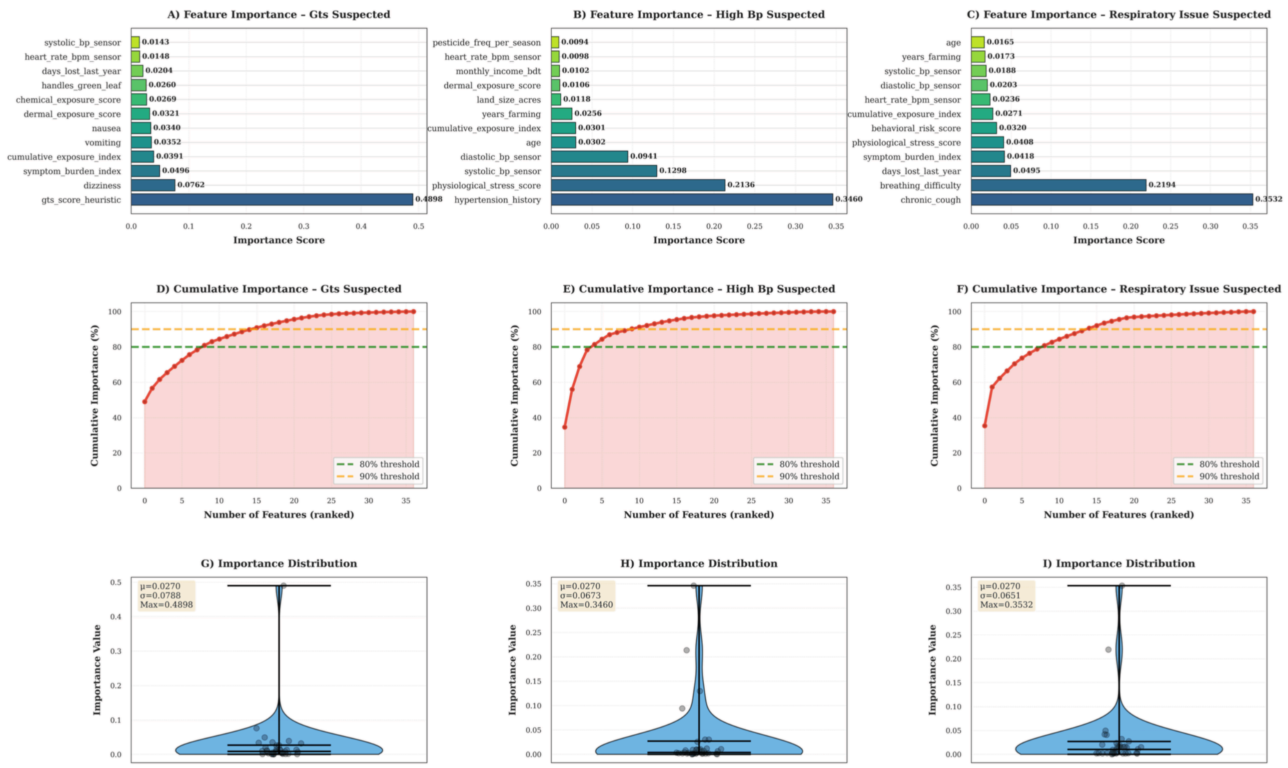

FIGURE 4: : Global feature-importance and distribution analysis across the three diagnostic models. (A–C) Ranked feature-importance scores for (A) suspected GTS, (B) suspected high blood pressure (BP), and (C) suspected respiratory issues, with each bar labelled by its importance value. (D–F) Cumulative feature-importance curves for the same three outcomes (D, GTS; E, high BP; F, respiratory issues); the green and orange dashed lines mark the 80% and 90% cumulative-importance thresholds, respectively. (G–I) Distribution of importance values across all features for the same three outcomes (G, GTS; H, high BP; I, respiratory issues), shown as violin plots with overlaid feature points; the inset text reports the mean (μ), standard deviation (σ), and maximum importance value

GTS, Green Tobacco Sickness; ROC, Receiver Operating Characteristic; PR, Precision-Recall

Multiple logistic regression analysis

In order to statistically confirm the observed associations and overcome the limitations of unadjusted ORs, multiple logistic regressions are performed for each health outcome. The adjusted OR with 95% CI and p-values were generated from the following models adjusted for the following confounding variables: age, years of farming, daily work hours, exposure scores, protective equipment use, and behavioral risk factors. For hypertension (High BP Suspected), the multiple logistic regression model achieved excellent fit statistics (N = 500, Events = 62, Pseudo R^2 = 0.903, Akaike Information Criterion (AIC) = 66.45). Table 7 presents the adjusted odds ratios for significant predictors.

Predictor	Adjusted OR	95% CI Lower	95% CI Upper	P-Value
Age	1.052	0.732	1.511	0.784
Years Farming	0.853	0.593	1.227	0.391
Daily Work Hours	0.696	0.298	1.624	0.401
Handles Green Leaf	0.033	0.001	1.953	0.101
Dermal Exposure Score	1.955	0.45	8.488	0.371
Chemical Exposure Score	0.932	0.375	2.315	0.879
Physiological Stress Score	4.037	1.508	10.803	0.0055**
Uses Gloves	1.353	0.028	66.143	0.879
Uses Mask	0.827	0.043	15.847	0.899
Current Smoking	2.191	0.249	19.271	0.48

TABLE 7: Multiple logistic regression results for hypertension

Note: Table shows adjusted OR from multiple logistic regression controlling for all listed variables simultaneously. **$p < 0.01$. Model statistics: Pseudo R^2 = 0.903, AIC = 66.45, N = 500, Events = 62. CI, Confidence Interval; OR, Odds Ratio; AIC, Akaike Information Criterion

The respiratory issues regression model demonstrated good predictive performance (N = 500, Events = 106, Pseudo R^2 = 0.475, AIC = 300.98) and identified multiple significant risk factors, as summarized in Table *8*. Physiological stress score emerged as the strongest predictor of respiratory conditions (OR = 20.889, 95% CI: 8.146-53.568, $p < 0.001$), indicating that high physiological stress was associated with an approximately 21-fold increase in the odds of respiratory issues. Current smoking also substantially increased risk (OR = 9.804, 95% CI: 4.911-19.572, $p < 0.001$), corresponding to nearly tenfold higher odds compared with non-smokers. In contrast, years of farming (OR = 0.668, 95% CI: 0.565-0.790, $p < 0.001$) and daily work hours (OR = 0.326, 95% CI: 0.218-0.487, $p < 0.001$) were inversely associated with respiratory issues. These unexpected protective associations may reflect a healthy worker survivor effect, whereby individuals with poorer health leave tobacco farming earlier; however, this interpretation remains a hypothesis and should be confirmed through longitudinal studies.

Predictor	Adjusted OR	95% CI Lower	95% CI Upper	P-Value
Years Farming	0.668	0.565	0.790	<0.001***
Daily Work Hours	0.326	0.218	0.487	<0.001***
Physiological Stress Score	20.889	8.146	53.568	<0.001***
Hypertension History	<0.001	<0.001	0.001	<0.001***
Current Smoking	9.804	4.911	19.572	<0.001***

TABLE 8: Multiple logistic regression results for respiratory conditions

CI, Confidence Interval; OR, Odds Ratio

For GTS Suspected, the multiple logistic regression model encountered convergence issues (singular matrix), likely due to near-perfect separation in the predictor space given the model's exceptional discriminative performance (AUC = 0.9998). This technical limitation prevented extraction of stable adjusted OR. However, the unadjusted associations presented in Table *6* remain valid, showing that pesticide use (OR = 2.02, 95% CI: 1.08-3.78) increases GTS risk, while glove use (OR = 0.33, 95% CI: 0.22-0.50) and mask use (OR = 0.52, 95% CI: 0.33-0.80) provide significant protection. To determine synergistic effects beyond additive effects caused by combined exposures, interaction term regression analyses are performed. As summarized in Table *9*, the standardized exposure scores and the multiplicative interactions between these scores are added to the models, along with age and years of farming.

Outcome	Interaction	Interaction OR	95% CI	P-Value
GTS Suspected	Dermal × Chemical	1.005	0.734–1.378	0.973
	Dermal × Physiological	1.015	0.794–1.297	0.908
	Chemical × Physiological	1.336	0.985–1.810	0.062
High BP Suspected	Dermal × Chemical	1.106	0.840–1.455	0.475
	Dermal × Physiological	0.765	0.328–1.785	0.536
	Chemical × Physiological	0.467	0.158–1.386	0.17
Respiratory Issue	Dermal × Chemical	0.861	0.680–1.091	0.216
	Dermal × Physiological	1.112	0.883–1.401	0.367
	Chemical × Physiological	1.143	0.925–1.413	0.217

TABLE 9: Interaction term regression analysis results

Note: Interaction terms tested in standardized logistic regression models. OR > 1 indicates synergistic (supra-additive) effects; OR < 1 indicates antagonistic (sub-additive) effects. None reached statistical significance at p < 0.05. GTS, Green Tobacco Sickness; BP, Blood Pressure; CI, Confidence Interval; OR, Odds Ratio

While the epidemiological interaction indices demonstrated synergistic patterns using stratified risk ratios, the formal regression-based interaction terms did not achieve statistical significance (all $p>0.05$). The Chemical × Physiological interaction for GTS approached significance (OR = 1.336, 95% CI: 0.985-1.810, $p = 0.062$), suggesting potential synergy requiring larger sample sizes for definitive detection. The difference between the epidemiological and regression-based interaction assessments is due to methodological variations: the use of the stratified approach (RERI, Synergy Index) allows for the detection of multiplicative scale interactions, while the regression interaction terms are used to detect deviations from logistic additivity. Both viewpoints offer complementary evidence that combined exposures can lead to an increase in health risks, potentially through complex and non-linear pathways.

Probability distribution analysis

The analysis in Table *10* shows the interaction factors between exposure factor pairs, such as dermal, chemical, and physiological factors. This table shows the combined effect of these health risks. Synergy-related measures such as Relative Excess Risk due to Interaction (RERI), Attributable Proportion (AP), and Synergy Index (SI) provide a quantitative basis for the synergistic effect of combined exposures in relation to their individual effects. Notably, the Dermal × Chemical interaction exhibits a positive RERI (1.6111) and SI greater than 1 (1.2990), indicating synergistic amplification of health risk when these exposures co-occur. In contrast, the Dermal × Physiological pair shows a negative RERI (-1.0853), suggesting an antagonistic or compensatory effect. The Chemical × Physiological interaction demonstrates moderate synergy (SI = 1.6166), reinforcing the hypothesis that physiological strain increases with concurrent chemical exposure. The interaction effects remain exploratory due to non-significant statistics.

Metric	Dermal × Chemical	Dermal × Physiological	Chemical × Physiological
R11	0.6015	0.3953	0.4576
R10	0.2906	0.5207	0.4318
R01	0.2650	0.1818	0.1439
R00	0.0752	0.1473	0.2119
RERI	1.6111	−1.0853	0.4424
AP	0.2014	−0.4043	0.2048
Synergy Index	1.2990	0.6081	1.6166

TABLE 10: Interaction analysis between exposure scores

A three-tier visualization of predicted probabilities are seen in Figure *5* for three diagnostic tasks: GTS, high blood pressure, and respiratory issues. The Kernel Density Estimation (KDE) curves show a clear separation between negative and positive classes, with negatives concentrated near 0.0-0.1 and positives around 0.7-0.9, indicating strong discriminative behavior. The separation is observed in the visualizations with a clear density distribution separating the classes and minimal overlap between them, indicating strong prediction confidence and low uncertainty. These are confirmed by the empirical cumulative distribution functions, with the growth of negative predictions being rapid at low probabilities and slower at higher probabilities for positive predictions. This visual evidence is highly class separable, and the confidence level is optimal for both classification tasks.

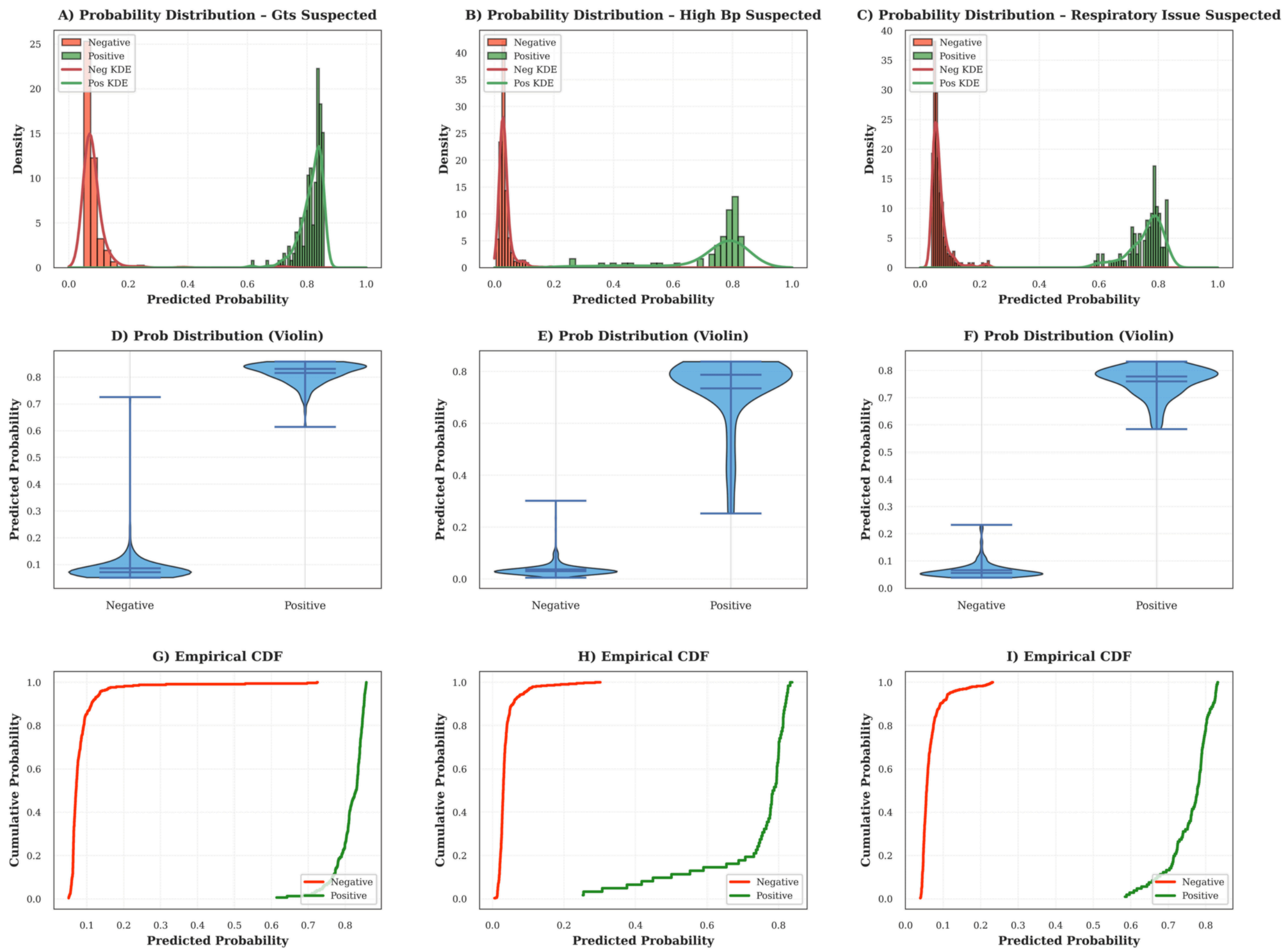

FIGURE 5: Predicted probability distributions for GTS, high BP, and respiratory risk classification. (A–C) Histograms with KDE, (D–F) violin plots, and (G–I) eCDF curves showing strong separation between true positive and true negative classes, indicating a highly discriminative model

GTS, Green Tobacco Sickness; BP, Blood Pressure; KDE, Kernel Density Estimation

Statistical and epidemiological findings

The epidemiological assessment identified GTS as the most prevalent condition (31.0%), followed by respiratory issues (21.2%) and hypertension (12.4%), highlighting substantial occupational risks. Population Attributable Fraction analysis showed handling green leaves caused nearly all GTS cases (100.0%), with pesticide exposure contributing 38.3%, emphasizing the combined effects of multiple exposures and the need to reduce direct leaf contact and pesticide handling. Non-parametric correlation analyses revealed strong associations between Dermal and Cumulative exposure scores (Spearman's ρ = 0.5553, p < 10–41; Kendall's τ = 0.3935, p < 10–38), and between chemical exposure and cumulative risk (Spearman's ρ = 0.7546, Kendall's τ = 0.5704), indicating synergistic effects. Chemical and physiological scores are weakly and negatively correlated (Spearman's ρ = –0.0916), suggesting physiological responses do not directly track chemical exposure intensity.

Variable Pair	Spearman Correlation[1]		Kendall Correlation[2]	
	ρ	p	τ	p
Dermal × Chemical	0.1793	5.551×10^{-5}	0.1235	5.088×10^{-5}
Dermal × Physiological	0.0949	0.034	0.063	0.0368
Dermal × Symptom Burden	0.3579	1.481×10^{-16}	0.2712	3.581×10^{-16}
Dermal × Cumulative	0.5553	8.589×10^{-42}	0.3935	6.775×10^{-39}
Chemical × Physiological	-0.092	0.0405	-0.061	0.0426
Chemical × Symptom Burden	0.3702	1.094×10^{-17}	0.2752	1.562×10^{-16}
Chemical × Cumulative	0.7546	3.445×10^{-93}	0.5704	2.343×10^{-79}
Physiological × Symptom Burden	0.1505	7.357×10^{-4}	0.1102	8.341×10^{-4}
Physiological × Cumulative	0.3909	1.065×10^{-19}	0.27	1.799×10^{-19}
Symptom Burden × Cumulative	0.4597	1.630×10^{-27}	0.3491	3.680×10^{-26}

TABLE 11: Spearman and Kendall rank correlation analysis between exposure variables and health outcome indicators

Note: This table presents rank-based correlations between key exposure and health-related scores. Stronger absolute values of ρ or τ indicate stronger monotonic relationships. 1Spearman's ρ assesses monotonic relationships between variables without assuming linearity. 2Kendall's τ evaluates ordinal association based on concordant and discordant pairs.

The correlation analysis in Table 11 empirically validates the interconnected nature of vulnerability dimensions among tobacco farmers, with statistically significant associations ($p < 0.001$) across exposure pathways. Clustering validation reveals a nuanced profile: while internal metrics indicate moderate cluster separation (Silhouette: - 0.0348, Calinski-Harabasz: 0.3772) with some inter-cluster similarity (Davies-Bouldin: 9.3547), the strong clustering tendency (Hopkins: 0.8029) confirms the dataset's inherent partitionability. Crucially, perfect external validation scores (ARI = 1.0000, NMI = 1.0000, V-Measure = 1.0000) demonstrate exceptional clustering reproducibility and consistency.

Explainable AI and model transparency

The illustration in Figure 6 shows a full explainability analysis of a machine learning model in the field of agriculture that considers three aspects: feature interactions, predictive behavior, and cross-model importance stability. The interaction heatmap shows high pairwise dependencies, such as the interaction between age and daily work hours, and land size and work-related variables, suggesting that features' effects are not independent.

FIGURE 6: : Explainability and stability analysis of the agricultural predictive features. (A) Feature interaction strength matrix: heatmap of pairwise interaction strength between features, with darker cells indicating stronger interactions. (B) Feature interaction strength expressed as Friedman's H-statistic for each feature, quantifying the proportion of that feature's effect attributable to interactions with other features. (C) Model PDP comparison for the top-ranked feature, showing predicted probability versus feature value for Logistic Regression, Random Forest, XGBoost, and Neural Net. (D) Feature-importance stability across models: feature importance (%) for the leading features as estimated by Logistic Regression (LR), Random Forest (RF), XGBoost (XGB), and Neural Network (NN)

PDP, Partial Dependence Plot

Friedman's H-statistic also indicates that the overall interaction effects between age and years of farming are the highest, suggesting that the impact of years of farming varies significantly depending on the age of the farmer. The partial dependence analysis for these four models (Logistic Regression, Random Forest, XGBoost, and Neural Network) indicates a similar monotonic association between the features and the outcome, which indicates general agreement of the feature-outcome association despite differences in architecture. Divergence is also evident in the feature importance comparisons: For example, usage of pesticides is not among the top 10 most important features, except for models that use the Ensemble, though other models include this feature in their top 10 most important. Similarly, for models other than the Ensemble, the number of years farmers had been farming is among the least important features, while other models include it among their top 10 most important features; this shows that the models differ in the way they learn underlying patterns.

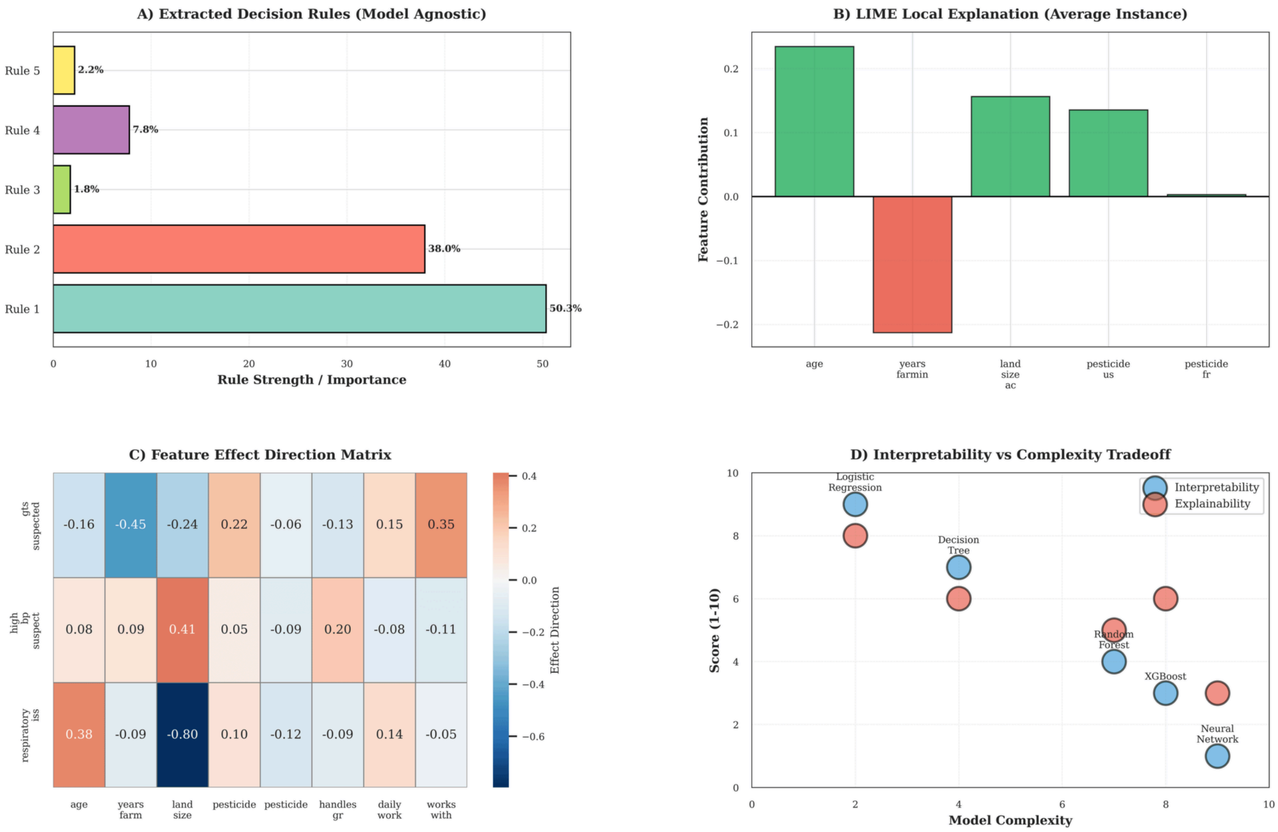

FIGURE 7: Model-agnostic explainability analysis of the agricultural health-prediction model. (A) Extracted decision rules: strength/importance of the five leading model-agnostic decision rules, each bar labelled with its percentage share of total rule importance. (B) LIME local explanation for a representative (average) instance: local interpretable model-agnostic explanation (LIME) feature contributions, where green bars denote positive contributions to the predicted probability and red bars negative contributions. (C) Feature effect direction matrix: heatmap of the direction and magnitude of each feature's effect on suspected GTS, suspected high blood pressure (BP), and suspected respiratory issues; red indicates a positive (risk-increasing) effect and blue a negative (risk-decreasing) effect. (D) Interpretability versus complexity tradeoff: interpretability score (blue) and explainability score (red), each on a 1–10 scale, plotted against model complexity for Logistic Regression, Decision Tree, Random Forest, XGBoost, and Neural Network

In Figure *7*, a multi-level explainability analysis of an agricultural health prediction model is seen using model-agnostic techniques. The model's global decision logic is very sparse: the top two rules have more than 88% of the rule's importance, indicating that few rules make an important decision in determining the prediction. The illustration shows the LIME application to a medium-sized example case, showing that features like age and land size have positive effects on predictions, and years of farming have a strong negative effect. It also shows the relation between the features and the health and well-being outcomes, displaying the impacts that are both meaningful and positive, such as pesticide use and high blood pressure, and meaningful and negative, such as land size and respiratory conditions. There is a tradeoff between model complexity and interpretability, with simpler models being more interpretable and more complex models being more difficult to interpret without the use of post-hoc explanations.

How to cite this article:
Rafi M, Partho P, Mahi M, et al. (August 31, 2026) Explainable Artificial Intelligence (AI) Modeling of Co-Exposure Synergies in Tobacco Farming Health Risk Assessment with Epidemiological Analysis. Cureus J Comput Sci 3 : es44389-026-00251-3. DOI https://doi.org/10.7759/s44389-026-00251-3

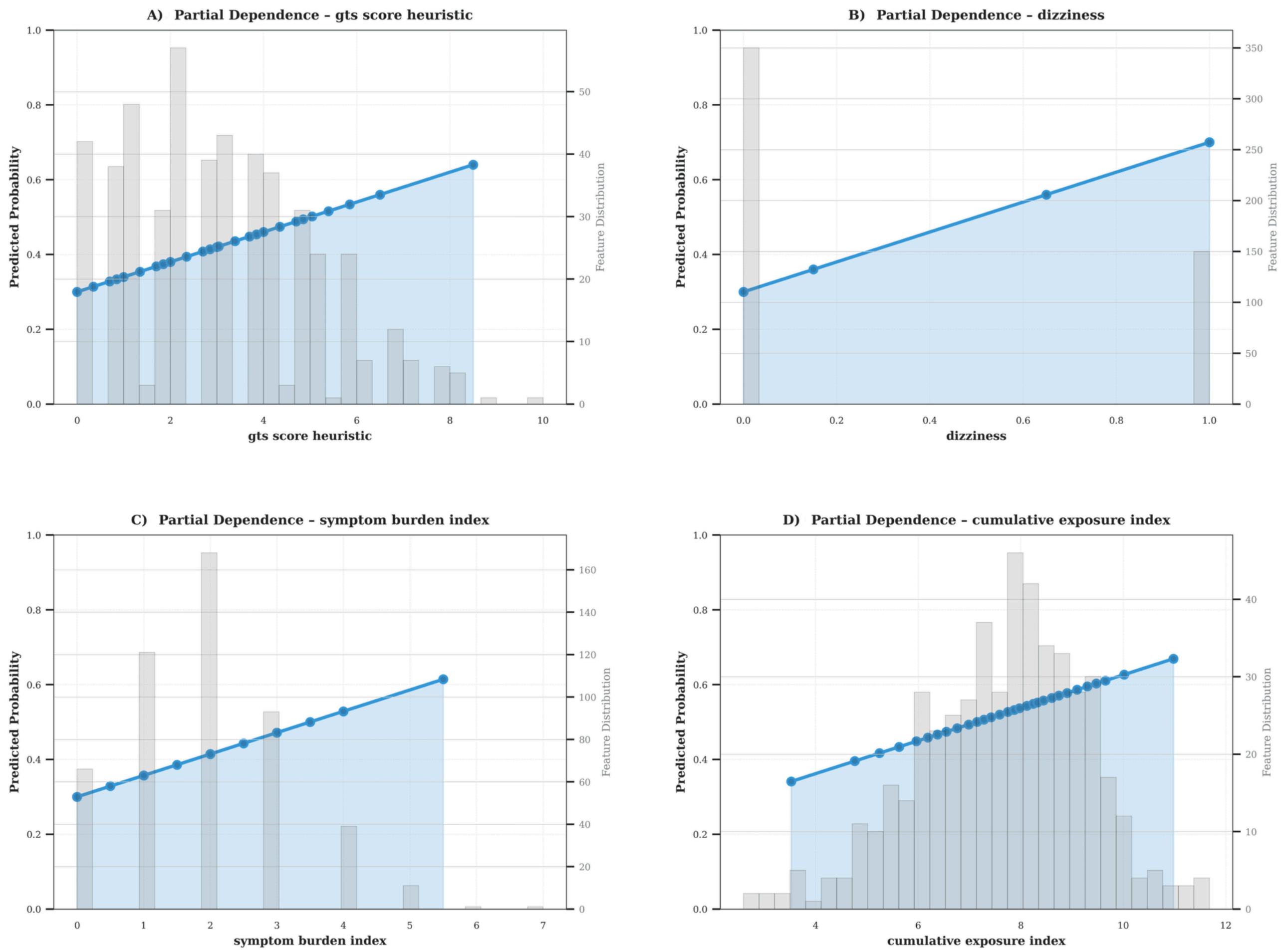

FIGURE 8: One-dimensional partial dependence plots (PDPs) showing the marginal effect of individual features on the model's predicted probability. Panels show the partial dependence for (A) GTS score heuristic, (B) dizziness, (C) symptom burden index, and (D) cumulative exposure index. In each panel the blue line and markers show the partial dependence (predicted probability, left axis) and the grey histogram shows the underlying feature distribution (right axis), indicating data density and hence the reliability of the model response across the feature range. All features show a consistent positive monotonic relationship with predicted risk

GTS, Green Tobacco Sickness

The one-dimensional Partial Dependence Plots in Figure *8* illustrate the marginal effect of individual features on the model's predicted probability, alongside their data distributions for interpretability. The model demonstrates a consistent positive monotonic relationship, meaning that as the values of the features increase, the risk predicted by the model also increases. The data are highly clustered at the low end of the GTS score heuristic and symptom burden index, where the reliability is higher, and the data increase gradually and discretely. Strong model sensitivity, despite a very imbalanced distribution, is evident in the binary effect of dizziness, which manifests itself in the form of a sharp jump in the probability. The cumulative exposure index shows a smoother, more continuous relationship with a more normal distribution over the feature range, indicating good model behavior in the middle range of the features.

Discussion and comparative analysis

The idea for the vulnerability mapping pipeline builds on occupational health assessment efforts among tobacco farmers and aims to consolidate ensemble machine learning techniques, epidemiological interaction analysis, and explainable AI into a single multi-hazard framework. This framework brings together co-exposure scoring and vulnerability archetyping, predictive modeling, and interpretable risk analysis, and is different from earlier research [2,24] that has been more focused on isolated exposures or descriptive analyses. The framework performed well in terms of predictive performance and could be used to detect 3 vulnerability archetypes with distinct exposure characteristics and GTS prevalence, enabling better risk stratification.

This work also builds on previous research in the agricultural health field [19,25,27] by integrating statistical inference with a machine learning-based vulnerability assessment. Regression analyses identified significant associations between physiological stress and hypertension (OR = 4.037, 95% CI: 1.508-10.803, $p = 0.0055$) as well as respiratory conditions (OR = 20.889, 95% CI: 8.146-53.568, $p < 0.001$). Exploratory interaction analyses further suggested potential additive effects for the Dermal × Chemical (SI = 1.30, RERI = 1.61) and Chemical × Physiological (SI = 1.62, RERI = 0.44) exposure combinations. However, because the interaction terms were not statistically significant, these findings should be interpreted cautiously and regarded as hypothesis-generating rather than confirmatory. The archetype-based vulnerability mapping nevertheless provides a practical framework for prioritizing interventions, including the distribution of protective equipment, occupational health screening, pesticide safety training, and heat-stress mitigation for potentially vulnerable farming populations.

Conclusions

The proposed Co-Exposure Vulnerability Mapping Model integrates ensemble machine learning, multi-hazard exposure assessment, vulnerability theory, and explainable artificial intelligence into a unified analytical framework for occupational health risk assessment. The framework demonstrated strong predictive performance and identified distinct vulnerability archetypes representing heterogeneous patterns of exposure and health risk among tobacco farmers. By simultaneously considering dermal, chemical, physiological, and behavioral factors, the model provides a more comprehensive understanding of occupational vulnerability than conventional approaches that evaluate hazards independently. The identified archetypes may help distinguish population groups with different combinations of risk factors and thereby support more context-specific risk assessment and intervention planning.

The exploratory analyses suggested potential interactions between chemical and physiological exposures and between chemical and behavioral exposures. However, these interaction effects were not statistically significant and should therefore be interpreted cautiously and validated using larger, more diverse, and longitudinal populations. The incorporation of model-agnostic explainability techniques enhances transparency by providing insights into feature importance, interaction patterns, and model consistency during predictive decision-making. This interpretability may facilitate the responsible use of machine learning findings by researchers, healthcare professionals, and policymakers. The proposed framework offers an interpretable, scalable, and data-driven approach to occupational health risk stratification and may support the early identification of vulnerable groups, efficient allocation of preventive resources, and development of targeted interventions in agricultural communities.

Future studies should build on this framework by adopting longitudinal cohort designs to examine temporal exposure patterns, disease trajectories, and seasonal variations across diverse agricultural settings and regions. Further research could also explore Bayesian and randomized hyperparameter optimization methods to improve model performance, computational efficiency, and generalizability.

Additional Information

Author Contributions

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

Concept and design: Md. Abid Hasan Rafi

Acquisition, analysis, or interpretation of data: Md. Abid Hasan Rafi, Pijush Kanti Roy Partho, Md. Najib Ul Azam Mahi, Pankaj Bhowmik, Tangina Sultana, Md. Delowar Hossain

Drafting of the manuscript: Md. Abid Hasan Rafi, Pijush Kanti Roy Partho, Md. Najib Ul Azam Mahi

Critical review of the manuscript for important intellectual content: Md. Abid Hasan Rafi, Pijush Kanti Roy Partho, Md. Najib Ul Azam Mahi, Pankaj Bhowmik, Tangina Sultana, Md. Delowar Hossain

Supervision: Pankaj Bhowmik, Tangina Sultana, Md. Delowar Hossain

Disclosures

Human subjects: Consent was obtained or waived by all participants in this study. Institute of Research and Training (IRT) issued approval SRG-252351. This study involved human participants through a questionnaire survey of 500 farmers. The study was reviewed and approved by the Institute of Research and Training (IRT), HSTU, Bangladesh. Participation was voluntary, informed consent was obtained from all participants prior to data collection, and all responses were anonymized and treated confidentially. **Animal subjects:** All authors have confirmed that this study did not involve animal subjects or tissue. **Conflicts of interest:** In compliance with the ICMJE uniform disclosure form, all authors declare the following: **Payment/services info:** All authors have declared that no financial support was received from any organization for the submitted work. **Financial relationships:** All authors have declared that they have no financial relationships at present or within the previous three years with any organizations that might have an interest in the submitted work. **Other relationships:** All authors have declared that there are no other relationships or activities that could appear to have influenced the submitted work.

Data Availability Statements

The data (and/or code) supporting this study are openly available in TF-MHOV: Tobacco Farmers' Multi-Hazard Occupational Health and Vulnerability Dataset at https://doi.org/10.7910/DVN/LAIQUT, reference 10.7910/DVN/LAIQUT.

References

1. Ali MY, Abdulla, Al Kafy A, et al.: Comparative occupational health risk between tobacco and paddy farming people in Bangladesh. SSM - Mental Health. 2022, 2:100061. 10.1016/j.ssmmh.2022.100061
2. Kumar AK, Kumar A, Vaidhyanathan R, Kushwaha HL, Chakraborty D, Tomar BS: Assessments of occupational health hazards of tobacco farmworkers in Andhra Pradesh, India. Toxicology and Environmental Health Sciences. 2023, 15:335-344. 10.1007/s13530-023-00186-5
3. Guindo Y, Parent ME, Richard H, Luce D, Barul C: Expert-based assessment of chemical and physical exposures, and organizational factors, in past agricultural jobs. Environmental Research. 2024, 263:120238. 10.1016/j.envres.2024.120238
4. Dalberto D, Alves J, Garcia ALH, et al.: Exposure in the tobacco fields: Genetic damage and oxidative stress in tobacco farmers occupationally exposed during harvest and grading seasons. Mutation Research - Genetic Toxicology and Environmental Mutagenesis. 2022, 878:503485. 10.1016/j.mrgentox.2022.503485
5. Martins-da-Silva AS, Torales J, Becker RFV, et al.: Tobacco growing and tobacco use. International Review of Psychiatry. 2022, 34:51-58. 10.1080/09540261.2022.2034602
6. Lencucha RA, Vichit-Vadakan N, Patanavanich R, Ralston R: Addressing tobacco industry influence in tobacco-growing countries. Bulletin of the World Health Organization. 2023, 102:58-64. 10.2471/blt.23.290219
7. Zhang J, Wang J, Chen S, Tang S, Zhao W: Multi-hazard meteorological disaster risk assessment for agriculture based on historical disaster data in Jilin Province, China. Sustainability. 2022, 14:7482. 10.3390/su14127482
8. Fassa AG, Faria NMX, Szortyka ALSC, Meucci RD, Fiori NS, Carvalho MPD: Child labor in family tobacco farms in southern Brazil: occupational exposure and related health problems. International Journal of Environmental Research and Public Health. 2021, 18:12255. 10.3390/ijerph182212255

9. Ali MY, Shahrier M, Abdulla-Al Kafy A, et al.: Environmental impact assessment of tobacco farming in northern Bangladesh. Heliyon. 2023, 9:e14505. 10.1016/j.heliyon.2023.e14505
10. Kumar AK, Kumar A, Vaidhyanathan R, Kushwaha HL, Chakraborty D, Tomar BS: Demographic and socio-economic of the tobacco farmworkers in Andhra Pradesh, India. The Pharma Innovation Journal. 2023, 12:1090-1092.
11. Mim NA, Hasan SS, Hoque MZ, Ahmed M, Chakma P: Tobacco farmers' perceptions of unsafe tobacco cultivation and its effect on health and environment: a case of Chittagong Hill Tracts, Bangladesh. Clean Technologies. 2024, 6:586-601. 10.3390/cleantechnol6020031
12. Piyankarage SC, McGahee E, Feng J, Blount BC, Wang L: Automated solid phase extraction and polarity-switching tandem mass spectrometry technique for high throughput analysis of urine biomarkers for 14 tobacco-related compounds. ACS Omega. 2021, 6:30901-30909. 10.1021/acsomega.1c02543
13. Akakuru OC, Adakwa CB, Ikoro DO, et al.: Application of artificial neural network and multi-linear regression techniques in groundwater quality and health risk assessment around Egbema, Southeastern Nigeria. Environmental Earth Sciences. 2023, 82:77. 10.1007/s12665-023-10753-1
14. Paul A, Sultana NN, Nazir N, Das BK, Jabed MA, Nath TK: Self-reported health problems of tobacco farmers in south-eastern Bangladesh. Journal of Public Health. 2021, 29:595-604. 10.1007/s10389-019-01159-0
15. Moradhaseli S, Ataei P, Van den Broucke S, Karimi H: The process of farmers' occupational health behavior by health belief model: evidence from Iran. Journal of Agromedicine. 2020, 26:231-244. 10.1080/1059924x.2020.1837316
16. Fiałkowska K, Matuszczyk K: Safe and fruitful? Structural vulnerabilities in the experience of seasonal migrant workers in agriculture in Germany and Poland. Safety Science. 2021, 139:105275. 10.1016/j.ssci.2021.105275
17. Burgess R, Nyhan K, Freudenberg N, Ransome Y: Corporate activities that influence population health: a scoping review and qualitative synthesis to develop the HEALTH-CORP typology. Globalization and Health. 2024, 20:77. 10.1186/s12992-024-01082-4
18. Loftus TJ, Shickel B, Balch JA, et al.: Phenotype clustering in health care: A narrative review for clinicians. Frontiers in Artificial Intelligence. 2022, 5:842306. 10.3389/frai.2022.842306
19. Alotaibi A: Ensemble deep learning approaches in health care: a review. Computers, Materials & Continua. 2025, 82:3741-3771. 10.32604/cmc.2025.061998
20. Nguyen DK, Lan CH, Chan CL: Deep ensemble learning approaches in healthcare to enhance the prediction and diagnosing performance: the workflows, deployments, and surveys on the statistical, image-based, and sequential datasets. International Journal of Environmental Research and Public Health. 2021, 18:10811. 10.3390/ijerph182010811
21. Park J, Ahn H, Youn K, Lee M, Hong S: Ensemble learning to identify depression indicators for Korean farmers. IEEE Access. 2023, 11:118787-118800. 10.1109/access.2023.3327126
22. Manners R, Hammond J, Umugabe DR, Sibomana M, Schut M: A farm typology development cycle: From empirical development through validation, to large-scale organisational deployment. Agricultural Systems. 2025, 224:104250. 10.1016/j.agsy.2024.104250
23. Sehgal N, Brennan PA, Dunlop AL, et al.: Associations of prenatal tobacco and insecticide co-exposures with neurobehavioral responses among children born to pregnant women exposed to cannabis. Neurotoxicology and Teratology. 2025, 111:107536. 10.1016/j.ntt.2025.107536
24. Han MK, Ye W, Wang D, et al.: Bronchodilators in tobacco-exposed persons with symptoms and preserved lung function. New England Journal of Medicine. 2022, 387:1173-1184. 10.1056/nejmoa2204752
25. Fayshal MA, Ullah MR, Fairooz HM, Rahman SMA, Siddique I Md: Evaluating multidisciplinary approaches within an integrated framework for human health risk assessment. Journal of Environmental Engineering and Studies. 2023, 8:30-41. 10.46610/joees.2023.v08i03.004
26. Chopin P, Mubaya CP, Descheemaeker K, Öborn I, Bergkvist G: Avenues for improving farming sustainability assessment with upgraded tools, sustainability framing and indicators. A review. Agronomy for Sustainable Development. 2021, 41:19. 10.1007/s13593-021-00674-3
27. Upadhayay VK, Chitara MK, Mishra D, et al.: Synergistic impact of nanomaterials and plant probiotics in agriculture: A tale of two-way strategy for long-term sustainability. Frontiers in Microbiology. 2023, 14:1133968. 10.3389/fmicb.2023.1133968
28. Etemadi A, Abnet CC, Dawsey SM, Freedman ND: Biomarkers of tobacco carcinogenesis in diverse populations: challenges and opportunities. Cancer Epidemiology, Biomarkers & Prevention. 2023, 32:289-291. 10.1158/1055-9965.epi-22-1289

29. TF-MHOV: Tobacco Farmers' Multi-Hazard Occupational Health and Vulnerability Dataset. (2026). Accessed: June 6, 2026: https://doi.org/10.7910/DVN/LAIQUT.

30. Rafi MAH, Foysal MAS: Statistical benchmarking of machine learning methods in fisheries morphological analysis. National Academy Science Letters. 2026, 10.1007/s40009-026-02314-4

How to cite this article:
Rafi M, Partho P, Mahi M, et al. (August 31, 2026) Explainable Artificial Intelligence (AI) Modeling of Co-Exposure Synergies in Tobacco Farming Health Risk Assessment with Epidemiological Analysis. Cureus J Comput Sci 3 : es44389-026-00251-3.
DOI https://doi.org/10.7759/s44389-026-00251-3