

Article

Research on an Improved SVM-RF-Based Risk Assessment Algorithm for Infectious Substances at Port of Entry

Jin Li ^{1,*}, Chen Li ^{1,*} , Yong Bian ², Fengze Wu ¹ and Jie Tian ²

¹ College of Energy Environment and Safety Engineering, China Jiliang University, Hangzhou 310018, China; jinli@cjl.u.edu.cn (J.L.); fengzewu@cjl.u.edu.cn (F.W.)

² China Customs Science and Technology Research Center, Beijing 100026, China

* Correspondence: lichen@cjl.u.edu.cn; Tel.: +86-15869197511

Abstract

The wide variety of infectious substances encountered at ports of entry, coupled with complex risk profiles and the challenges of subjective identification, make it difficult for assessors to conduct rapid, accurate, and objective evaluations, particularly given limitations in expertise and experience. To address this challenge and to develop a highly generalizable risk assessment model for infectious substances, this study draws on a five-year case database of risk incidents at Beijing Customs ports. Frequency analysis was first employed to identify key risk factors associated with infectious substances entering ports. Subsequently, risk assessment models were constructed using decision tree, random forest, and support vector machine algorithms, as well as an improved SVM-RF algorithm, followed by an analysis of feature importance. The results demonstrate that the improved SVM-RF algorithm achieved superior generalization performance, with an evaluation accuracy of 0.93. To further validate its applicability and feasibility, the improved model was applied to real cases of infectious substances intercepted at Beijing Customs ports, where its risk level classifications were consistent with expert assessments. These findings provide a valuable reference for improving the customs safety assessment system for special biological resources and for mitigating the risks posed by infectious substances at ports of entry.

Keywords: port infectious substance risk; risk level assessment; SVM-RF; machine learning



Received: 5 October 2025

Revised: 30 October 2025

Accepted: 30 October 2025

Published: 6 November 2025

Citation: Li, J.; Li, C.; Bian, Y.; Wu, F.; Tian, J. Research on an Improved SVM-RF-Based Risk Assessment Algorithm for Infectious Substances at Port of Entry. *Sustainability* **2025**, *17*, 9900. <https://doi.org/10.3390/su17219900>

Copyright: © 2025 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

1. Introduction

In recent years, the rapid growth of the global biopharmaceutical industry has accelerated the cross-border movement of special biological resources. Among them, infectious substances present a particularly significant challenge for the precise and efficient supervision of customs ports, owing to the complexity and diversity of their associated risks. Ensuring efficient risk assessment and accurate risk identification during port supervision has therefore become a pressing technological challenge in the regulation of infectious substances at ports of entry. Addressing this challenge is essential for customs authorities to safeguard national border security and to support the sustainable development of the biotechnology industry [1].

Machine learning methods have been widely employed in security risk assessment and risk grading research because of their strong capability to process complex data structures and to uncover potential risk patterns. These advantages make them an important technical approach for building scientific and rational risk assessment systems. Previous studies [2–8] have integrated Fault Tree Analysis (FTA) and frequency statistics with methods such as

Event Sequence Diagrams (ESD), Bayesian Networks (BN), risk matrix approaches, and fuzzy theory. These combined methodologies have been applied in safety risk assessments across diverse domains, including lightning strike incidents, urban gas systems, desulfurization processes, and hemodialysis. Such applications have effectively strengthened the systematic nature of risk identification and improved the accuracy of assessment outcomes. Studies [9–14] have extensively applied decision tree algorithms in risk assessment across diverse domains, including agricultural supply chains, diabetes prediction, underground engineering permeability incidents, tree collapse, and cognitive impairment among the elderly. Findings suggest that pruning optimization and integration with algorithms such as random forest or Bayesian networks can significantly enhance the generalization ability, accuracy, and sensitivity of decision tree models, thereby improving their practicality and reliability in various risk prediction scenarios. Studies [15–19] highlight the widespread application and strong performance of the Random Forest (RF) algorithm in risk assessment across multiple domains, including urban water pipeline bursts, earthquake information release safety, power cable repair decisions, landslide hazard evaluation, and wildfire risk prediction. By integrating diverse factors such as weather conditions, vegetation types, and infrastructure, RF models have demonstrated both high accuracy and robust generalization ability in various case studies. In some applications, the model achieved a maximum fitting rate of 99.70%, offering effective decision support for infrastructure maintenance, disaster management, and ecosystem risk identification. Collectively, these studies underscore the advantages of RF in handling nonlinear, high-dimensional, and imbalanced data, making it particularly suitable for multi-source risk modeling and prediction in complex scenarios. Recent studies have demonstrated that Support Vector Machine (SVM) algorithms exhibit remarkable performance in infectious disease prediction and clinical risk modeling. For instance, Lilhore et al. developed a hybrid random forest–SVM model for hepatitis C virus (HCV) classification, achieving > 96% accuracy through the integration of feature ranking and oversampling techniques to address data imbalance [20]. Similarly, Ghazal et al. applied a fine Gaussian SVM to predict hepatitis C staging, obtaining 97.9% accuracy and effectively assisting in early diagnosis [21]. Wu et al. constructed multiple machine learning models, including SVM, for the preoperative identification of urinary infection stones, demonstrating robust discrimination performance with AUC values up to 0.77 [22]. In addition, He et al. employed SVM to predict HIV infection risk among men who have sex with men (MSM), providing a data-driven approach to enhance HIV surveillance and prevention strategies [23]. Sun et al. further used SVM and other machine learning methods to predict post-chemotherapy lung infection in cancer patients, illustrating the effectiveness of SVM-based algorithms in clinical infection risk assessment [24]. Collectively, these studies highlight the robust generalization ability and strong classification performance of SVM in complex infectious disease prediction tasks, providing a solid foundation for its application in port-of-entry infectious substance risk assessment. These advances provide effective technical support for risk prevention and control in meteorological disasters, urban safety management, healthcare, and industrial production. However, despite the maturity of SVM-based risk assessment models in multiple domains, research addressing the complex risk characteristics associated with imbalanced and multidimensional datasets of infectious substances at ports of entry remains limited.

Meanwhile, recent studies have demonstrated that hybrid models integrating Support Vector Machine (SVM) and Random Forest (RF) algorithms can further enhance prediction accuracy and robustness across various complex domains. For instance, Xu et al. improved rural land-cover classification performance by combining SVM and RF using the C5.0 algorithm, achieving higher accuracy than conventional classifiers such as XGBoost [25]. Similarly, Sivaraja et al. developed an RF–Least Squares SVM model to predict the anti-carbonation

performance of concrete with an R^2 of 0.999 [26], while Gelmez et al. applied an iterative SVM-RF approach to identify autism-related gene signatures with 92% diagnostic accuracy [27]. These findings collectively indicate that integrating SVM and RF can effectively exploit the strengths of both algorithms—leveraging RF's feature selection and ensemble learning capabilities alongside SVM's high-dimensional nonlinear classification power—to improve model generalization, stability, and interpretability. Building upon these insights, the present study introduces an improved SVM-RF model tailored to the complex, imbalanced, and multidimensional datasets characteristic of infectious substances at ports of entry.

Unlike previous research achievements in fields such as meteorological disasters, industrial production, and infrastructure safety, existing machine learning methods all exhibit certain limitations in the risk assessment of infectious substances at ports of entry. The primary reason lies in the fact that such risk subjects exhibit both high complexity and multidimensional characteristics. Simultaneously, certain high-risk or special-category infectious substances occur with extremely low frequency, resulting in significantly imbalanced sample data. Therefore, this paper employs the accident case risk factor frequency statistical method to systematically organize infectious substance risk factors, logically and intuitively presenting safety risk failure pathways to clarify potential risks at each stage. By comparing simulation results from decision tree algorithms, random forests, support vector machines, and hybrid RF-SVM models, an innovative improved SVM-RF model was constructed. The model's generalization performance was analyzed using four metrics: precision, recall, specificity, and F1-score, followed by comparative evaluation of the prediction models.

Therefore, this study aims to develop a data-driven and generalizable risk assessment model for infectious substances intercepted at ports of entry by integrating the strengths of Support Vector Machine (SVM) and Random Forest (RF) algorithms. By combining the ensemble learning and feature extraction capability of RF with the nonlinear classification advantage of SVM, the proposed approach seeks to overcome the limitations of traditional machine learning models that struggle with high-dimensional, imbalanced, and heterogeneous data typically observed in port supervision scenarios. The ultimate goal is to enhance both the accuracy and interpretability of infectious substance risk classification, thereby supporting more scientific and objective decision-making in customs biosecurity management.

To achieve this objective, the study focuses on addressing the following research questions.

First, what are the key risk factors that significantly influence the classification and assessment of infectious substances at ports of entry? Understanding these factors is essential for constructing a representative feature set and improving the transparency of the model's decision process.

Second, how can a hybrid SVM-RF algorithm be optimized to enhance classification accuracy and generalization performance compared with conventional single-algorithm approaches? This question aims to explore algorithmic optimization and model integration strategies to improve predictive robustness.

Third, to what extent can the proposed model align with expert evaluations and enhance the objectivity and efficiency of customs risk management for infectious substances? This inquiry evaluates the model's practical applicability and its contribution to intelligent risk-based supervision.

By systematically addressing these research questions, this study establishes a scientific and interpretable framework for risk identification and decision-making in customs biosecurity supervision, contributing to the advancement of data-driven approaches for port-based infectious substance management.

2. Machine Learning Model Theory and Optimization

Decision tree algorithms are characterized by an intuitive tree structure and clear rule interpretability, enabling effective handling of mixed numerical and categorical data while maintaining high training efficiency. The random forest (RF) algorithm utilizes a bagging framework that integrates multiple decision trees, substantially reducing variance through dual randomization of samples and feature subsets, thereby achieving both high accuracy and strong resistance to overfitting. Support Vector Machine (SVM) algorithms construct optimal classification hyperplanes based on the principle of structural risk minimization, demonstrating strong generalization performance in small-sample, high-dimensional datasets and showing particular suitability for complex pattern recognition with well-defined decision boundaries. The fusion of SVM and Random Forest (SVM-RF) combines the capability of SVM to construct optimal separating hyperplanes in high-dimensional spaces with the ensemble advantages of RF in reducing variance and enhancing generalization, aiming to further improve overall model performance. To systematically evaluate the advantages and applicability of these algorithms, this study constructs classification models using four methods—decision trees, random forests, SVMs, and the improved SVM-RF—for comparative analysis.

2.1. Fundamental Theory of Machine Learning Algorithms

To facilitate a clear understanding of the theoretical basis underlying the model development process, this section first introduces the fundamental concepts of several representative machine learning algorithms used in this study. These include the decision tree, random forest, and support vector machine (SVM) models, each of which provides distinct advantages in handling classification and prediction tasks involving complex, multidimensional data. A schematic overview of their theoretical relationships and structural characteristics is presented in Figure 1, serving as the foundation for subsequent model construction and performance comparison.

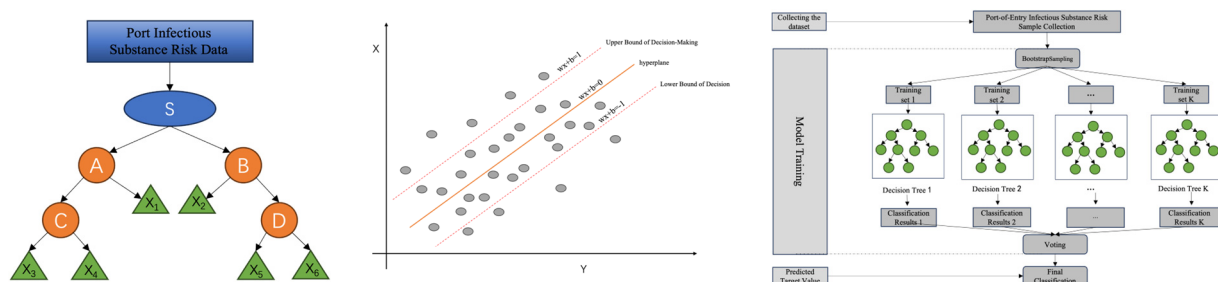


Figure 1. Schematic Diagram of the Theoretical Model for Machine Learning Models.

- (1) A decision tree is a supervised machine learning algorithm that represents decision rules and classification outcomes using a hierarchical tree structure. Structurally, it consists of nodes and edges arranged in a top-down, tree-like format, where internal nodes correspond to feature-based decision criteria, branches denote decision paths, and leaf nodes indicate classification results.
- (2) Support Vector Machine (SVM) is a supervised learning algorithm grounded in the principle of structural risk minimization. Its central concept is to construct a maximum-margin hyperplane that separates classes in the feature space, thereby achieving strong generalization performance for classification or regression tasks. The optimal hyperplane maximizes the margin between sample classes, enhancing the model's robustness. The classifier is defined by support vectors, which are the data points closest to the separating hyperplane that determine the decision boundary.

- (3) Random Forest (RF) is an ensemble learning algorithm based on the Bagging framework, employing decision trees as base learners. The algorithm generates multiple training subsets through bootstrap sampling from the original dataset and selects random feature subsets as candidates for splitting during the construction of each tree. By aggregating the predictions of all trees, RF effectively reduces overfitting and improves generalization, predictive accuracy, and model stability.

2.2. Improved SVM-RF Algorithm

Support Vector Machine-Random Forest (SVM-RF) is an innovative ensemble learning framework that constructs robust predictive models by combining the high-dimensional mapping capability of Support Vector Machines (SVM) with the ensemble decision-making advantages of Random Forests (RF). In the context of border inspection risk assessment for infectious substances, the data are characterized by high dimensionality, multi-source origins, and significant class imbalance. Direct application of conventional machine learning models often results in overfitting or insufficient recognition of minority-class risks, thereby undermining predictive performance. To address these challenges, this study proposes an improved SVM-RF algorithm that integrates RF with SVM to enable precise modeling and efficient identification of infectious substance risks at border ports.

To ensure the reliability and generalization of the improved SVM-RF model, a systematic hyperparameter optimization process was implemented. Specifically, a grid search combined with 10-fold cross-validation was employed to identify the optimal parameter settings for both the RF and SVM components. For the Random Forest, the number of trees and the maximum depth of trees were tuned within the ranges of 50–500 and 3–15, respectively. For the SVM with an RBF kernel, the penalty coefficient (C) and kernel width (γ) were explored within the ranges of 0.1–1000 and 10^{-4} –10, respectively. The combination yielding the highest average F1-score on the validation set was selected as the final model configuration. This approach ensures that the optimized parameters are not overfitted to the training data and that the model maintains robust predictive performance across different risk categories of infectious substances.

The random forest component shown in Figure 2 is primarily used for feature selection and importance evaluation. The prediction function of the random forest can be expressed as

$$f(x) = \frac{1}{T} \sum_{t=1}^T f_t(x). \quad (1)$$

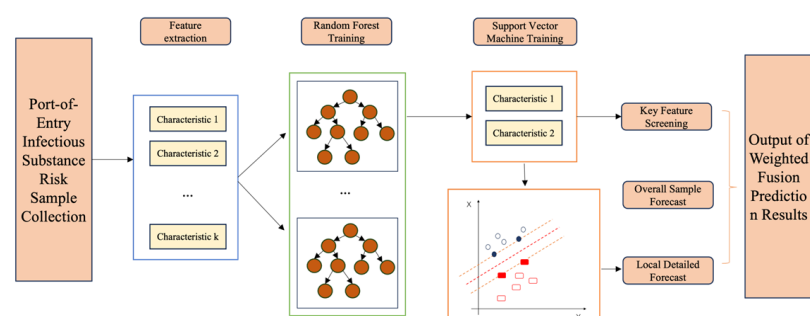


Figure 2. Risk Assessment Model for Infectious Substances at Ports of Entry Based on SVM-RF.

Here, $f(x)$ represents the final prediction result, T denotes the number of decision trees, and $f_t(x)$ is the output of the t -th decision tree. During training, the random forest measures feature importance by calculating each feature's contribution to reducing Gini impurity across all decision trees. Leveraging the infectious substance case repository, this study employs this mechanism for recursive feature selection of risk factors. Variables

with minimal contribution to model classification are progressively eliminated, yielding an optimal feature subset that effectively reflects port risk characteristics.

After completing feature selection, support vector machines are employed to construct the final classification model. Its optimization objective is

$$\min_{w,b} \frac{1}{2} \|w\|^2, \text{ s.t. } y_i(w \cdot x_i + b) \geq 1, i = 1, 2, \dots, n. \quad (2)$$

Here, w denotes the weight vector, b represents the bias term, and (x_i, y_i) signifies the input sample and its corresponding risk label. Through kernel function mapping, the SVM projects the multidimensional risk factors of infectious substances into a high-dimensional feature space. Within this space, it locates the optimal classification hyperplane, enabling effective differentiation between high-risk and low-risk samples.

2.3. Algorithm

Platform: Python 3.11; Libraries: Scikit-learn, Pandas, NumPy

Input: Complete dataset $D = [X, Y]$

Feature matrix $X = (x_1, x_2, \dots, x_n)^T$; Label vector $Y = (y_1, y_2, \dots, y_n)^T$

Output: Optimized model parameters (C, γ), Confusion Matrix, Accuracy (A), Precision (P), Recall (R), and F1-score (F1)

- (1) Data Partitioning: Split X and Y into training and testing subsets: $X_{train}, X_{test}, Y_{train}, Y_{test}$.
- (2) Initialize Random Forest (RF): Set a fixed random seed and train the Random Forest Classifier on X_{train}, Y_{train} .
- (3) Feature Importance Evaluation: Compute Gini-based feature importance; eliminate redundant or low-contribution variables.
- (4) Feature Selection: Retain the top-ranked features to form the optimized feature subset F .
- (5) Generate RF Predictions: Obtain preliminary prediction probabilities and feature importance weights from the trained RF model.
- (6) Construct SVM Input Set: Combine the selected feature subset F and RF-derived prediction probabilities as new input data F' .
- (7) Parameter Optimization: Perform grid search with cross-validation to tune SVM parameters C and γ under an RBF kernel.
- (8) Model Training: Train the optimized SVM model using F'_{train}, Y_{train} .
- (9) Model Evaluation: Apply the trained model to F'_{test}, Y_{test} ; compute A, P, R, and F1.
- (10) Result Output: Return the confusion matrix, accuracy, precision, recall, and F1-score.

As illustrated in Figure 3, the improved SVM-RF model adopts a two-stage fusion framework for assessing infectious substance risks at border ports. The figure presents the overall workflow of the algorithm, where the Random Forest (RF) component performs feature selection and importance evaluation, and the Support Vector Machine (SVM) component conducts the final classification.

This model employs a two-stage fusion mechanism for assessing infectious substance risks at border ports. In the first stage, a Random Forest (RF) is trained on multidimensional risk factors. By aggregating predictions from multiple decision trees through majority voting, it generates preliminary prediction probabilities. Feature importance for each risk factor is calculated based on Gini impurity reduction, effectively removing redundant variables and mitigating noise caused by sample imbalance. In the second stage, the RF-generated predictions and importance weights are incorporated as new features into a Support Vector Machine (SVM). Utilizing a Radial Basis Function (RBF) kernel allows nonlinear mapping, enabling the construction of an optimal separating hyperplane in high-dimensional space that maximizes classification margins among different categories of infectious substances. This staged fusion mechanism preserves the strengths of RF in

feature selection and generalization while leveraging the precision of SVM for nonlinear classification in high-dimensional spaces. Compared with alternative models, it achieves more efficient identification and accurate prediction of infectious substance risk levels at ports.

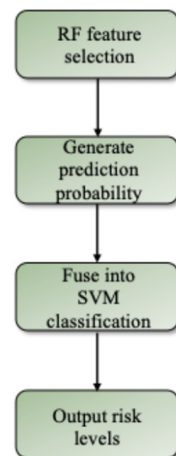


Figure 3. Improved SVM-RF fusion algorithm for infectious substance risk assessment at ports of entry.

3. Sample Data Sources and Data Processing

The import and export of infectious substances involve multiple dimensions of operational information, including biological characteristics, inspection approvals, packaging features, external environmental conditions, and emergency management protocols. The sources of risk in this process are highly interconnected and multidimensional. Risk factors encompass not only technical elements, such as sensitivity deviations in detection equipment and algorithmic limitations in customs data analysis systems, but also management and societal components. These include the professional competence of frontline quarantine personnel, interdepartmental coordination mechanisms, adherence to international standards for infectious substances, and public awareness regarding biohazardous materials.

3.1. Case Statistics and Risk Factor Selection

This study employs a frequency-based statistical approach to analyze risk factors in accident cases. For risk tracing of special biological resources, it is essential to systematically integrate equipment operation logs, quarantine anomaly records, historical violation cases, and routine monitoring data. The workflow of this assessment method is illustrated in Figure 4.

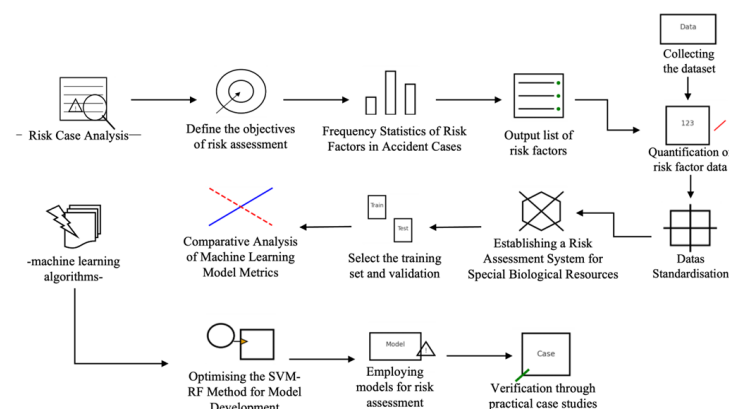


Figure 4. Port-of-Entry Infectious Substance Risk Assessment Process.

By analyzing the case database of risk incidents at Beijing Customs over the past five years, this study identifies eight critical factors that should be prioritized in constructing a risk analysis model: the intrinsic characteristics of infectious substances, transportation capacity requirements, risk profiles of their origins, laboratory safety levels at user facilities, intended import purposes, completeness of customs declaration materials, emergency response speed, and equipment readiness at user facilities. A multi-parameter coupled dynamic assessment system is then established to enable comprehensive risk analysis and tiered early warning for infectious substances. The classification of primary influencing factors is illustrated in Figure 5.

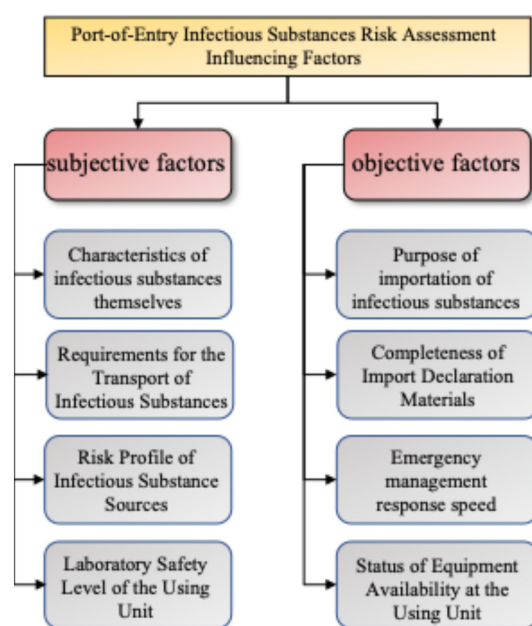


Figure 5. Classification of Risk Factors for Infectious Substances.

3.2. Dataset Quantification and Standardization

The infectious substance risk assessment dataset constructed in this study primarily draws on risk cases involving infectious substances intercepted at Beijing Customs over the past five years. Due to incomplete records in certain risk factor dimensions, a portion of the raw inspection data contains missing values. Direct use of such incomplete data for model training may compromise the accuracy and stability of predictive models due to insufficient valid samples.

To address this issue, a detailed analysis of missing value distribution was first conducted. Missing values were primarily concentrated in three feature dimensions—emergency response speed (8%), equipment readiness (6%), and completeness of customs declaration materials (5%), while the remaining variables exhibited less than 2% missingness. The missingness pattern followed a Missing at Random (MAR) distribution, indicating no systematic bias in data collection.

To ensure data integrity, three imputation strategies—mean imputation, k-nearest neighbors (KNN) imputation, and correlation-based interpolation—were comparatively evaluated using ten-fold cross-validation. The results showed that KNN imputation produced slightly better performance, improving model accuracy by approximately 0.8–1.2% compared with the other two methods, while maintaining consistent classification trends across all algorithms. Based on these results, KNN imputation was adopted as the final preprocessing approach, as it effectively preserved the local data structure and minimized bias.

Furthermore, key indicators for high-risk sample types were manually supplemented and corrected through field research and expert consultation. This comprehensive data

refinement process ultimately produced 500 complete and standardized data samples suitable for machine learning modeling. In accordance with the nationally prescribed risk classification table for special biological resources, the risks of infectious substances at ports were categorized using the grading criteria presented in Table 1.

Table 1. Classification Criteria for Infectious Substances at Ports of Entry.

Risk Level	Risk Level	Comprehensive Score Range
1	Low risk	[0–30]
2	Moderate risk	[31–60]
3	High risk	[61–80]
4	Extremely high risk	[81–100]

Through the collection and analysis of 500 infectious substance risk assessment cases, eight key risk factors were identified. Based on the specific characteristics of infectious substance risks, the corresponding risk factor indicators were quantified, as summarized in Table 2.

Table 2. Quantification of Risk Factors for Infectious Substances.

Risk Factors	Quantification Method
Characteristics of Infectious Substances Themselves	Refer to China’s Catalogue of Pathogenic Microorganisms Transmissible Among Humans, Categories I to IV (4 = Extremely High Risk (A), 3 = High Risk (B), 2 = Medium Risk (C), 1 = Low Risk (D))
Laboratory Safety Level for Units Handling Infectious Substances	Refer to P1, P2, P3, P4 laboratory standards (5 = P4 laboratory, 4 = P3 laboratory, 3 = P2 laboratory, 2 = P1 laboratory, 1 = other general laboratories)
Requirements for the Transport of Infectious Substances	Reference for Transporting Infectious Substances (5 = Highest Requirements (Class A Packaging), 4 = High Requirements (Class A Packaging), 3 = Standard Requirements (Class A Packaging), 2 = Low Requirements (Class B Packaging), 1 = Minimum Requirements (Class B Packaging))
Risk Assessment of Infectious Substance Sources	Refer to the remote sensing map of infectious disease distribution released by the WHO over the past six months (5 = highest-risk region (Africa), 4 = high-risk region (Southeast Asia), 3 = moderate-risk region (Mediterranean region), 2 = low-risk region (Pacific region), 1 = lowest-risk region (North America and Europe)).
Purpose of Importing Infectious Substances	Reference to Annual Surveys on the Purpose of Importing Infectious Materials (Classified into Five Levels: V = Vaccine Research and Drug Development, IV = Scientific Research by Schools and Enterprises, III = Clinical Disease Diagnosis and Testing, II = Reference Materials or Quality Control, I = School Education and Training)
Completeness of Materials for Import Declaration of Infectious Substances	Classified into five levels: I–V (V = No data available, IV = Incomplete data, III = Important data missing, II = Relatively complete data, I = Complete data)
Emergency Response Speed for Infectious Substances	Classified into five levels: I–V (V = slow speed, IV = relatively slow speed, III = normal speed, II = relatively fast speed, I = fast speed)
Infectious Material Handling Unit Equipment Readiness Status	Classified into five levels: I–V (V = Long-term lack of inspection, equipment missing or damaged; IV = Recent lack of inspection, equipment partially missing; III = No regular inspection, equipment relatively intact (high risk); II = Regular inspection, equipment relatively intact; I = Regular inspection, equipment intact)

Prior to model construction, the dataset underwent a comprehensive data quality inspection. Missing values were identified and addressed through appropriate imputation methods, and outliers were examined using interquartile range (IQR) and z-score analysis to ensure data consistency. The dataset was further analyzed for class distribution to assess balance among different risk levels of infectious substances. The results indicated that the dataset was moderately imbalanced. To mitigate this imbalance and improve model generalization, a combination of stratified sampling and class-weight adjustment strategies

was employed during model training. These preprocessing steps ensured the robustness and reliability of the subsequent machine learning analysis.

In the machine learning workflow, dataset normalization is a crucial preprocessing step. It constrains data within a defined range through mathematical transformations, eliminating differences in scale and magnitude among feature variables and producing a dimensionless dataset. This ensures that all features are on a comparable scale, facilitating analysis and the development of reliable predictive models. In practical applications, dataset features often exhibit varying magnitudes and distribution patterns; normalization effectively mitigates the impact of these differences on model performance. In this study, Z-score normalization (standard deviation normalization) is employed, as defined in Equations (3) and (4). In this process, X represents the original data, and Z denotes the normalized value. For clarity, the variables are defined as follows: x_i represents the i -th data point in the dataset. μ is the population mean of the dataset, calculated over all n data points; and σ is the standard deviation of the dataset. This approach transforms the data into a standard normal distribution with a mean of 0 and a standard deviation of 1.

$$z = \frac{x_i - \mu}{\sigma} \quad (3)$$

$$\mu = \frac{1}{n} \sum_{i=1}^n x_i, \sigma = \sqrt{\frac{1}{n} \sum_{i=1}^n (x_i - \mu)^2} \quad (4)$$

4. Hyperparameter Tuning and Model Evaluation

4.1. Model Hyperparameter Selection

Overfitting, a phenomenon in which machine learning models achieve high performance on training data but fail to generalize to unseen test sets, often arises when models capture noise in the training data rather than underlying patterns [28]. To improve model generalization, cross-validation techniques are commonly employed. In the context of infectious substance risk assessment, the complexity of the evaluation process and the multitude of influencing factors make outcomes more sensitive to the randomness of data partitioning than simple training-validation splits. To address this issue, this study adopts a 10-fold cross-validation approach to optimize model training and evaluation. The specific workflow is illustrated in Figure 6.

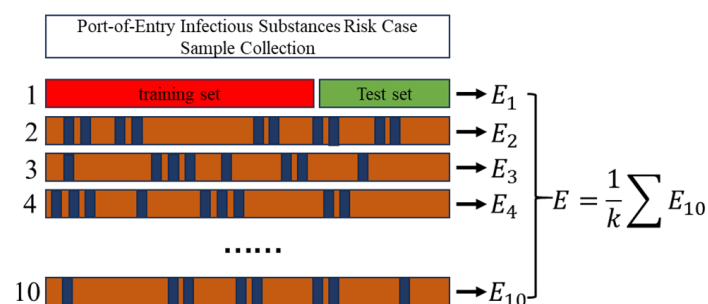


Figure 6. 10-fold Cross-Validation for Seeking Optimal Model Parameters.

The primary advantage of 10-fold cross-validation lies in its efficient use of limited data. By performing multiple iterations and averaging the results, it effectively reduces the sensitivity of evaluation outcomes to individual data splits. This robust evaluation mechanism provides a solid foundation for systematically comparing the performance of different model parameter configurations, thereby enabling precise selection of the optimal hyperparameter combination. The optimal hyperparameters obtained through 10-fold cross-validation are presented in Table 3.

Table 3. Optimal Parameter Table for Machine Learning Models.

Risk Assessment Model	Hyperparameters	Hyperparameter Value Range	Optimal Parameters
Decision Tree Model	min_samples_leaf	Default is 1	3
	max_depth	{3, 5, 8, 10, 15, 25, 30}	15
Support Vector Machine Model	Regularization parameter (C)	{0.1, 1, 10, 100, 1000, 10,000}	100
	gamma	{0.0001, 0.001, 0.01, 0.1, 1, 10}	0.1
	n_estimators	{10, 30, 50, 70, 90, 100}	500
Random Forest Model	n_estimators	{10, 30, 50, 70, 90, 100}	500
	min_samples_leaf	Default to 1	3
	max_depth	{3, 5, 8, 10, 15, 25, 30}	15
SVM-RF Model	Regularization parameter (C)	{0.1, 1, 10, 100, 1000, 10,000}	100
	gamma	{0.0001, 0.001, 0.01, 0.1, 1, 10}	0.1
	n_estimators	{10, 30, 50, 70, 90, 100}	500
	max_depth	{3, 5, 8, 10, 15, 25, 30}	15

4.2. Model Classification Performance Metrics

In machine learning, the confusion matrix is a fundamental tool for evaluating the performance of classification models, particularly in supervised learning. It systematically presents the correspondence between model predictions and the true classes of samples in a matrix format, serving as a key visualization method for assessing classification performance. In a confusion matrix, the rows correspond to the true classes of the samples, with the sum of each row representing the total number of samples in that true class. The columns correspond to the predicted classes generated by the model, with the sum of each column indicating the total number of samples predicted for that class.

The core components of the confusion matrix consist of four fundamental parameters: True Positives (TP), True Negatives (TN), False Positives (FP), and False Negatives (FN). Based on these parameters, several key evaluation metrics can be derived from the confusion matrix, including the following:

- (1) Precision

$$Precision = \frac{TP}{TP + FP}, \quad (5)$$

- (2) Recall

$$Recall = \frac{TP}{TP + FN}, \quad (6)$$

- (3) Specificity

$$Specificity = \frac{TN}{TN + FP}, \quad (7)$$

- (4) F1-score

$$F1\ Score = \frac{2(Precision \cdot Recall)}{Precision + Recall}. \quad (8)$$

In general, precision, specificity, and recall range from 0% to 100%, with higher values indicating better predictive performance of the model. The F1 score ranges from 0 to 1 and is also converted to percentages in this study for comparison purposes; similarly, higher values reflect stronger predictive capability, whereas lower values indicate reduced model effectiveness.

5. Comparative Analysis of Machine Learning Results

In machine learning, model complexity alone does not always determine classification performance; the feature attributes of the dataset also substantially influence model outcomes. To comprehensively assess the applicability of different models for predicting infectious substance risks at ports, this study constructs each model using an 8:2 training-to-test split based on identical dataset partitioning criteria. Four classification models are compared: decision tree, support vector machine, random forest, and SVM-RF. All models were trained and tested under a unified data preprocessing workflow to ensure comparability of evaluation results. Notably, all classification models in this study were developed and implemented using the Java programming language, providing scalability and portability aligned with practical engineering requirements for subsequent system integration and deployment.

Figure 7 presents the evaluation results of the decision tree model. The test set comprises 100 cases, of which 90 were correctly predicted, resulting in an overall prediction accuracy of 90%. Among the true categories, there are 32 samples at Level 1, 28 at Level 2, 23 at Level 3, and 17 at Level 4. In the predicted categories, 28 samples are classified as Level 1, 24 as Level 2, 21 as Level 3, and 17 as Level 4.

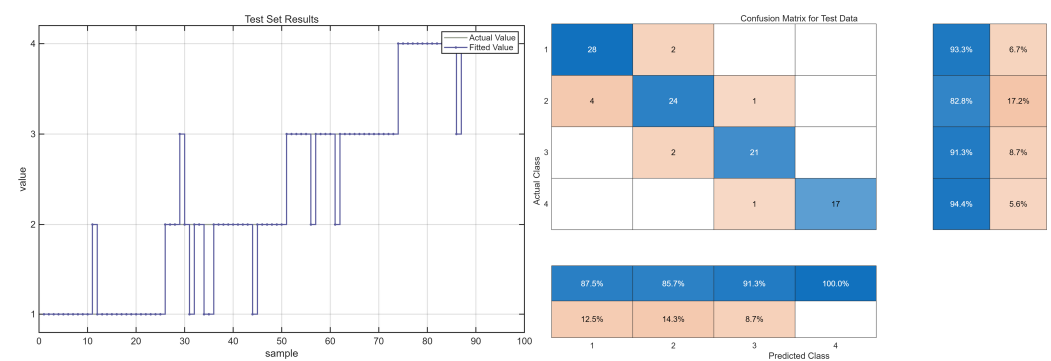


Figure 7. Decision Tree Model Results.

Figure 8 presents the evaluation results of the support vector machine (SVM) model. The test set comprises 100 cases, of which 90 were correctly predicted, resulting in an overall prediction accuracy of 90%. Among the true categories, there are 32 samples at Level 1, 28 at Level 2, 23 at Level 3, and 17 at Level 4. In the predicted categories, 28 samples are classified as Level 1, 26 as Level 2, 21 as Level 3, and 17 as Level 4.

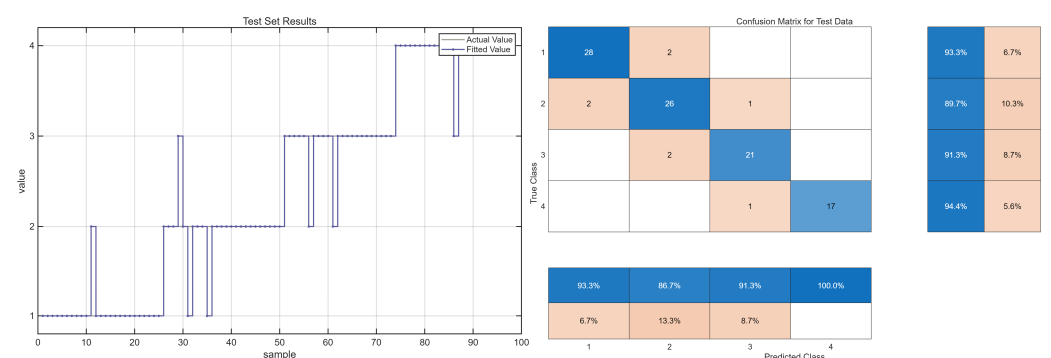


Figure 8. Support Vector Machine Model Results.

Figure 9 presents the evaluation results of the random forest (RF) model. The test set comprises 100 cases, of which 90 were correctly predicted, resulting in an overall prediction accuracy of 90%. Among the true categories, there are 32 samples at Level 1, 28 at Level

2, 23 at Level 3, and 17 at Level 4. In the predicted categories, 28 samples are classified as Level 1, 25 as Level 2, 21 as Level 3, and 17 as Level 4.

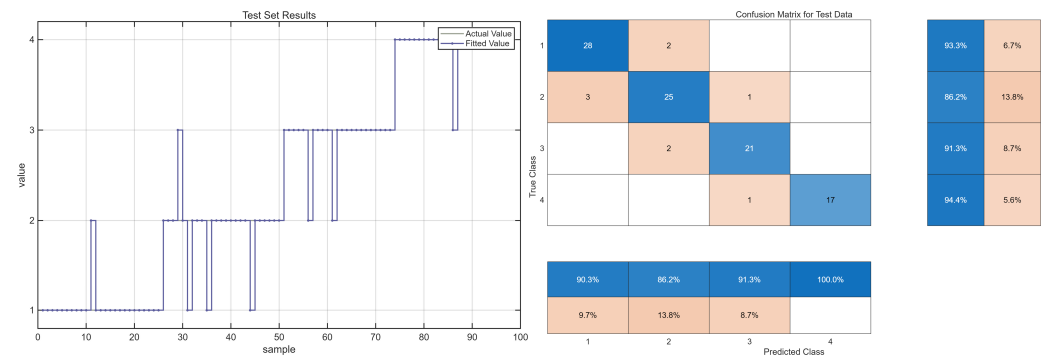


Figure 9. Random Forest Model Results.

Figure 10 presents the evaluation results of the SVM-RF model. The test set comprises 100 cases, of which 90 were correctly predicted, resulting in an overall prediction accuracy of 90%. Among the true categories, there are 32 samples at Level 1, 28 at Level 2, 23 at Level 3, and 17 at Level 4. In the predicted categories, 27 samples are classified as Level 1, 27 as Level 2, 21 as Level 3, and 17 as Level 4. The detailed number of correctly predicted samples in the confusion matrix is provided in Table 4.

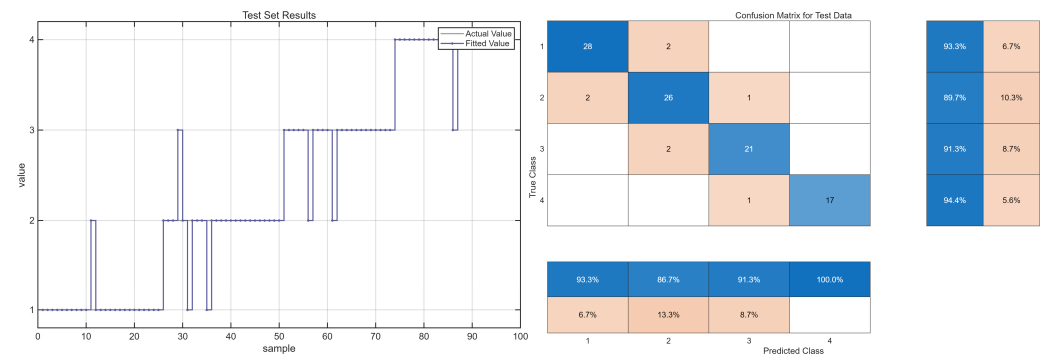


Figure 10. SVM-RF Model Results.

Table 4. Confusion Matrix Prediction Count Statistics.

Indicator Risk Level		Confusion Matrix Metrics						
		TP	FN	FP	TN	Precision	Recall	Specificity
Decision Tree	Grade I	28	2	4	66	0.9113	0.9046	0.9651
	Grade II	24	5	4	67			
	Grade III	21	2	2	75			
	Grade IV	17	1	0	82			
Support Vector Machine	Grade I	28	2	2	68	0.9283	0.9218	0.9723
	Grade II	26	3	4	67			
	Grade III	21	2	2	75			
	Grade IV	17	1	0	82			
Random Forest	Grade I	28	2	1	69	0.9196	0.9132	0.9687
	Grade II	25	4	5	66			
	Grade III	21	2	2	75			
	Grade IV	17	1	0	82			

Table 4. Cont.

Indicator Risk Level		Confusion Matrix Metrics							
		TP	FN	FP	TN	Precision	Recall	Specificity	F1 Score
SVM-RF	Grade I	27	3	1	69	0.9302	0.9221	0.9724	0.9262
	Grade II	27	2	5	66				
	Grade III	21	2	2	74				
	Grade IV	17	1	0	82				

The accuracy rates of the four models are 0.91, 0.92, 0.91, and 0.93, respectively. Based solely on prediction accuracy, no significant difference is observed between the support vector machine (SVM) and the optimized SVM-RF model. Therefore, further comparison using additional evaluation metrics is conducted. Figure 11 presents the specific values of these performance metrics, clearly demonstrating that the improved SVM-RF model outperforms all other models across all indicators. To further validate the classification performance, Receiver Operating Characteristic (ROC) analysis was conducted. The ROC curves presented in Figure 12 provide a clear visual comparison of the classification capability of the four models. Each ROC curve represents the trade-off between the true positive rate and the false positive rate under varying thresholds, with the area under the curve (AUC) serving as a quantitative indicator of model performance.

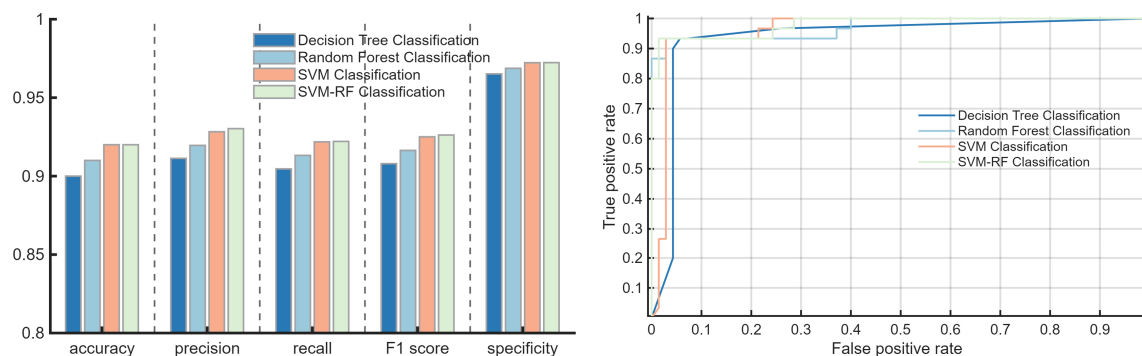


Figure 11. Comparison of Model Performance Metrics.

As shown in Figure 12, the AUC values for DT, RF, SVM, and SVM-RF are 0.9041, 0.9592, 0.9670, and 0.9828, respectively. The improved SVM-RF model exhibits the largest AUC, indicating its superior ability to discriminate between high-risk and low-risk samples. Its ROC curve lies closest to the upper-left corner, signifying an optimal balance between sensitivity and specificity. These results provide strong visual and quantitative evidence of the enhanced predictive accuracy of the improved model.

Overall, the improved SVM-RF algorithm demonstrates greater stability, precision, and adaptability compared with single models. By effectively leveraging the nonlinear feature extraction of SVM and the anti-overfitting capability of RF, the model offers a reliable framework for data-driven risk identification and classification of infectious substances at ports of entry. This contributes to improving the efficiency and scientific basis of port biosecurity risk assessment and management.

To further ensure the reliability of the comparative results, a paired *t*-test was performed on the accuracy and AUC values obtained from ten-fold cross-validation. The statistical results showed that the performance improvement of the SVM-RF model over the DT, RF, and SVM models was significant at the 0.05 level ($p < 0.05$). This finding confirms that the observed enhancement in predictive accuracy is statistically meaningful rather than a result of random variation, thereby reinforcing the robustness of the proposed model.

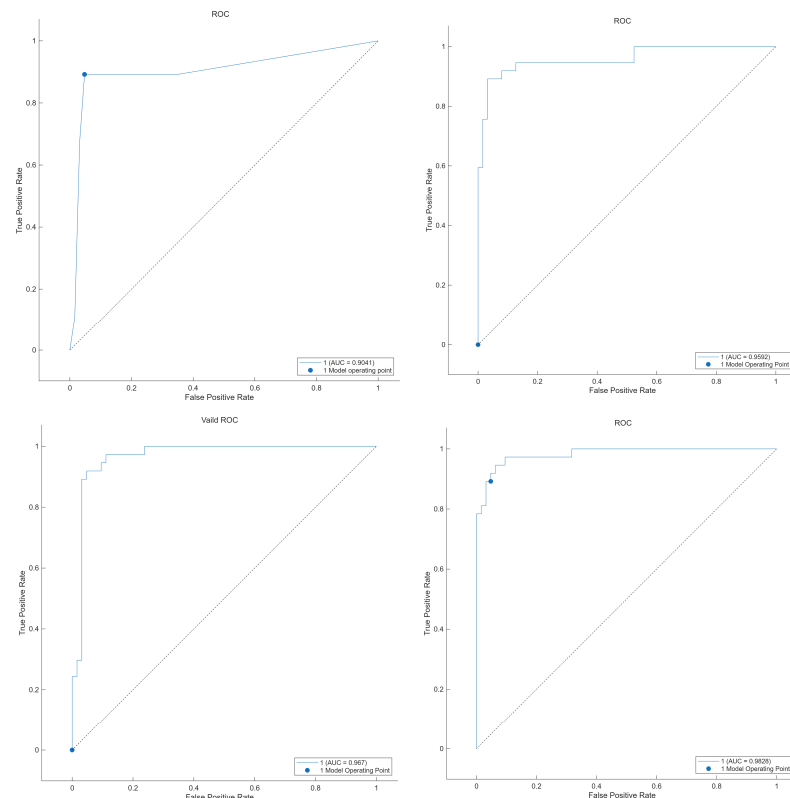


Figure 12. ROC Curve Comparison Chart.

6. Discussion and Conclusions

This study employs machine learning algorithms to construct a risk assessment model for infectious substances at ports of entry, aiming to enhance the efficiency of risk identification and evaluation for such materials through a data-driven approach. By comparing multiple machine learning methods—including decision trees, support vector machines, random forests, and the improved SVM-RF algorithm—the enhanced SVM-RF model demonstrated superior accuracy and generalization capabilities compared to individual algorithms. It combines the high-dimensional classification ability of support vector machines with the anti-overfitting advantages of random forests, making the model more robust in practical applications and capable of addressing complex infectious substance risk assessment challenges. Understanding how these factors influence risk levels aids in further optimizing risk assessment models and strengthening control measures.

Although the improvement in overall accuracy of the SVM-RF model over the baseline algorithms appears modest (from 0.91 to 0.93), a paired *t*-test on ten-fold cross-validation results confirmed that this gain is statistically significant ($p < 0.05$). Importantly, this 2% increase in accuracy corresponds to an approximately 18% reduction in the misclassification rate of high-risk infectious substance samples, which is practically meaningful in the context of customs risk management. Even small improvements in predictive precision can substantially enhance early warning systems, inspection prioritization, and the efficient allocation of limited safety resources. From a sustainability perspective, optimizing resource use while maintaining or improving safety aligns with broader goals of sustainable public health management and risk mitigation.

Nevertheless, this study relies on historical case data from Beijing Customs, and the dataset remains relatively small and geographically constrained. These factors may limit the generalizability of the proposed model to other ports or regions with distinct operational patterns, data structures, or inspection protocols. Furthermore, rare or emerging infectious substances may be underrepresented in the current dataset, potentially affecting the predic-

tive robustness for such cases. Future research should expand the dataset to include samples from multiple ports and diverse temporal contexts, thereby enhancing the external validity, adaptability, and sustainability of the proposed risk assessment framework.

Beyond generalizability, another important aspect concerns the interpretability of the proposed model in practical applications. While feature importance is derived from the Random Forest component, the interpretability of the final SVM-RF framework also deserves attention. Although the SVM-RF hybrid model primarily focuses on improving predictive performance, its decisions can still be explained to end-users through several means. Specifically, the Random Forest sub-model provides feature importance rankings that reveal which variables contribute most to risk classification, offering an interpretable basis for understanding model behavior. In addition, representative case analyses can be used to visualize decision boundaries between risk levels, enabling customs officers to understand why specific samples are classified as high or low risk. In future research, integrating explainable AI (XAI) tools such as SHAP or LIME could further enhance interpretability, helping domain experts trace model outputs back to specific features and strengthening trust in model-driven risk assessments.

In future research, the integration of singular pooling [29] could further enhance the feature representation capability of the SVM-RF framework. Singular pooling, as a spectral-domain pooling paradigm based on singular value decomposition (SVD), enables efficient compression of redundant features while preserving key discriminative information. Incorporating this mechanism into the SVM-RF fusion process may facilitate more compact and information-rich feature representations before classification, effectively mitigating the challenges posed by high-dimensional and multi-source risk data. Therefore, future work will focus on designing an extended “SVM-RF + Singular Pooling” framework or applying singular pooling directly to the infectious substance risk dataset to explore its potential in improving model robustness, generalization, and computational efficiency.

Author Contributions: Conceptualization, J.L.; methodology, J.L.; validation, C.L. and Y.B.; formal analysis, J.L. and C.L.; investigation, J.L., Y.B., F.W. and J.T.; resources, J.L.; data curation, F.W. and J.L.; writing—original draft preparation, J.L.; writing—review and editing, J.L. and C.L.; visualization, J.L. and Y.B.; supervision, J.L. and Y.B.; project administration, J.L. funding acquisition, Y.B. and C.L. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported by the National Key Research and Development Program of China (2023YFC2605801).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The data from this study involve public safety and privacy concerns. Therefore, if data are required, they may be requested from the corresponding author and obtained upon approval.

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

References

1. Huo, C.; Yang, Z.; Li, X.; Huo, F. Research on Policy Automatic Evaluation Method Based on PMC Index Model: A Case Study of China’s Biosafety Policy. *Inf. Stud. Theory Appl.* **2025**. Available online: https://en.cnki.com.cn/Article_en/CJFDTOTAL-QBLL20250705001.htm (accessed on 3 August 2025).
2. Yang, Z.; Liang, J.; Guo, L.; Dong, X. Risk Assessment of Lightning Strikes on Chemical Plant Storage Tanks Based on Hybrid Causal Logic. *China Saf. Sci. J.* **2024**, *34*, 174–182.
3. Song, B.; Yao, B.; Luo, Y. Risk Assessment of Natech Events in Urban Gas Supply Systems. *J. Saf. Environ.* **2023**, *23*, 3822–3829.
4. Zhang, Y.; Lin, G.; Li, Y. Risk Assessment of Poisoning and Asphyxiation Accidents in Desulfurization Processes Based on Fault Tree and Risk Matrix Method. *J. Xi’an Univ. Sci. Technol.* **2020**, *40*, 40–48.

5. Liu, Z.; Li, M. Risk Assessment of Hemodialysis Infection Based on FTA and Bayesian Network. *Ind. Eng. Manag.* **2017**, *22*, 106–113.
6. Wang, C.; Lü, S. Study on Mixed Probability Risk Assessment of Urban Gas Pipeline Leakage Disasters. *China Saf. Sci. J.* **2016**, *26*, 146–151.
7. Hulme, A.; McLean, S.; Dallat, C.; Walker, G.H.; Waterson, P.; Stanton, N.A.; Salmon, P.M. Systems Thinking-Based Risk Assessment Methods Applied to Sports Performance: A Comparison of STPA, EAST-BL, and Net-HARMS in the Context of Elite Women's Road Cycling. *Appl. Ergon.* **2021**, *91*, 103297. [[CrossRef](#)] [[PubMed](#)]
8. Wang, H.; Wang, Z.; Bai, Y.; Wan, W.; Wang, W.; Wang, K. Risk Assessment Method for Soil Secondary Salinization Based on Hydrological Frequency Analysis. *Ecol. Indic.* **2025**, *176*, 113688. [[CrossRef](#)]
9. Liu, R.; Qu, Y.; Liu, X.; Jiang, Y. Application of Ensemble Learning and Decision Tree in Prospective Risk Assessment of Type 2 Diabetes. *Chin. J. Prev. Control. Chronic Dis.* **2023**, *31*, 278–283.
10. Liu, J.; Xu, X.; Wu, F.; Li, F.; Wang, Y. Quantitative Risk Assessment of Water Inrush Accidents in Underground Engineering Based on Decision Tree. *Met. Mine* **2023**, *2*, 217–224.
11. Ding, G.; Liu, S.; Li, J. Evaluation of the Influence of Lifestyle Risk Factors on Cognitive Impairment Among Community-Dwelling Elderly Using Decision Tree and Random Forest. *Chin. J. Dis. Control. Prev.* **2020**, *24*, 485–488.
12. Xu, S.; Cai, H.; Zhao, L.; Xu, Y. A Bayesian Decision Tree Algorithm Model for Risk Prediction and Assessment in Agricultural Supply Chains. *J. Southwest Univ. (Nat. Sci. Ed.)* **2024**, *46*, 189–200.
13. Srivanit, M.; Kaewkhaw, S. A Machine Learning-Based Protocol to Support Visual Tree Assessment and Risk of Failure Classification on a University Campus. *Urban For. Urban Green.* **2024**, *99*, 128420. [[CrossRef](#)]
14. Recommender System Using Decision Tree and Neural Network for Disaster Risk Reduction Management Services. IEEE Conference Publication, IEEE Xplore. Available online: <https://ieeexplore.ieee.org/document/10847506> (accessed on 3 August 2025).
15. Jiang, H.; Tao, J.; Lü, Y.; Gao, X.; He, J. Risk Assessment and Application of Urban Water Supply Networks Based on Random Forest. *Water Wastewater Eng.* **2025**, *61*, 176–180.
16. Li, Y.; He, L.; Wan, J.; Pan, D.; Sun, J. Application of Improved K-SMOTE Random Forest Algorithm in Seismic Information Release Safety Risk Assessment. *J. Earthq. Eng.* **2025**, *47*, 168–177. [[CrossRef](#)]
17. Shan, Y.; Li, W.; Zhou, S.; Qu, G.; Xu, Q. Ecological Risk Assessment of Landslide Disasters in the Upper Reaches of the Yangtze River. *Acta Ecol. Sin.* **2024**, *44*, 11437–11449.
18. Yang, F.; Wang, H.; Fang, J.; He, J.; Huang, B.; Wang, L. Cable Risk Assessment and Repair Decision-Making Based on K-Means Clustering and Random Forest. *J. Nanjing Univ. Aeronaut. Astronaut.* **2024**, *56*, 892–899. [[CrossRef](#)]
19. Nikolaychuk, O.; Pestova, J.; Yurin, A. Wildfire Susceptibility Mapping in Baikal Natural Territory Using Random Forest. *Forests* **2024**, *15*, 170. [[CrossRef](#)]
20. Lilhore, U.K.; Manoharan, P.; Sandhu, J.K.; Simaiya, S.; Dalal, S.; Baqasah, A.M.; Alsafyani, M.; Alroobaea, R.; Keshta, I.; Raahemifar, K. Hybrid Model for Precise Hepatitis-C Classification Using Improved Random Forest and SVM Method. *Sci. Rep.* **2023**, *13*, 12473. [[CrossRef](#)] [[PubMed](#)]
21. Ghazal, T.M.; Anam, M.; Hasan, M.K.; Hussain, M.; Farooq, M.S.; Ali, H.M.A.; Ahmad, M.; Soomro, T.R. Hep-Pred: Hepatitis C Staging Prediction Using Fine Gaussian SVM. *Comput. Mater. Contin.* **2021**, *69*, 191–203. [[CrossRef](#)]
22. Wu, Y.; Mo, Q.; Xie, Y.; Zhang, J.; Jiang, S.; Guan, J.; Qu, C.; Wu, R.; Mo, C. A Retrospective Study Using Machine Learning to Develop Predictive Model to Identify Urinary Infection Stones In Vivo. *Urolithiasis* **2023**, *51*, 84. [[CrossRef](#)]
23. He, J.; Li, J.; Jiang, S.; Cheng, W.; Jiang, J.; Xu, Y.; Yang, J.; Zhou, X.; Chai, C.; Wu, C. Application of Machine Learning Algorithms in Predicting HIV Infection Among Men Who Have Sex with Men: Model Development and Validation. *Front. Public Health* **2022**, *10*, 967681. [[CrossRef](#)] [[PubMed](#)]
24. Sun, T.; Liu, J.; Yuan, H.; Li, X.; Yan, H. Construction of a Risk Prediction Model for Lung Infection After Chemotherapy in Lung Cancer Patients Based on the Machine Learning Algorithm. *Front. Oncol.* **2024**, *14*, 1403392. [[CrossRef](#)]
25. Xu, S.; Zhao, Q.; Yin, K.; Zhang, F.; Liu, D.; Yang, G. Combining Random Forest and Support Vector Machines for Object-Based Rural Land-Cover Classification Using High Spatial Resolution Imagery. *J. Appl. Remote Sens.* **2019**, *13*, 014521. [[CrossRef](#)]
26. Sivaraja, M.; Swaminathan, A.N.; Kuttimarks, M.S.; Rajprasad, J.; Sakthivel, M.; Rex, J. Prediction of the Anti-Carbonation Performance of Concrete Based on Random Forest-Least Squares Support Vector Machine Model. *J. Environ. Eng. Landsc. Manag.* **2025**, *33*, 226–241.
27. Gelmez, P.; Karakoc, T.E.; Ulucan, O. Autism Spectrum Disorder and Atypical Brain Connectivity: Novel Insights from Brain Connectivity-Associated Genes by Combining Random Forest and Support Vector Machine Algorithm. *OMICS J. Integr. Biol.* **2024**, *28*, 563–572. [[CrossRef](#)] [[PubMed](#)]

28. Zhao, Q.; Bi, K.; Qiu, T. Comparison and Integration of Machine Learning Models for Ethylene Cracking Processes. *J. Tsinghua Univ. (Sci. Technol.)* **2022**, *62*, 1450–1457.
29. Zhu, S.; Cai, J.; Xiong, R.; Zheng, L.; Ma, D. Singular Pooling: A Spectral Pooling Paradigm for Second-Trimester Prenatal Level II Ultrasound Standard Fetal Plane Identification. In *IEEE Transactions on Circuits and Systems for Video Technology*; IEEE: New York City, NY, USA, 2025; p. 1.

Disclaimer/Publisher’s Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.