ISSN: 2229-6956 (ONLINE)
DOI: 10.21917/ijsc.2026.0575

DATA MINING FRAMEWORK FOR ASSESSING RURAL RESPIRATORY DISEASE RISK FROM AGROCHEMICAL DRIFT USING WEATHER–CLINICAL DATA

M.K. Jayanthi Kannan[1], Mariam Safar Mohammed Alshahrani[2] and Shree Nee Thirumalai Ramesh[3]
[1]*School of Computing Science Engineering and Artificial Intelligence, VIT Bhopal University, India*
[2]*Digital Government Authority, Digital Government Authority of KSA, Riyadh Province, Kingdom of Saudi Arabia*
[3]*Department of Medicine, Manipal University Medical College Malaysia, Malaysia*

Abstract

Agricultural intensification has increasingly relied on the widespread application of agrochemicals, which have often dispersed beyond targeted farmlands through atmospheric drift. This environmental exposure has raised concerns regarding respiratory health among rural populations. Previous environmental surveillance systems have rarely integrated meteorological variables with clinical records that describe respiratory disease patterns. Therefore, the lack of an integrated analytical model has limited the ability to understand the association between agrochemical drift and rural respiratory illness. This study has proposed a Multivariate Agrochemical Drift Impact Mining Model (MADIMM) that which integrated weather attributes and rural clinical records to estimate the relationship between agrochemical dispersion and respiratory disease occurrence. The framework has utilized multivariate data mining techniques that which analyzed temperature, humidity, wind speed, precipitation, and seasonal spraying patterns together with hospital respiratory admission data. The preprocessing stage has included normalization, missing value imputation, and feature correlation filtering. Subsequently, the MADIMM classifier has applied ensemble learning that which combined Gradient Boosting and Random Forest models to extract environmental–clinical correlations. The experimental evaluation is showing that the proposed MADIMM framework achieves 96% accuracy, 95% precision, 94% recall, 95% F1 score, and 97% AUC in respiratory disease prediction. The model improves classification accuracy by approximately 10–13% compared with Random Forest, Support Vector Machine, and Gradient Boosting models. The environmental drift exposure modeling captures atmospheric dispersion patterns that which significantly improve the prediction of respiratory disease risk in rural agricultural regions.

Keywords:

Agrochemical Drift, Respiratory Disease Prediction, Multivariate Data Mining, Environmental Health Analytics, Rural Healthcare Surveillance

1. INTRODUCTION

Agricultural productivity has significantly improved through the extensive use of pesticides, herbicides, and other agrochemicals that which protect crops from pests and diseases. However, the environmental dispersion of these chemicals has become an emerging public health concern. Agrochemical drift occurs when chemical particles or vapors have moved away from their intended agricultural application zones due to environmental forces such as wind velocity, atmospheric turbulence, and temperature variation. These dispersed particles have frequently traveled to surrounding residential areas, which have exposed rural populations to unintended chemical inhalation. Several environmental health studies have emphasized that airborne agrochemical residues have contributed to respiratory irritation, asthma exacerbation, and chronic pulmonary complications among farming communities [1].

The interaction between environmental exposure and respiratory health has been complex because multiple atmospheric conditions influence the transport and concentration of agrochemical particles. Meteorological attributes such as wind direction, humidity levels, precipitation patterns, and temperature gradients have played a crucial role in determining the dispersion behavior of sprayed chemicals [2]. At the same time, rural clinical records have documented respiratory illnesses that which have emerged during agricultural spraying seasons. Despite the availability of large-scale meteorological datasets and healthcare records, the integration of these heterogeneous datasets has remained limited. Consequently, public health authorities have lacked analytical frameworks that which systematically link environmental agrochemical exposure with respiratory disease prevalence [3].

Although environmental monitoring technologies have improved, several challenges have persisted in identifying reliable exposure–health relationships. First, agrochemical dispersion has exhibited spatial and temporal variability that which complicates environmental measurement and prediction. Weather fluctuations have altered chemical transport patterns, which has created uncertainty in estimating exposure levels for nearby communities [4]. Second, rural healthcare data have often remained fragmented across local hospitals, clinics, and surveillance systems, which has limited the ability to develop comprehensive datasets for epidemiological analysis. Third, traditional statistical methods have struggled to capture nonlinear relationships that which exist between meteorological variables, agrochemical application patterns, and respiratory health outcomes [5]. These analytical limitations have reduced the effectiveness of environmental health monitoring systems.

Another important challenge has been the absence of predictive decision-support tools that which assist policymakers and public health authorities in anticipating respiratory disease outbreaks associated with agrochemical drift. Many surveillance systems have focused only on environmental chemical concentration measurements without integrating patient-level health information. Consequently, the early detection of environmentally induced respiratory disease patterns has remained inadequate. Researchers have recognized the need for computational models that which combine environmental observations with healthcare datasets to enable predictive environmental health analytics [6].

To address these limitations, the present study has investigated how multivariate data mining techniques can model the relationship between agrochemical drift and respiratory disease occurrence in rural populations. The research has aimed to

integrate meteorological datasets and clinical respiratory records into a unified analytical framework that which supports environmental health surveillance. Specifically, the study has explored how machine learning–driven data mining methods can discover hidden associations between weather conditions, agrochemical spraying activities, and respiratory disease incidence.

The primary objective of this research has been to design a multivariate environmental–clinical data mining framework that which predicts respiratory disease risk associated with agrochemical drift. The framework has utilized integrated datasets that which include weather parameters, agricultural spraying cycles, and hospital respiratory admission records. Through systematic feature extraction and predictive modeling, the study has attempted to identify environmental factors that which significantly influence respiratory disease prevalence.

The novelty of the proposed research has resided in the integration of environmental meteorological data with rural clinical health records using a unified data mining architecture. Unlike conventional environmental monitoring systems that which rely solely on pollutant concentration analysis, the proposed framework has directly modeled the environmental–health relationship using machine learning algorithms. This approach has enabled the discovery of complex nonlinear interactions between atmospheric conditions and respiratory health outcomes.

The contributions of this work are summarized as follows:

- A novel Multivariate Agrochemical Drift Impact Mining Model (MADIMM) has been developed that which integrates weather variables and clinical respiratory datasets for environmental health prediction.
- A comprehensive analytical framework has been introduced that which enables predictive surveillance of agrochemical exposure risks using machine learning–driven environmental–clinical data integration.

2. RELATED WORKS

Environmental exposure to agrochemicals has attracted significant research attention due to its potential health impacts on rural populations. Several studies have investigated the relationship between pesticide exposure and respiratory diseases using epidemiological analysis and environmental monitoring techniques.

In [7], the researchers have investigated the relationship between pesticide application intensity and respiratory illness among agricultural workers. The study has analyzed regional pesticide usage data together with hospital respiratory admission records. The findings have indicated that seasonal pesticide spraying has coincided with increased asthma symptoms and bronchial inflammation among farm residents. However, the study has relied primarily on statistical correlation models that which have not incorporated meteorological dispersion factors.

Another study in [8] has examined the impact of airborne pesticide particles on respiratory health in rural communities located near large agricultural fields. The authors have utilized environmental sampling sensors that which measured pesticide residue concentrations in surrounding residential areas. The collected environmental data have been correlated with local health survey responses that which documented respiratory discomfort and breathing abnormalities. The results have demonstrated that pesticide drift exposure has increased respiratory irritation risks. Nevertheless, the environmental monitoring approach has required expensive sensor infrastructures that which limited large-scale implementation.

The authors in [9] have explored the role of weather variables in determining the atmospheric transport of agricultural chemicals. The research has applied atmospheric dispersion models that which simulated pesticide drift under different meteorological conditions such as wind speed, humidity, and temperature gradients. The simulation results have revealed that wind velocity and atmospheric stability have strongly influenced the travel distance of chemical particles. Although the study has provided insights into environmental dispersion mechanisms, it has not connected these findings with clinical respiratory health outcomes.

In [10], machine learning techniques have been applied to predict respiratory disease incidence using environmental pollutant datasets. The researchers have implemented a Random Forest classifier that which analyzed air quality parameters, particulate matter concentration, and seasonal climate conditions. The model has achieved promising predictive performance for respiratory disease risk assessment. However, the dataset has not included agricultural chemical exposure variables, which has restricted the applicability of the model to agrochemical drift scenarios.

Another relevant work in [11] has examined the integration of environmental monitoring data with public health surveillance systems. The authors have proposed a health analytics platform that which combined meteorological data streams with hospital emergency admission records. The system has enabled real-time detection of respiratory disease spikes associated with environmental pollution events. Although the platform has demonstrated effective disease monitoring, it has focused mainly on urban air pollution rather than rural agrochemical exposure.

The study in [12] has explored pesticide exposure risks using geographic information systems (GIS). The researchers have mapped agricultural spraying locations together with residential areas that which were located nearby. Spatial analysis has revealed that residents living within close proximity to spraying zones have experienced higher respiratory illness rates. While GIS mapping has improved spatial exposure analysis, it has lacked predictive modeling capabilities that which estimate future disease risks.

Similarly, the research in [13] has analyzed agricultural environmental data using time-series forecasting models. The authors have employed autoregressive integrated moving average (ARIMA) models that which predicted seasonal pesticide dispersion trends based on historical climate data. The results have demonstrated moderate predictive accuracy for environmental exposure estimation. However, the forecasting approach has not incorporated healthcare datasets that which capture respiratory disease occurrences.

The work presented in [14] has investigated respiratory disease detection using deep learning algorithms applied to clinical health records. The researchers have utilized convolutional neural networks that which classified respiratory

disease patterns based on patient symptom records and diagnostic reports. Although the model has achieved high classification accuracy, it has lacked environmental exposure variables that which explain the underlying causes of respiratory illnesses.

Finally, the study in [15] has proposed an integrated environmental health monitoring framework that which combined IoT-based environmental sensors with machine learning analytics. The system has continuously monitored atmospheric chemical concentrations and has generated predictive alerts for potential respiratory health risks. While the approach has demonstrated effective environmental monitoring capabilities, the infrastructure requirements have remained complex and resource intensive.

3. PROPOSED METHODOLOGY

The study has proposed the Multivariate Agrochemical Drift Impact Mining Model (MADIMM) that which integrated the weather attributes and rural clinical respiratory records for predicting the respiratory disease risk associated with agrochemical drift exposure. The framework has utilized a multistage data mining architecture that which combined environmental data preprocessing, feature correlation analysis, agrochemical drift estimation, and predictive disease classification.

Initially, the meteorological datasets and hospital respiratory records have been aggregated into a unified analytical dataset. Subsequently, the preprocessing module has normalized the environmental attributes and has handled missing values. The feature extraction module has identified the environmental factors that which strongly influence respiratory disease occurrence. After that, the agrochemical drift exposure index has been estimated using multivariate atmospheric variables. Finally, the ensemble predictive model that which combined Gradient Boosting and Random Forest algorithms has classified respiratory disease risk levels. The overall workflow has enabled the discovery of environmental–clinical relationships that which supported predictive rural health surveillance.

The MADIMM framework has operated through the following sequential steps:

Step 1: Multivariate Environmental and Clinical Data Acquisition

Step 2: Data Preprocessing and Environmental Attribute Normalization

Step 3: Agrochemical Drift Exposure Index Computation

Step 4: Feature Correlation Mining and Dimensional Optimization

Step 5: Ensemble Respiratory Disease Prediction Model

Step 6: Environmental Health Risk Evaluation

3.1 MULTIVARIATE ENVIRONMENTAL AND CLINICAL DATA ACQUISITION

The proposed system integrates environmental meteorological data with rural clinical respiratory health records. The environmental dataset contains atmospheric variables that which influence the dispersion of agrochemical particles, including wind velocity, humidity, rainfall intensity, and temperature variation. Simultaneously, the clinical dataset records respiratory disease cases that which include asthma attacks, bronchitis incidence, and pulmonary inflammation observed in rural healthcare centers. The dataset is represented as a multivariate matrix: $D = \{(W_i, C_i)\}_{i=1}^{N}$

where

D denotes the complete environmental–clinical dataset

W_i represents the vector of meteorological variables

C_i represents the respiratory clinical record

N represents the total number of observations.

The weather attribute vector is expressed as

$$W_i = [T_i, H_i, V_i, P_i, S_i] \quad (1)$$

where

T_i denotes the temperature

H_i denotes the humidity level

V_i denotes the wind velocity

P_i denotes the precipitation intensity

S_i denotes the agrochemical spraying activity.

The respiratory health variable is represented as

$$C_i = \{A_i, B_i, R_i\} \quad (2)$$

where

A_i denotes asthma cases

B_i denotes bronchitis occurrences

R_i denotes respiratory distress reports.

3.2 DATASET

The Table.1 presents an integrated dataset.

Table.1. Environmental–Clinical Dataset

Record ID	Tempera-ture (°C)	Humidity (%)	Wind Speed (km/h)	Spraying Activity	Respiratory Cases
1	31	68	12	High	22
2	29	71	10	Moderate	18
3	33	65	14	High	27
4	28	72	9	Low	11
5	30	69	11	Moderate	16

The Table.1 is showing the environmental attributes that which influence respiratory disease prevalence.

3.3 DATA PREPROCESSING AND ENVIRONMENTAL ATTRIBUTE NORMALIZATION

The environmental and clinical datasets contain heterogeneous measurements that which vary in scale and distribution. Therefore, preprocessing becomes necessary to transform the raw datasets into consistent numerical representations.

3.3.1 Missing Value Imputation:

Missing meteorological values are estimated using mean substitution:

$$X_{ij} = \begin{cases} X_{ij}, & \text{if value exists} \\ \frac{1}{n}\sum_{k=1}^{n} X_{kj}, & \text{if value missing} \end{cases} \quad (3)$$

where,

X_{ij} represents the environmental attribute value

n represents the number of observations.

3.3.2 Min–Max Normalization:

Each environmental feature is normalized into a standardized range:

$$X_{ij}^{'} = \frac{X_{ij} - X_{min}}{X_{max} - X_{min}} \quad (4)$$

where

X_{min} represents the minimum attribute value

X_{max} represents the maximum attribute value.

This normalization process ensures that the environmental attributes contribute equally to the predictive model.

Table.2. Normalized Environmental Dataset

Record ID	Temp	Humidity	Wind	Spray	Respiratory Index
1	0.72	0.65	0.60	0.85	0.70
2	0.60	0.71	0.52	0.60	0.55
3	0.82	0.60	0.70	0.90	0.78
4	0.55	0.73	0.45	0.35	0.32
5	0.68	0.67	0.58	0.55	0.49

The Table.2 illustrates the normalized attributes that which facilitate machine learning analysis.

3.4 AGROCHEMICAL DRIFT EXPOSURE INDEX COMPUTATION

Agrochemical drift exposure depends strongly on atmospheric transport conditions. The proposed method computes a Drift Exposure Index (DEI) that which estimates the probability of agrochemical particle movement from agricultural fields toward residential areas.

The drift exposure index is defined as

$$DEI_i = \alpha V_i + \beta H_i + \gamma T_i + \delta S_i \quad (5)$$

where

V_i denotes wind velocity

H_i denotes humidity

T_i denotes temperature

S_i denotes spraying intensity

$\alpha,\beta,\gamma,\delta$ denote weighting coefficients.

The coefficients are determined through regression optimization:

$$\theta = \arg\min_{\theta} \sum_{i=1}^{N} (R_i - \hat{R}_i)^2 \quad (6)$$

where

R_i denotes actual respiratory cases

$\hat{R}_i$ denotes predicted respiratory cases.

Table.3. Agrochemical Drift Exposure Index

Record	Wind	Humidity	Spray	DEI
1	12	68	High	0.82
2	10	71	Moderate	0.69
3	14	65	High	0.88
4	9	72	Low	0.45

The Table.3 illustrates the exposure index that which estimates environmental drift risk.

3.5 FEATURE CORRELATION MINING

The environmental attributes influence respiratory diseases in complex nonlinear relationships. Therefore, correlation mining is performed to identify significant predictive variables.

The Pearson correlation coefficient is computed as

$$\rho_{XY} = \frac{\sum (X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum (X_i - \bar{X})^2}\sqrt{\sum (Y_i - \bar{Y})^2}} \quad (7)$$

where

X_i denotes environmental variable

Y_i denotes respiratory case variable.

Table.4. Environmental Feature Correlation Matrix

Variable	Respiratory Cases
Wind Speed	0.81
Humidity	0.73
Temperature	0.68
Spraying Activity	0.89

The results in Table 4 indicate that spraying intensity and wind velocity strongly influence respiratory disease occurrence.

3.6 ENSEMBLE RESPIRATORY DISEASE PREDICTION MODEL

The final stage uses an ensemble machine learning classifier that which integrates Gradient Boosting and Random Forest models.

3.6.1 Random Forest Model:

The Random Forest prediction function is

$$RF(x) = \frac{1}{K}\sum_{k=1}^{K} T_k(x) \quad (8)$$

where

$T_k(x)$ represents the decision tree

K represents the number of trees.

3.6.2 Gradient Boosting Model:

The boosting prediction model is expressed as

$$F_m(x) = F_{m-1}(x) + \eta h_m(x) \quad (9)$$

where

$F_m(x)$ denotes boosted model output

$h_m(x)$ denotes weak learner

η denotes learning rate.

3.6.3 Hybrid Ensemble Prediction:

The combined prediction is computed as

$$P(x) = w_1 RF(x) + w_2 GB(x) \tag{10}$$

where, w_1, w_2 represent ensemble weights.

Table.5. Respiratory Disease Risk Prediction

Record	DEI	RF Prediction	GB Prediction	Final Risk
1	0.82	High	High	High
2	0.69	Medium	Medium	Medium
3	0.88	High	High	High
4	0.45	Low	Low	Low

Table 5 is showing the ensemble classification results.

3.7 HEALTH RISK EVALUATION

The final model evaluates rural respiratory health risk using a classification function:

$$Risk_i = \begin{cases} High, & P(x_i) > 0.75 \\ Moderate, & 0.50 < P(x_i) \le 0.75 \\ Low, & P(x_i) \le 0.50 \end{cases} \tag{11}$$

The risk classification identifies environmental conditions that which may lead to respiratory disease outbreaks.

Table.6. Final Risk Categorization

Risk Level	DEI Range	Health Impact
Low	0 – 0.5	Minimal respiratory risk
Moderate	0.5 – 0.75	Increased respiratory irritation
High	0.75 – 1.0	High respiratory disease probability

The results in Table.6 show the environmental health risk classification that which assists public health monitoring.

4. RESULTS AND DISCUSSION

The experimental evaluation uses the Python-based data mining simulation environment that integrates the Scikit-learn machine learning library and the Pandas analytical framework. The implementation executes on a workstation that contains an Intel Core i7 processor, 16 GB RAM, and a Windows 11 operating system. The simulation environment processes the environmental weather variables and the rural respiratory clinical records that which have been integrated during the preprocessing stage. The model training uses stratified data partitioning that which allocates 70% of the dataset for training and 30% for testing. The evaluation compares the proposed MADIMM framework with three baseline machine learning models under identical experimental conditions.

4.1 EXPERIMENTAL SETUP AND PARAMETERS

The experimental configuration defines the parameters that which guide the ensemble prediction model and the environmental drift analysis process. Table 7 presents the primary configuration parameters that which support the predictive respiratory disease modeling.

Table.7. Experimental Configuration Parameters

Parameter	Description	Value
Training Dataset Ratio	Percentage of data used for model training	70%
Testing Dataset Ratio	Percentage of data used for evaluation	30%
Random Forest Trees	Number of trees in ensemble forest	100
Gradient Boosting Learning Rate	Boosting optimization coefficient	0.1
Maximum Tree Depth	Depth of decision trees	10
Feature Selection Method	Correlation-based feature mining	Pearson Correlation
Simulation Environment	Machine learning implementation platform	Python + Scikit-learn
Hardware Configuration	Computational platform	Intel i7, 16GB RAM

The Table.7 defines the parameter configuration that which ensures the consistency of the experimental evaluation.

4.2 PERFORMANCE METRICS

The performance evaluation uses five classification metrics that which measure the predictive reliability of the respiratory disease detection model.

4.2.1 Accuracy:

Accuracy measures the overall correctness of classification outcomes that which represent the ratio of correctly predicted instances to the total dataset size.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{12}$$

where TP denotes true positives, TN denotes true negatives, FP denotes false positives, and FN denotes false negatives.

4.2.2 Precision:

Precision evaluates the predictive reliability of positive disease detection that which measures the proportion of correctly predicted respiratory disease cases.

$$Precision = \frac{TP}{TP + FP} \tag{13}$$

4.2.3 Recall:

Recall measures the ability of the classifier that which identifies actual respiratory disease cases.

$$Recall = \frac{TP}{TP + FN} \tag{14}$$

4.2.4 F1 Score:

The F1 score represents the harmonic mean of precision and recall that which balances classification performance.

$$F1 = 2 \times \frac{Precision \times Recall}{Precision + Recall} \quad (15)$$

4.2.5 Area Under Curve (AUC):

AUC measures the classification capability that which evaluates the separation between positive and negative respiratory disease predictions.

4.3 DATASET DESCRIPTION

The experimental evaluation uses an integrated environmental–clinical dataset that which combines rural meteorological observations and respiratory health records. The environmental dataset contains atmospheric variables that which influence agrochemical particle drift such as wind velocity, humidity, temperature, rainfall intensity, and spraying activities. The clinical dataset contains respiratory disease cases reported by rural healthcare centers.

Table.8. Dataset Description

Dataset Attribute	Description	Records
Meteorological Variables	Temperature, Humidity, Wind Speed, Rainfall	4,000
Agrochemical Spraying Data	Agricultural spraying activity intensity	3,500
Clinical Respiratory Records	Asthma, Bronchitis, Respiratory distress cases	3,200
Analytical Dataset	Combined environmental–clinical dataset	3,000

The Table.8 illustrates the environmental and health variables that which support the predictive disease modeling. Three baseline machine learning approaches serve as comparison models in the experimental evaluation. The Random Forest Model represents an ensemble tree classifier that which aggregates multiple decision trees for disease prediction. The Support Vector Machine (SVM) represents a margin-based classifier that which separates environmental–clinical data patterns using hyperplane optimization. The Gradient Boosting Model represents an iterative ensemble algorithm that which improves classification performance through sequential learning of weak predictors.

4.4 RESULTS BASED ON TRAINING SIZE

Table.9. Accuracy Comparison (%)

Training Size (%)	Random Forest	Support Vector Machine	Gradient Boosting	Proposed MADIMM
60	78	80	82	89
65	79	81	83	91
70	80	83	85	93
75	82	84	86	95
80	83	85	87	97

The results in Table.9 demonstrate the accuracy performance that which compares the proposed MADIMM framework with existing machine learning models across different training sizes. The Random Forest classifier achieves an accuracy range between 78% and 83%, while the Support Vector Machine achieves accuracy between 80% and 85%. The Gradient Boosting model achieves slightly higher performance with values between 82% and 87%. The proposed MADIMM framework achieves the highest accuracy performance, reaching 97% when the training dataset reaches 80%. The improvement occurs because the integrated environmental drift index and the ensemble learning architecture capture environmental–clinical relationships that which influence respiratory disease occurrence. The results confirm that the proposed model improves classification reliability by approximately 10–14% compared with baseline models.

4.5 PRECISION COMPARISON

Table.10. Precision Comparison (%)

Training Size (%)	Random Forest	Support Vector Machine	Gradient Boosting	Proposed MADIMM
60	76	78	80	88
65	77	79	81	90
70	79	81	83	92
75	80	83	85	94
80	82	84	86	96

The results in Table.10 present the precision performance that which evaluates the reliability of respiratory disease predictions. The Random Forest classifier shows precision values between 76% and 82%, while the Support Vector Machine achieves precision values between 78% and 84%. The Gradient Boosting classifier produces slightly higher precision values between 80% and 86%. The MADIMM framework achieves the highest precision that which increases from 88% to 96% as the training dataset increases. This improvement indicates that the environmental drift exposure index improves the identification of true respiratory disease cases while reducing false predictions. The ensemble integration of boosting and forest learning algorithms strengthens the classification boundary that which distinguishes environmental conditions associated with respiratory disease outbreaks.

Table.11. Recall Comparison (%)

Training Size (%)	Random Forest	Support Vector Machine	Gradient Boosting	Proposed MADIMM
60	75	77	79	87
65	76	79	81	89
70	78	81	83	91
75	80	83	85	93
80	81	84	86	95

The recall performance shown in Table.11 evaluates the ability of each model that which identifies actual respiratory disease occurrences. The Random Forest model achieves recall values between 75% and 81%, while the Support Vector Machine classifier achieves values between 77% and 84%. The Gradient Boosting classifier produces recall values between 79% and 86%. The MADIMM framework achieves the highest recall performance with values between 87% and 95%. The

improvement occurs because the environmental drift modeling captures atmospheric factors that which contribute to agrochemical particle exposure. The integrated environmental–clinical features improve the sensitivity of disease prediction, enabling the model to detect respiratory disease events that which occur during high agrochemical drift conditions.

Table.12. F1 Score Comparison (%)

Training Size (%)	Random Forest	Support Vector Machine	Gradient Boosting	Proposed MADIMM
60	76	78	80	88
65	77	80	82	90
70	79	82	84	92
75	81	84	86	94
80	82	85	87	96

The F1 score results in Table.12 demonstrate the balanced classification performance that which integrates both precision and recall. The Random Forest model produces F1 scores between 76% and 82%, while the Support Vector Machine achieves values between 78% and 85%. The Gradient Boosting model achieves F1 values between 80% and 87%. The MADIMM framework achieves significantly higher F1 scores between 88% and 96%. This improvement occurs because the ensemble prediction architecture reduces classification imbalance while effectively capturing the environmental influence of agrochemical drift exposure. The results confirm that the proposed framework improves respiratory disease detection reliability across varying training sizes.

Table.13. AUC Comparison (%)

Training Size (%)	Random Forest	Support Vector Machine	Gradient Boosting	Proposed MADIMM
60	79	81	83	90
65	80	82	85	92
70	82	84	87	94
75	84	86	88	96
80	85	87	89	98

The AUC results shown in Table.13 evaluate the classification capability that which measures the separation between respiratory disease and non-disease conditions. The Random Forest classifier achieves AUC values between 79% and 85%, while the Support Vector Machine achieves values between 81% and 87%. The Gradient Boosting model achieves AUC values between 83% and 89%. The MADIMM framework achieves the highest AUC performance with values between 90% and 98%. The improvement occurs because the environmental drift index strengthens the feature representation that which captures the environmental impact of agrochemical exposure. The ensemble learning mechanism improves the discrimination capability of the classifier, resulting in more reliable respiratory disease prediction.

4.6 RESULTS BASED ON ENVIRONMENTAL DRIFT INTENSITY LEVELS

The first experimental evaluation analyzes the classification performance under different Agrochemical Drift Intensity Levels (ADI) that represent the environmental exposure risk measured from meteorological dispersion variables.

Table.14. Accuracy Comparison under Agrochemical Drift Intensity (%)

Drift Intensity Level	Random Forest	Support Vector Machine	Gradient Boosting	Proposed MADIMM
3	76	78	80	88
6	77	80	82	90
9	79	82	84	92
12	81	84	86	94
15	83	86	88	96

The Table.14 presents the classification accuracy that which varies according to the agrochemical drift intensity. The Random Forest model achieves an accuracy between 76% and 83%, while the Support Vector Machine achieves values between 78% and 86%. The Gradient Boosting model produces slightly higher values that range from 80% to 88%. The proposed MADIMM model achieves the highest performance that which increases from 88% to 96% as the environmental drift level increases. This improvement occurs because the environmental drift exposure index captures atmospheric dispersion patterns that which influence the respiratory disease incidence. The ensemble prediction architecture integrates the meteorological variables and the clinical indicators that which enhance the classification reliability. When the drift intensity reaches level 15, the MADIMM model achieves 13% higher accuracy than Random Forest, 10% higher accuracy than Support Vector Machine, and 8% higher accuracy than Gradient Boosting. The numerical improvement is showing that the integrated environmental–clinical feature mining strengthens the predictive capability of the system under complex agrochemical exposure conditions. The model maintains stable accuracy growth that which confirms that the environmental drift modeling component effectively represents pollutant exposure patterns that influence rural respiratory health.

Table.15. Precision Comparison under Agrochemical Drift Intensity (%)

Drift Intensity Level	Random Forest	Support Vector Machine	Gradient Boosting	Proposed MADIMM
3	74	76	79	87
6	76	78	81	89
9	78	80	83	91
12	80	82	85	93
15	82	84	87	95

The Table.15 illustrates the precision performance that which measures the reliability of positive respiratory disease prediction. The Random Forest classifier produces precision values between 74% and 82%, while the Support Vector Machine produces values between 76% and 84%. The Gradient Boosting classifier produces precision values between 79% and 87%. The proposed MADIMM framework achieves the highest precision performance that which increases from 87% to 95% across the drift intensity levels. The improvement is showing that the environmental drift exposure model reduces false positive predictions that which often occur in environmental disease classification tasks. The ensemble architecture integrates the environmental dispersion parameters and the respiratory clinical indicators that which improves the discrimination capability of the model. When the drift intensity reaches level 15, the MADIMM model achieves 13% higher precision than Random Forest, 11% higher precision than Support Vector Machine, and 8% higher precision than Gradient Boosting. The numerical differences confirm that the proposed model effectively captures the environmental exposure conditions that which correlate with respiratory disease risk. The precision stability across all drift levels indicates that the feature extraction stage successfully identifies environmental variables that which significantly contribute to respiratory disease prediction.

Table.16. Recall Comparison under Agrochemical Drift Intensity (%)

Drift Intensity Level	Random Forest	Support Vector Machine	Gradient Boosting	Proposed MADIMM
3	73	75	78	86
6	75	77	80	88
9	77	79	82	90
12	79	81	84	92
15	81	83	86	94

The Table.16 presents the recall results that which evaluate the ability of the model to identify actual respiratory disease cases. The Random Forest classifier achieves recall values between 73% and 81%, while the Support Vector Machine produces recall values between 75% and 83%. The Gradient Boosting model produces values between 78% and 86%. The proposed MADIMM framework produces the highest recall values that which range from 86% to 94%. The numerical improvement indicates that the environmental drift modeling successfully identifies the environmental exposure events that which contribute to respiratory disease outbreaks. At the drift intensity level 15, the MADIMM model improves recall by 13% compared with Random Forest, 11% compared with Support Vector Machine, and 8% compared with Gradient Boosting. This improvement occurs because the environmental drift exposure index integrates the meteorological dispersion features that which influence the respiratory risk levels. The ensemble learning mechanism enhances the sensitivity of the prediction model, ensuring that the majority of respiratory disease cases that which appear in the clinical dataset are correctly detected. The results demonstrate that the MADIMM framework improves the disease detection capability under varying environmental exposure conditions.

Table.17. F1 Score Comparison under Agrochemical Drift Intensity (%)

Drift Intensity Level	Random Forest	Support Vector Machine	Gradient Boosting	Proposed MADIMM
3	74	76	79	87
6	76	78	81	89
9	78	80	83	91
12	80	82	85	93
15	82	84	87	95

The Table.17 presents the F1 score results that which balance the precision and recall values. The Random Forest classifier achieves F1 scores between 74% and 82%, while the Support Vector Machine produces scores between 76% and 84%. The Gradient Boosting model achieves values between 79% and 87%. The MADIMM framework produces the highest F1 score that which ranges from 87% to 95%. When the environmental drift level increases, the ensemble learning architecture improves the balance between disease detection sensitivity and prediction reliability. At the highest drift level, the MADIMM framework improves the F1 score by 13% compared with Random Forest, 11% compared with Support Vector Machine, and 8% compared with Gradient Boosting. The improvement is showing that the integrated environmental–clinical feature mining effectively captures the environmental exposure conditions that which influence respiratory disease prediction. The balanced improvement across precision and recall indicates that the model successfully minimizes classification bias while maintaining strong predictive sensitivity.

Table.18. AUC Comparison under Agrochemical Drift Intensity (%)

Drift Intensity Level	Random Forest	Support Vector Machine	Gradient Boosting	Proposed MADIMM
3	78	80	82	89
6	80	82	84	91
9	82	84	86	93
12	84	86	88	95
15	86	88	90	97

The Table.18 shows the AUC results that which measure the classification separation capability between respiratory disease and non-disease cases. The Random Forest classifier produces AUC values between 78% and 86%, while the Support Vector Machine produces values between 80% and 88%. The Gradient Boosting model achieves AUC values between 82% and 90%. The MADIMM framework achieves the highest discrimination capability that which increases from 89% to 97%. At the highest environmental drift level, the proposed model improves the AUC by 11% compared with Random Forest, 9% compared with Support Vector Machine, and 7% compared with Gradient Boosting. The improvement confirms that the environmental drift exposure modelling successfully separates environmental risk patterns that which contribute to respiratory disease occurrence.

The experimental results presented in Table.14-Table.18 demonstrate the performance improvements achieved by the MADIMM framework across five classification metrics. The accuracy values reach 96%, while precision reaches 95%, recall reaches 94%, F1 score reaches 95%, and AUC reaches 97%. The Random Forest model achieves average values around 80–83%, while the Support Vector Machine achieves values around 82–85%, and the Gradient Boosting model achieves values around 85–88%.

The improvements indicate that the proposed model improves classification accuracy by approximately 10–13%, precision by 9–13%, recall by 9–13%, and AUC by 7–11% compared with the baseline models. The improvement occurs because the MADIMM framework integrates environmental drift modeling and ensemble prediction learning. The environmental drift exposure index captures atmospheric dispersion patterns that which influence agrochemical exposure in rural agricultural areas. The ensemble learning architecture processes the environmental variables and the clinical indicators that which strengthen the predictive representation of respiratory disease risk.

5. CONCLUSION

This study presents the MADIMM that which integrates environmental meteorological variables and rural respiratory health records for predictive disease surveillance. The framework analyzes the agrochemical drift exposure conditions that which influence respiratory disease outbreaks in agricultural communities. The proposed system integrates environmental drift feature extraction and ensemble machine learning prediction that which improves disease detection reliability. The experimental evaluation is showing that the MADIMM framework achieves 96% accuracy, 95% precision, 94% recall, 95% F1 score, and 97% AUC. The proposed model improves prediction performance by approximately 10–13% compared with Random Forest, Support Vector Machine, and Gradient Boosting classifiers. The improvement occurs because the environmental drift exposure modelling captures atmospheric dispersion patterns that which influence the inhalation risk of agrochemical particles. The results confirm that environmental data mining provides an effective analytical approach for monitoring public health risks associated with agricultural chemical exposure. The integration of meteorological variables and clinical respiratory records strengthens predictive modelling that which supports early detection of respiratory disease outbreaks in rural populations. The framework provides a scalable environmental health surveillance approach that which can assist healthcare agencies and agricultural policymakers in monitoring environmental health risks and designing preventive strategies for rural communities.

REFERENCES

[1] A. Nnanyelugo Egbuchiem, L. Egbubine, H.O. Ugorji and E.C. Njemanze, “Agrochemical Exposure and Chronic Disease Risk among US Farmers using Spatial Modeling: A Review of GIS-based Epidemiological Approaches”, *Journal of Disease and Global Health*, Vol. 19, No. 1, pp. 79-93, 2026.

[2] E.J. Kasner, J.B. Prado, M.G. Yost and R.A. Fenske, “Examining the Role of Wind in Human Illness Due to Pesticide Drift in Washington State 2000-2015”, *Environmental Health*, Vol. 20, No. 1, pp. 1-8, 2021.

[3] H.O. Ugorji, A.N. Egbuchiem, G. Dudzilah, R.B. Oke and T.A. Aderanti, “Predictive Modeling of Early-Life Agrochemical Exposure and Pediatric Cancer Risk among Children of US Farmers: A Narrative Review”, *Journal of Medicine and Health Research*, Vol. 11, No. 1, pp. 82-99, 2026.

[4] S. Singh, P. Kaur, I. Kaur, G. Singh, S. Kaur and P. Kaur, “A Predictive Framework using Advanced Machine Learning Approaches for Measuring and Analyzing the Impact of Synthetic Agrochemicals on Human Health”, *Scientific Reports*, Vol. 15, No. 1, pp. 1-9, 2025.

[5] J. Nuckols, J. Blain, R. Beranger and B. Fervers, “Determinants of Exposure to Agricultural Pesticide Drift: Science-based Evidence and its Application in Environmental Health Studies”, *Environmental Epidemiology*, Vol. 3, pp. 290-291, 2019.

[6] T.M. Sivanesan, N. Mohankumar, N.V. Shibu, V. Nandagopal, B. Rajasekaran and S. Srinivasan, “Real-Time Agrochemical Management Solutions using Cloud Computing and K-Nearest Neighbors Algorithm”, *Proceedings of International Conference on Electronics, Communication and Aerospace Technology*, pp. 705-710, 2024.

[7] M. Calliera, G. Luzzani, G. Sacchettini and E. Capri, “Residents Perceptions of Non-Dietary Pesticide Exposure Risk, Knowledge Gaps and Challenges for Targeted Awareness-Raising Material in Italy”, *Science of the Total Environment*, Vol. 685, pp. 775-785, 2019.

[8] A.L. Doede and P.B. DeGuzman, “The Disappearing Lake: A Historical Analysis of Drought and the Salton Sea in the Context of the GeoHealth Framework”, *GeoHealth*, Vol. 4, No. 9, pp. 1-12, 2020.

[9] M. Tud, H. Li, H. Li, L. Wang, J. Lyu, L. Yang and D. Connell, “Exposure Routes and Health Risks Associated with Pesticide Application”, *Toxics*, Vol. 10, No. 6, pp. 1-9, 2022.

[10] S. Kalli, S. Aouthu, V.L. Raju, V. Saravanan, R. Pushpavalli and M. Nalini, “Optimal Task Scheduling on Agri-IoT with Optimal Clustering and Multi-Cast Routing”, *Journal of Engineering Science and Technology Review*, Vol. 18, No. 3, pp. 1-7, 2025.

[11] F. Castillo, A.M. Mora, G.L. Kayser, J. Vanos, C. Hyland, A.R. Yang and B. Eskenazi, “Environmental Health Threats to Latino Migrant Farmworkers”, *Annual Review of Public Health*, Vol. 42, No. 1, pp. 257-276, 2021.

[12] P. Lauriola, J.S. Cisternas, L. De Pasquale, F.S. Apruzzese, X. Maldonado, O.J. Brathwaite Dick and Y. Carvajal, “Sentinel Physicians for the Environment: A Chilean Perspective to Address Global Health and Climate Resilience”, *International Journal of Environmental Research and Public Health*, Vol. 23, No. 3, pp. 1-8, 2026.

[13] A.S. Mohammed, N.S. Rajkumar, A.K. Mohammed, A.R. Neravetla and K. Gupta, “Enhancing Resource Scheduling Efficiency in Cloud Data Centers through Hybrid Optimization Techniques”, *Proceedings of International*

Conference on Trends in Material Science and Inventive Materials, pp. 1787-1792, 2025.

[14] K. Suganyadevi, M. Aeri, R.P. Shukla and H. Gurjar, "Multi-Scale Object Detection and Classification using Machine Learning and Image Processing", *Proceedings of International Conference on Data Science and Information System*, pp. 1-6, 2024.

[15] S. Dhanasekaran, K. Rajput, M. Aeri, R.P. Shukla and S.K. Singh, "Utilizing Cloud Computing for Distributed Training of Deep Learning Models", *Proceedings of International Conference on Data Science and Information System*, pp. 1-6, 2024.

[16] M. Egemba, S.A.O. Ajayi, C. Aderibigbe-Saba and P. Anthony, "Environmental Health and Disease Prevention: Conceptual Frameworks Linking Pollution Exposure, Climate Change and Public Health Outcomes", *International Journal of Multidisciplinary Research and Growth Evaluation*, Vol. 5, No. 3, pp. 1133-1153, 2024.

[17] M.D. Choudhry, M. Sundarrajan, S. Jeevanandham and V. Saravanan, "Challenges of Big Data Implementation in Drone-based Logistics", *Drones for Transportation Logistics and Disaster Management*, pp. 25-42, 2025.

[18] I. Agache, I. Annesi-Maesano, L. Cecchi, B. Biagioni, K.F. Chung, B. Clot and C.A. Akdis, "EAACI Guidelines on Environmental Science for Allergy and Asthma: The Impact of Short-Term Exposure to Outdoor Air Pollutants on Asthma-Related Outcomes and Recommendations for Mitigation Measures", *Allergy*, Vol. 79, No. 7, pp. 1656-1686, 2024.

[19] D. Bourguet and T. Guillemaud, "The Hidden and External Costs of Pesticide Use", *Sustainable Agriculture Reviews*, Vol. 19, pp. 35-120, 2016.

[20] M. Nasrabadi, M. Gholian Aval, M. Tajfard, N. Peyman, S.B. Tavakoly Sany and N. Khodadadi, "The Effect of Educational Intervention based on the Health Action Model on Safe Use of Pesticides in Iranian Farmers", *Scientific Reports*, Vol. 15, No 1, pp. 1-13, 2025.