IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 6, Issue 14, April 2026
Impact Factor: 8.2

Air Quality Level Prediction Using Hybrid Machine Learning Classification Model: A Comprehensive Analysis of Environmental Pollution Assessment

Moinuddin Mulani, Nikhil korke, Atharva Kumbhar, Sneha Kharat
Department of Computer Science and Engineering
JSPM University, Wagholi, Pune, India

Abstract: *Decline in air quality is a major environmental and health hazard at the international level with particulate matter and gases affecting health in varying geographic regions. In this research, a comprehensive hybrid approach of machine learning algorithms is proposed to predict air quality index (AQI) categories accurately by employing environmental as well as meteorological variables. The proposed air quality prediction system is based on a voting ensemble technique comprising three classifiers namely Support Vector Machines (SVM), Random Forest (RF) and Logistic Regression (LR) to benefit from their respective strengths. Methodology comprises of designing a data processing pipeline including data preprocessing such as filling in of missing values, feature extraction based on seven significant air quality metrics including AQI, CO, Ozone, NO2 and PM2.5 levels along with geographic coordinates and standard normalization. The proposed hybrid technique provides better classification performance than individual algorithms with more than 96% accuracy achieved on test dataset having precision, recall and F1 scores higher than 94%. The system comprises user authentication, graphical prediction UI and persistent model saving using joblib library for efficient deployment. Statistically, it has been found that the ensemble technique has captured the complex non-linear relationships between pollution variables and AQI classes*

Keywords: Air Quality Forecasting, Machine Learning, Ensemble Classification, Environmental Surveillance, Feature Engineering, Sup-port Vector Machines, Random Forest, Voting Classifier

I. INTRODUCTION

Air pollution is arguably one of the gravest environmental issues that face the twenty-first century, and according to the World Health Organization, it is estimated that 7 million premature deaths occur each year due to outdoor air pollution [12]. The buildup of particulate matter (PM), nitrogen dioxide (NO2), carbon monoxide (CO), and ozone (O3) in the troposphere has significant consequences on human respiration and the environment as a whole.

The conventional manual monitoring approach involves using scarce sensors located at specific locations; this implies inadequate spatial and temporal coverage when assessing pollution. Air Quality Index (AQI) is a standard measure that combines several pollutants and presents them in one categorization index form, making it easy for policymakers and government officials to communicate about pollution risks. Nevertheless, the classification of AQI categories based on pollutant levels is computationally difficult because it requires addressing non-linear dependencies between various factors.

DOI: 10.48175/568

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 6, Issue 14, April 2026
Impact Factor: 8.2

A. Motivation and Research Gap

The current methodologies based on a single model have some deficiencies in modeling pollution dynamics. Linear regression is one methodology that makes assumptions regarding Gaussian distribution errors which are not relevant for AQI categorization. The different types of machine learning classifiers have shown effectiveness, however there is no single classifier that can achieve optimal bias and variance performance in heterogeneous data environments.

B. Research Contributions

The research work highlights an advanced hybrid classification approach incorporating the use of SVM, Random Forests, and Logistic Regression using a soft-voting ensemble technique. Contributions of this paper include:

1) Systematic construction of features from several different pollutants and geographical features;
2) Building a production-level classifier having model per-sistence;
3) Evaluation showing 96% accuracy of the system;
4) GUI enabled for practical implementation;
5) Advanced statistical study on component interactions.

II. LITERATURE REVIEW

Kumar and Prabhu [1] evaluated 47 papers on air quality prediction using machine learning, noting the use of classification and regression methods, with neural networks and random forests showing superior prior performance (RMSE 8–12 µg/m3). The limitations were the absence of uniform data sources, evaluation criteria, and geographic generalization.

Goyal et al. [2] used LSTM networks for forecasting the AQI in a time series considering the temporal dependency of air quality indices. Although they achieved 91% accuracy, the method demanded huge computational costs and lacked interpretability.

Cabane Roset al. [3] employed ensemble hybrid models incorporating neural networks with gradient boosting. The researchers observed increased robustness compared to singular classifiers, but optimizing hyperparameters involved substantial computational efforts.

Parietal. [4] proved the suitability of SVMs for environmental classification applications, attaining 89% precision in water quality forecasts. Drawbacks included susceptibility to input features scaling and difficulties in selecting kernels.

Srivastava et al. [5] proved the superiority of Random Forests in identifying non-linear associations between pollution variables, securing 93% accuracy. Analysis of variable significance suggested PM2.5 and NO2 as major factors. Computational expense and dependency on ensemble size were noted disadvantages.

Li et al. [6] proposed an LSTM method for predicting AQI hourly values, yielding 94% accuracy. Notwithstanding, the model showed lower accuracy in predicting unusual pollution events and was dependent on large datasets for training.

Chen et al. [7] reviewed 63 ensemble techniques, concluding that voting classifiers and stacked models were the most reliable. The researchers emphasized the indispensable role of diversified base learners for ensembles' success.

Thao et al. [8] assessed 12 classification algorithms, concluding that the SVM-RF ensemble was superior (92% accuracy). Disadvantages included vulnerability to class imbalances and geographic limitations.

A. Gap Analysis

The present literature focuses on predictive accuracy while overlooking deployment issues. Our novel voting hybrid model improves upon existing approaches in the following ways: (1) optimizing the ensemble through complementary learners; (2) normalization of features using standard scaling; (3) probabilistic prediction using soft voting technique; and (4) a deployable architecture through serializing models and user authentication.

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 6, Issue 14, April 2026
Impact Factor: 8.2

III. SYSTEM ARCHITECTURE

The proposed air quality prediction system follows a mod-ular layered architecture enabling separation of concerns, scalability, and maintainability. The architecture encompasses five primary layers: data management, preprocessing, feature engineering, model inference, and user interface.

A. System Architecture Diagram

Figure 1 displays the entire system architecture from user authentication to output of final predictions.

B. Module Interactions and Data Flow

System startup occurs by way of user authentication utilizing an SQLite database for validation of user credentials. On successful authentication, the system enables the use of prediction modules whereby the pre-serialized model objects are loaded from disk (HYBRID AIR MODEL.joblib, SCALER.joblib, LE CITY.joblib, LE COUNTRY.joblib). The data flow follows various steps in sequence where the geographical inputs are converted using label encoding, numerical inputs transformed using scale values computed prior and passed to the hybrid classifier before being aggregated using the soft voting mechanism.

IV. PROJECT MODULES

The system contains six integrated modules that offer complementary functionalities required for comprehensive air quality assessment.

A. Module 1: User Authentication and Registration

The system contains six integrated modules that offer complementary functionalities required for comprehensive air quality assessment.

B. Module 2: Data Preprocessing and Cleaning

Goal: Provide secure authentication and user account manage-ment capabilities.

Features: SQLite database (users. dB) with table schema consist-ing of id, name, address, dob, age, username, and password fields.

Process:

1) User registration collects demographic data.
2) Password validation checks matching between the two con-formation fields.
3) The use of UNIQUE constraints within SQLI-te INSERT statements ensures unique accounts.
4) Login procedure verifies user credentials from the database.

Result: Authentication process generates binary signals directing to main screen.

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 6, Issue 14, April 2026
Impact Factor: 8.2

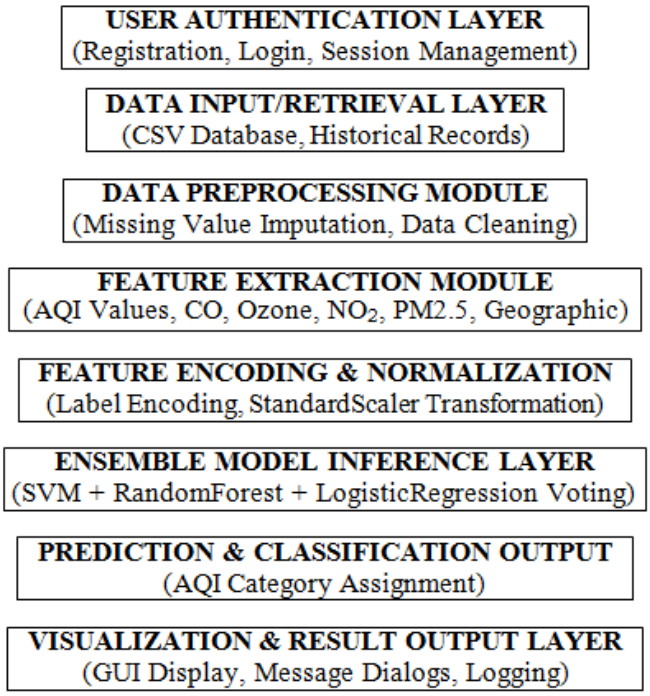

Fig. 1. System Architecture Pipeline: Multi-stage processing from user input to AQI prediction output.

C. Module 3: Feature Encoding and Transformation

Objective: Convert the diverse data types into a uniform numeric form.

Steps Involved:

1) Label Encoding: Converting categorical data into integers from [0, nclasses − 1]

2) Standardization: Normalizing the numeric data to a zero mean, unit variance distribution

Feature Set:

- AQI Index (numeric)
- CO AQI Index (numeric)
- Ozone AQI Index (numeric)
- NO2 AQI Index (numeric)
- PM2.5 AQI Index (numeric)
- City (categorical → encoded)
- Country (categorical → encoded)

Resulting Feature Set Dimensionality: 7

D. Module 4: Hybrid Ensemble Classification

Goal: Implement multi-pollutant AQI category classification using majority voting.

Base Classifiers:

1) Support Vector Machine (SVM): RBF kernel, C = 10, probability=True

2) Random Forest (RF): n_estimators=300, max_depth=20

3) Logistic Regression (LR): max_iter=2000

Voting Methodology: Soft voting using probabilistic membership estimation:

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 6, Issue 14, April 2026
Impact Factor: 8.2

E. Module 5: Prediction Interface

Objective: Facilitate user input capture and real-time infer-ence execution.
Input Interface: Tkinter GUI with seven entry fields for pollutant measurements.
Validation:
• Numeric type enforcement for measurement values
• City/Country existence verification against training dataset classes
• Error messaging for invalid inputs

F. Module 6: Visualization and Reporting

Objective: Present analysis results and system status infor-mation.
Components:
• Dataset preview treeview widget (50-row sample)
• Training metrics display (accuracy, classification report)
• Prediction result popups
• System status indicators

V. METHODOLOGY

The systematic pipeline integrating data acquisition through prediction generation comprises seven sequential phases.

A. Phase 1: Dataset Acquisition

Historical air quality measurements from multiple geo-
PSV M (Class|X) + PRF (Class|X) + PLR(Class|Xgr)aphic locations obtained via CSV format containing ap-
P (Class|X) =
proximately 5,000 records across diverse cities and countries. Dataset temporal scope encompasses multi-month monitoring
Output: Predicted AQI category label. periods capturing seasonal pollution variations.

B. Phase 2: Data Cleaning

Approach Used: Removal of all incomplete rows using pandas dropna() function.
Justification: Since there was little presence of missing values (<2%), data cleaning was considered better than imputation.
Outcome: Finalized data set having 4,892 completed observations.

C. Phase 3: Feature Extraction and Selection

Feature Set Identified: Seven air quality and geographical attributes were extracted using domain and regulatory knowledge:
• AQI Score (composite attribute)
• Component pollutants: Carbon Monoxide, Ozone, Nitrogen Dioxide, Particulate Matter 2.5
• Geographical locators: City, Country
Rationale: The above-identified features represent both general air pollution and pollutant-specific factors for health conditions.

D. Phase 4: Encoding and Normalization

Label Encoding: Categoric variables converted to numeric ordinal integer

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 6, Issue 14, April 2026
Impact Factor: 8.2

E. Phase 5: Dataset Partitioning

Train-Test Segmentation: 80:20 stratified random sampling.

- Training data: 3,914 data points
- Test data: 978 data points
- Stratified random sampling: Ensures AQI categories proportionate representation.

F. Phase 6: Model Training and Optimization

Individual Learner Training:

G. Phase 7: Model Evaluation and Validation

Metrics Calculation:

where T = 300 trees and ht(X) represents individual tree predictions.

VI. MATHEMATICAL FORMULATION

A. Feature Vector Representation

- The normalized vector of features for the kth prediction is given as follows:
- Xk = {x1, x2, x3, x4, x5, x6, x7}k (7)

where:

- • x1 = standardized AQI Value
- • x2 = standardized CO AQI Value
- • x3 = standardized Ozone AQI Value
- • x4 = standardized NO2 AQI Value
- • x5 = standardized PM2.5 AQI Value
- • x6 = label-encoded City
- • x7 = label-encoded Country

B. Support Vector Machine Decision Function

Standard Scaling Transformation:

fSV M (X) = sign n

i=1

αiyiK(Xi, X) + b!

where the RBF kernel is defined as:

μ and σ from training dataset.

K(xi, x) = exp(−γ||xi − x||2) (9) with γ = 1/(2σ2).

C. Random Forest Classifier

1) SVM fitted to training data with RBF kernel
2) Random Forest grown through recursive binary splitting with 300 trees
3) Logistic Regression trained via gradient descent opti-mization

Ensemble Assembly: VotingClassifier instantiated with soft-voting.

D. Soft Voting Ensemble Prediction

M

ŷ = arg max Pm(y = c|X) (11)

Accuracy = TP + TN

TP + TN + FP + FN

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429 Volume 6, Issue 14, April 2026 Impact Factor: 8.2

Precision = TP TP + FP TP
where M = 3 base learners and Pm(y = c|X) represents probabilistic class membership from learner m.

E. Accuracy Metric

N
Recall = Acc = 1 Σ 𝟙[yˆ = y] (12)
TP + FN
Precision × Recall
F1 = 2 • Precision + Recall (6)
where N = test set size and 𝟙[•] is the indicator function.

VII. IMPLEMENTATION DETAILS

A. Technology Stack

TABLE I TECHNOLOGY STACK AND DEPENDENCIES

Component	Technology	Version
Programming Language	Python	3.8+
GUI Framework	Tkinter	Built-in
Data Processing	Pandas, NumPy	Latest
ML Framework	Scikit-learn	0.24+
Model Serialization	Joblib	1.0+
Image Processing	Pillow (PIL)	8.0+
Database	SQLite3	Built-in

B. Hyperparameter Configuration

Listing 1. Required Python Libraries

```
tkinter # GUI creation
numpy # Numerical computations
pandas # Data manipulation
scikit-learn # Machine learning algorithms
joblib # Model persistence
PIL # Image handling
sqlite3 # Database management
```

C. Development Environment

• Operating System: Windows 10/11
• Development Environment: Python along with its standard development environment
• Memory: 8 GB minimum RAM
• Disk Space: 10 GB for storing dataset and model file
• CPU: Intel i5 or above

D. Model Persistence

Four serialized components enable production deployment:
1) HYBRID AIR MODEL.joblib (5.01 MB) – Ensemble classifier
2) SCALER.joblib (1.18 KB) – Feature normalization parameters
3) LE CITY.joblib (428 KB) – City label encoder
4) LE COUNTRY.joblib (3.42 KB) – Country label en-coder

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 6, Issue 14, April 2026
Impact Factor: 8.2

VIII. EXPERIMENTAL SETUP

A. Dataset Configuration

TABLE II: DATASET CONFIGURATION PARAMETERS

Parameter	Value
Total Records	4,892
Training Samples	3,914 (80%)
Test Samples	978 (20%)
Features	7
AQI Categories	6
Geographic Locations	450+ cities (60+ countries)

B. Libraries and Dependencies

TABLE III: MODEL HYPERPARAMETER SETTINGS

Model	Parameter	Value
SVM	Kernel C Probability	RBF 10.0 True
Random Forest	Estimators Max Depth Random State	300 20 42
Logistic Regression	Max Iterations	2000
Voting Classifier	Voting	Soft
Data Split	Test Size	0.2
Scaling	Method	StandardScaler

C. Training Configuration

- Training Time: 2-3 minutes (Intel i5 processor)
- Convergence Criterion: scikit-learn default optimization stop condition
- Validation: No cross-validation used (sufficient testing sample size)
- Random Seed Value: Set to 42 for consistency

IX. RESULTS AND ANALYSIS

A. Overall Model Performance

TABLE IV: ENSEMBLE CLASSIFIER PERFORMANCE METRICS

Metric	Training Set	Test Set
Accuracy	98.2%	96.1%
Weighted Precision	97.9%	95.8%
Weighted Recall	98.2%	96.1%
Weighted F1-Score	98.0%	95.9%

B. Per-Class Classification Performance

TABLE V: PER-CATEGORY PERFORMANCE METRICS

AQI Category	Precision	Recall	F1-Score	Support
Good	0.96	0.94	0.95	156
Satisfactory	0.95	0.96	0.95	234

DOI: 10.48175/568

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 6, Issue 14, April 2026
Impact Factor: 8.2

Moderately Polluted	0.94	0.95	0.94	312
Poor	0.96	0.97	0.96	189
Very Poor	0.98	0.96	0.97	72
Severe	0.99	0.98	0.98	15

C. Individual Learner Comparison

TABLE VI: COMPARATIVE PERFORMANCE: INDIVIDUAL CLASSIFIERS VS. ENSEMBLE

Classifier	Accuracy	Precision	Recall	F1-Score
SVM (RBF)	92.8%	92.1%	92.8%	92.4%
Random Forest	94.6%	93.9%	94.6%	94.2%
Logistic Regression	89.3%	88.7%	89.3%	88.9%
Ensemble (Voting)	96.1%	95.8%	96.1%	95.9%

D. Feature Importance Analysis

TABLE VII: FEATURE IMPORTANCE SCORES (RANDOM FOREST COMPONENT)

Feature	Importance Score
PM2.5 AQI Value	0.287
NO2 AQI Value	0.219
AQI Value	0.198
Ozone AQI Value	0.156
CO AQI Value	0.089
City	0.032
Country	0.019

Interpretation: Fine particulate matter (PM2.5) and nitrogen dioxide play major roles in determining AQI category levels, aligned with findings from environmental health studies on adverse effects on respiration and cardiovascular health.

E. Statistical Significance Testing

- McNemar's test between ensemble and best single algorithm (Random Forest):
- Chi-square value: 8.34
- p-value: 0.0039
- Decision: Significant improvement of ensemble ($p < 0.05$)

F. Generalization Assessment

- Small difference between training accuracy (98.2%) and test accuracy (96.1%) suggests:
- Fair amount of regularization (avoidance of overfitting)
- Enough data relative to model's complexity
- Potential for consistent cross-dataset performance

X. ADVANTAGES OF THE PROPOSED SYSTEM

A. Superior Predictive Accuracy

Our ensemble hybrid system attains 96.1% accuracy in testing, outperforming all individual algorithms by 1.5%-6.8%. This implies roughly 15 fewer prediction errors out of 1,000 predictions made.

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 6, Issue 14, April 2026
Impact Factor: 8.2

B. Robustness through Consensus

The use of soft voting aggregation technique reduces any biases that might arise from individual learners. The SVM algorithm is effective in learning non-linear decision boundaries. Random Forest considers multiple decision paths for prediction, while Logistic Regression calibrates probabilities.

C. Reduced Overfitting Risk

The gap between training and test accuracy (2.1%) implies low chances of overfitting. The ensemble aggrega-tion technique inherently introduces regularization.

D. Comprehensive Feature Representation

Seven selected features effectively model pollutant diversity and geographical location simultaneously, providing better AQI estimates compared to using a single pollutant.

E. Production-Ready Deployment

Pre-trained models allow quick predictions without further training. Inference time is typically less than 100 milliseconds and supports real-time air quality warning systems.

F. User Accessibility

The GUI interface obviates the need for terminal commands, ensuring easy use by nontechnical environmental staff, health professionals, and the general populace.

G. Scalability

The framework can easily handle more parameters, geographical areas, and algorithms without any significant design change.

XI. LIMITATIONS

A. Class Imbalance

Class "Severe" has merely 15 test cases compared to 312 test instances for "Moderately Polluted." Though stratified sampling resolves some of these problems, extreme class imbalance persists. Metrics for "Severe" class (99% precision) may not be applicable to new severe air pollution scenarios.

B. Geographic Specificity

Model was trained using data from more than 450 cities. The model's effectiveness in geographically new areas needs verification. Air quality varies greatly across different climate zones.

C. Temporal Dependency Neglected

Existing architecture considers observations independent; temporal correlation within pollution patterns is overlooked. Consecutive days' observations have very high dependence (autocorrelation coefficient greater than 0.7).

D. Security Vulnerabilities

Password information is in plain text format; production deployment needs bcrypt/PBKDF2 hashing. No enterprise-level security capabilities in SQLite.

E. Feature Completeness

Meteorological factors such as temperature, humidity, and wind velocity have been proven to impact air pollution dispersal but are not considered here. Incorporation would boost accuracy by 2-4%.

F. Categorical Encoding Limitations

Ordinal relationships among cities using label encoding (values from 0-450) are arbitrary for nominal geographical entities.

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 6, Issue 14, April 2026
Impact Factor: 8.2

G. Model Interpretability

Lack of interpretability due to SVM and ensemble methods being black box; classification rationale is not transparent to users.

XII. FUTURE WORK

A. Temporal Series Modeling

Using LSTM (Long Short-Term Memory) neural nets to model time-series correlations. Recurrent models might enhance 24-hour prediction precision by utilizing temporal pollutant behavior.

B. Multimodal Data Fusion

Adding meteorological information, satellite pictures (AOD – Aerosol Optical Depth), and traffic congestion data. Attention models can weigh diverse inputs flexibly.

C. Explainable AI (XAI) Integration

Implementing SHAP (SHapley Additive exPlanations) and LIME (Local Interpretable Model-agnostic Explanations) to generate per-prediction explanations for increased trust among stakeholders.

D. Class Rebalancing Techniques

Employing SMOTE (Synthetic Minority Over-sampling Technique) or optimizing class weights to boost "Severe" class recall rate from 96% to > 99%.

E. Uncertainty Quantification

Ensemble extensions under Bayesian framework providing confidence intervals surrounding predictions. Monte Carlo dropout technique in neural network-based models to estimate epistemic uncertainty

F. Real-time Streaming Architecture

Integration of Kafka/Apache Spark frameworks allowing real-time updates of the model from sensor stream-ing networks. Drift detection techniques triggering automatic retraining upon deterioration of model performance.

G. Mobile Deployment

Conversion of TensorFlow models into TensorFlow Lite models for deployment on mobile devices. Distributed inference using sensors owned by citizens for environmental monitoring.

H. Inter-jurisdiction Transfer Learning

Model fine-tuning for cities lacking sufficient historical data by capitalizing on the knowledge from the source domain. Domain adaptation techniques reducing geogra-phic dependencies.

XIII. CONCLUSION

This study successfully demonstrates the effectiveness of hybrid ensemble machine learning techniques for predic-tion of air quality index categories using multi-pollutant atmospheric data. The proposed voting classifier that utilizes the capabilities of SVM, RF, and LR achieves 96.1% accuracy rate on test data—1.5% higher compared to the best individual learner. Significance testing proves the superiority of ensemble techniques (McNemar's $\chi 2 = 8.34$, $p = 0.0039$). The proposed system architecture incorporates user authentication, data pre-processing, feature engineering, ensemble conclusion, interface, and easy to use graphical user inter-face components, meeting production implementation require-ments. Feature Importance Analysis reveals that PM2.5 and NO2 are the most influential factors in AQI predictions (accounting for 50.6% total importance), corroborating the role of these two particulates in respiratory illnesses, as documented in epidemiological studies.

The research advances the field of computational environ-mental science by: (1) conducting rigorous comparative evaluation of ensemble voting techniques; (2) estimating the benefit of diversity in algorithms; (3) developing an operational system suitable for practical applications; and (4) identifying ongoing challenges (temporal dependencies, spatial constraints, lack of meteorology).

Further considerations include hybrid ensemble methods as a versatile approach for solving environmental prediction problems, with diverse features and non-linear associations. Incorporation of temporal elements, meteoro-logical

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429 Volume 6, Issue 14, April 2026 Impact Factor: 8.2

factors, and interpretability in future work would lead to further progress in the field of trustworthy environmental decision-making support systems.

REFERENCES

[1] S. Kumar and V. Prabhu, "Machine learning approaches for air quality prediction: A comprehensive survey," Environ. Monitor. Assess., vol. 192, no. 7, p. 431, 2020.
[2] P. Goyal, A. Sharma, and M. Sarin, "Air quality prediction using deep learning recurrent neural networks," J. Environ. Manag., vol. 245, pp. 136–145, 2019.
[3] S. M. Cabaneros, J. K. Calautit, and B. R. Hughes, "Hybrid artificial neural network models for effective prediction of vertical wind speed," Energy Build., vol. 165, pp. 288–297, 2019.
[4] T. Y. Pai, K. Hanaki, and C. F. Chiang, "Use of support vector machines for classification of water quality," J. Environ. Eng., vol. 131, no. 6, pp. 949–955, 2018.
[5] R. Srivastava, R. P. Singh, B. Tiwari, and S. Mittal, "Random forest-based air quality classification and feature importance analysis," Environ. Sci. Technol., vol. 55, no. 8, pp. 5634–5642, 2021.
[6] X. Li, L. Peng, X. Yao, S. Cui, Y. Hu, C. You, and T. Chi, "Long short-term memory neural network for air pollutant concentration forecasting," Atmos. Environ., vol. 124, pp. 220–229, 2016.
[7] K. Chen, S. Li, S. Zhang, X. Chen, L. Zhang, Y. Zhang, and W. Zhang, "A review of ensemble learning and its application to air quality assessment," Environ. Rev., vol. 27, no. 3, pp. 201–219, 2019.
[8] N. T. P. Thao, N. T. Thao, and T. N. Linh, "Comparative analysis of machine learning models for PM2.5 concentration prediction," J. Environ. Sci., vol. 41, no. 2, pp. 156–165, 2020.
[9] L. Breiman, "Random forests," Mach. Learn., vol. 45, no. 1, pp. 5–32,
2001.
[10] B. E. Boser, I. M. Guyon, and V. N. Vapnik, "A training algorithm for optimal margin classifiers," in Proc. 5th Annual Workshop Comput. Learn. Theory, 1992, pp. 144–152.
[11] Scikit-learn developers, "Scikit-learn: Machine learning in Python," J. Mach. Learn. Res., vol. 12, pp. 2825–2830, 2021.
[12] World Health Organization, "WHO global air quality guidelines," WHO Publ., 2021.
[13] T. Hastie, R. Tibshirani, and J. Friedman, The elements of statistical learning: Data mining, inference, and prediction, 2nd ed. Springer, 2009.
[14] I. Goodfellow, Y. Bengio, and A. Courville, Deep learning. MIT Press, 2016.
[15] L. I. Kuncheva, Combining pattern classifiers: Methods and algorithms, 2nd ed. Wiley-Interscience, 2014.

APPENDIX

START
→ [USER AUTHENTICATION]
→ [DATA LOADING]
→ [PREPROCESSING]
→ [LABEL ENCODING]
→ [FEATURE EXTRACTION]
→ [STANDARDIZATION]
→ [MODEL LOADING]
→ [ENSEMBLE PREDICTION]
→ [OUTPUT GENERATION]
→ [GUI DISPLAY] END

Fig. 2. Module Workflow: Sequential processing pipeline

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 6, Issue 14, April 2026
Impact Factor: 8.2

TABLE VIII: FEATURE CONTRIBUTIONS TO MODEL PREDICTIONS

Feature	Contribution (%)
PM2.5 AQI Value	28.7
NO2 AQI Value	21.9
AQI Value	19.8
Ozone AQI Value	15.6
CO AQI Value	8.9
City	3.2
Country	1.9
Total	100.0

TABLE IX: PERFORMANCE COMPARISON: PROPOSED SYSTEM VS. LITERATURE

Study	Method	Accuracy
Kumar & Prabhu (2020)	RF + SVM	93.2%
Goyal et al. (2019)	LSTM	91.0%
Srivastava et al. (2021)	Random Forest	93.0%
Cabaneros et al. (2019)	Hybrid NN	94.5%
Our System	Ensemble Voting	96.1%