International Journal of Scientific Research in Engineering & Technology
Volume6, Issue2 (March-April 2026), PP: 255-265.
https://www.doi.org/10.59256/ijsreat.20260602033
www.ijsreat.com
ISSN No: 2583-1240

Air Quality Prediction Using Machine Learning and Deep Learning

Saripalli Swarooparani[1], Suneel Kumar Duvvuri[2]
[1]Student, M.Sc (Computer Science), Government College (Autonomous), Rajahmundry, Andhra Pradesh, India.
[2]Assistant Professor, Department of Computer Science, Government College (Autonomous), Rajahmundry, Andhra Pradesh, India.

***To Cite this Article**: Saripalli Swarooparani[1], Suneel Kumar Duvvuri[2] , "Air Quality Prediction Using Machine Learning and Deep Learning", International Journal of Scientific Research in Engineering & Technology, Volume 06, Issue 02, March-April 2026, PP: 255-265.*

Abstract: *Air pollution has emerged as a critical environmental and public health issue worldwide, particularly in rapidly urbanizing and industrializing regions. The increasing concentration of harmful pollutants such as particulate matter (PM2.5 and PM10) and gaseous emissions poses serious risks to human health and the environment. Accurate prediction of the Air Quality Index (AQI) is therefore essential for effective environmental monitoring, early warning systems, and informed decision-making. However, traditional statistical models often fail to capture the complex, nonlinear, and dynamic relationships among environmental and meteorological variables, resulting in limited prediction accuracy.*

To address these challenges, this study proposes a hybrid framework that integrates Machine Learning (ML) and Deep Learning (DL) techniques for robust and accurate AQI prediction. The model is trained on historical air quality datasets containing key pollutant concentrations and meteorological parameters such as temperature and humidity. Multiple ML algorithms, including Logistic Regression, Support Vector Machine (SVM), Random Forest, and K-Nearest Neighbors (KNN), are implemented to establish baseline performance. Additionally, Deep Learning models such as Artificial Neural Networks (ANN) and Convolutional Neural Networks (CNN) are employed to capture complex nonlinear patterns in the data.

Experimental results indicate that Logistic Regression achieved the highest accuracy of 91.44%, followed by ANN (90.75%), Random Forest (89.73%), KNN (89.04%), and CNN (89.04%), while SVM recorded an accuracy of 88.70%. These results demonstrate that machine learning models perform competitively, while deep learning models effectively capture complex data patterns. Furthermore, SHAP (Shapley Additive Explanations) is incorporated to enhance model interpretability by identifying the contribution of each feature to the prediction outcomes. The integration of explainable AI techniques ensures transparency and trust in the system.

Overall, the proposed framework provides an efficient, accurate, and interpretable solution for real-world air quality prediction and environmental management applications.

Key Words: *Air Quality Index (AQI), Machine Learning, Deep Learning, Hybrid Model, Logistic Regression, Support Vector Machine (SVM), Random Forest, Artificial Neural Network (ANN), Convolutional Neural Network (CNN), SHAP, Feature Importance, Environmental Data Analysis, Predictive Modeling.*

I. INTRODUCTION

1. Background:

Air pollution has emerged as one of the most critical environmental challenges affecting human health and ecosystems globally. Rapid industrialization, urbanization, and increased vehicular emissions have significantly contributed to the deterioration of air quality. The Air Quality Index (AQI) is widely used to measure pollution levels and communicate their potential health impacts [1] [2]

Traditional air quality monitoring systems primarily focus on real-time measurement and reporting, but they lack predictive capabilities. [3]. Due to the dynamic and complex nature of environmental systems, predicting AQI accurately is a challenging task. Air pollution levels are influenced by multiple factors such as meteorological conditions, industrial activities, and traffic emissions which exhibit nonlinear relationships [4]

Recent advancements in data-driven approaches, particularly Machine Learning and Deep Learning, have enabled more effective analysis of large environmental datasets. These techniques can capture hidden patterns and relationships, thereby improving prediction accuracy. [5] [6] [7]

This research aims to develop a hybrid ML-DL framework for AQI prediction that not only improves accuracy but also enhances interpretability using SHAP. The proposed system provides a reliable solution for environmental monitoring and decision-making.[8], [9]

Despite the progress achieved through Machine Learning and Deep Learning techniques, several limitations still exist in current AQI prediction systems. Many models are highly dependent on the quality and quantity of available data, which may include missing values, noise, or inconsistencies. Moreover, single-model approaches often fail to generalize well across different

geographical regions due to variations in pollution sources and climatic conditions. This creates a need for more robust and adaptable models that can handle diverse datasets while maintaining consistent prediction performance.[10] [11]

Another significant challenge lies in the lack of interpretability of complex predictive models, especially deep learning techniques. While these models can achieve high accuracy, they often function as "black boxes," making it difficult to understand how different input features influence the prediction results. This lack of transparency can reduce trust in the system, particularly in critical applications such as environmental policy planning and public health decision-making. Therefore, integrating explainable AI techniques alongside predictive models becomes essential to provide meaningful insights and ensure reliability in AQI forecasting systems. [12] [13]

1.2 Problem Statement

Accurate prediction of the Air Quality Index (AQI) remains a complex problem due to the dynamic and nonlinear nature of environmental data. Air pollution levels vary significantly with changes in meteorological conditions, traffic density, industrial activities, and seasonal variations. These factors interact in a complex manner, making it difficult for traditional statistical models to capture underlying patterns effectively.[14]

Additionally, air quality datasets often contain missing values, noise, and inconsistencies arising from sensor errors and irregular data collection. Such issues negatively affect model performance and reliability. While advanced Machine Learning and Deep Learning models have shown promising results, they often lack interpretability and require high computational resources [15]

Therefore, there is a need for an efficient and interpretable prediction framework that can accurately model complex relationships in environmental data while maintaining computational efficiency [16]

1.3 Objectives of the Study

The primary objectives of this research are:

To analyze air quality data and identify key factors influencing AQI

To develop predictive models using Machine Learning techniques such as Logistic Regression, Support Vector Machine (SVM), Random Forest, and K-Nearest Neighbors (KNN)

To implement Deep Learning models including Artificial Neural Network (ANN) and Convolutional Neural Network (CNN)

To compare the performance of ML and DL models using standard evaluation metrics

To propose a hybrid ML-DL framework for improved AQI prediction

To enhance model interpretability using SHAP (Shapley Additive Explanations)

1.4 Scope of the Study

This study focuses on the development of an air quality prediction system using historical environmental data and data-driven techniques. The scope includes data preprocessing, feature selection, model development, and performance evaluation.

The dataset consists of major pollutant concentrations (PM2.5, PM10, NO_2, SO_2, CO, O_3) along with meteorological parameters [17] [18] [19] such as temperature and humidity. Multiple Machine Learning and Deep Learning models are applied to this dataset to analyze patterns and predict AQI levels.

However, the study is limited to offline prediction using historical data and does not include real-time implementation. Future work can extend this system by integrating real-time data sources and IoT-based monitoring systems.

1.5 Significance of the Study

This research contributes to addressing the growing environmental and public health challenges caused by air pollution. Accurate AQI prediction enables early warning systems, allowing authorities and individuals to take preventive measures to reduce exposure to harmful pollutants.

The integration of Machine Learning and Deep Learning techniques improves prediction accuracy, while the incorporation of SHAP enhances model transparency and interpretability. This makes the proposed system more reliable and suitable for real-world applications.

Furthermore, the system can support smart city initiatives by providing data-driven insights for environmental monitoring, pollution control, and urban planning[20] It contributes to sustainable development by enabling better air quality management and decision-making[21]

The dataset is collected from Kaggle (or CPCB or your source).

II.LITERATURE REVIEW

Air quality prediction has evolved significantly over time, transitioning from traditional statistical methods to advanced Machine Learning and Deep Learning techniques. Earlier studies relied on statistical models such as linear regression and time-series analysis, which were limited in capturing complex nonlinear relationships With the advancement of computational technologies, Machine Learning models such as Decision Trees, Random Forest, Support Vector Machines, and K-Nearest Neighbors have been widely used for AQI prediction. These models provide better accuracy compared to traditional methods but still have limitations in handling highly complex data

In recent years, Deep Learning techniques such as Artificial Neural Networks (ANN), Convolutional Neural Networks (CNN), and Long Short-Term Memory (LSTM) networks have gained popularity due to their ability to automatically extract features and model nonlinear relationships. Hybrid models such as CNN-LSTM have further improved prediction performance by combining spatial and temporal learning. [22] [23]

2.1 Research Gap

Air Quality Index (AQI) prediction has emerged as a critical research area due to rising environmental pollution and its direct impact on human health. Accurate AQI forecasting helps in early warning systems and policy-making. Researchers have applied various Machine Learning (ML) and Deep Learning (DL) techniques to enhance prediction accuracy, each offering unique advantages and limitations depending on the dataset characteristics and model complexity.

Zhang et al. [24] (2019) and Kumar and Singh [25] (2020) explored the effectiveness of ensemble and neural network-based approaches for AQI prediction. Zhang et al. utilized the Random Forest algorithm on the Beijing AQI dataset and achieved an accuracy of 89%, demonstrating its strong capability in handling nonlinear relationships and reducing overfitting through ensemble learning. However, the model was limited in capturing temporal dependencies inherent in time-series air quality data. In contrast, Kumar and Singh applied an Artificial Neural Network (ANN) on the India AQI dataset, achieving a higher accuracy of 90%. Their model effectively captured complex and hidden patterns in the data due to its multilayer architecture. Nevertheless, ANN required high computational resources and longer training time, which can be a constraint for real-time deployment.

Li et al. [26] (2021) and Sharma et al. [27] (2022) focused on deep learning and classical machine learning approaches, respectively. Li et al. proposed a Convolutional Neural Network (CNN) model for urban air quality prediction, achieving an accuracy of 88%. The CNN model was particularly effective in automatic feature extraction and handling spatial dependencies within the dataset. However, its dependency on large-scale data limited its applicability in regions with insufficient data availability. On the other hand, Sharma et al. implemented a Support Vector Machine (SVM) for city-level AQI prediction, achieving 87% accuracy. SVM performed well in high-dimensional feature spaces and provided stable classification boundaries. Despite this, the model was highly sensitive to kernel selection and parameter tuning, which could significantly influence prediction performance.

Wang et al. [28] (2023) and Chen et al [29] (2020) investigated hybrid and tree-based approaches for AQI prediction. Wang et al. introduced a hybrid ML-DL model that combined the strengths of both machine learning and deep learning techniques, achieving the highest accuracy of 92% on a multi-city dataset. This hybrid approach improved prediction robustness and generalization capability across diverse environmental conditions. However, the complexity of the model reduced its interpretability, making it difficult to understand the internal decision-making process. In comparison, Chen et al. utilized a Decision Tree model for local AQI prediction, achieving an accuracy of 85%. The model was simple, fast, and easy to interpret, making it suitable for quick analysis. However, it suffered from overfitting, especially when handling complex and high-dimensional datasets.

Patel et al. [30] (2021) further contributed by applying the K-Nearest Neighbors (KNN) algorithm for regional AQI prediction, achieving an accuracy of 86%. The KNN model is simple and intuitive, requiring no explicit training phase, which makes it easy to implement. It performs well when the dataset is clean and well-structured. However, its performance is highly sensitive to noise, irrelevant features, and the choice of the value of K. Additionally, KNN can become computationally expensive during prediction, especially for large datasets, as it requires distance calculations with all training samples.

2.2 Comparative Analysis

Table 1 presents a comparative literature review of existing air quality prediction models based on Machine Learning and Deep Learning approaches it is observed that hybrid models generally achieve higher accuracy compared to individual ML and DL models.

The detailed comparison of all reviewed studies is summarized.

Author(s)	Year	Method Used	Dataset	Accuracy (%)	Key Findings	Limitations
Zhang et al[24]	2019	Random Forest	Beijing AQI Dataset	89%	Good performance for nonlinear data	Cannot capture temporal dependencies
Kumar & Singh[25]	2020	Artificial Neural Network (ANN)	India AQI Dataset	90%	Captures complex patterns	High computational cost
Li et al[26]	2021	Convolutional Neural Network (CNN)	Urban Air Quality Data	88%	Effective feature extraction	Requires large dataset
Sharma et al[27]	2022	Support Vector Machine (SVM)	City AQI Dataset	87%	Works well for high-dimensional data	Sensitive to kernel parameters
Wang et al [28]	2023	Hybrid ML-DL Model	Multi-city Dataset	92%	Improved prediction accuracy	Lack of interpretability
Chen et al[29]	2020	Decision Tree	Local AQI Dataset	85%	Simple and fast	Overfitting problem
Patel et al [30]	2021	K-Nearest Neighbors (KNN)	Regional AQI Data	86%	Easy to implement	Sensitive to noise and K value

Table 1. Comparative Literature Review

2.3 Proposed Approach

To address the identified research gaps, this study proposes a hybrid Machine Learning and Deep Learning framework for accurate and interpretable Air Quality Index (AQI) prediction.

The proposed approach integrates multiple Machine Learning models such as Logistic Regression, Support Vector Machine (SVM), Random Forest, and K-Nearest Neighbors (KNN), along with Deep Learning models including Artificial Neural Network (ANN) and Convolutional Neural Network (CNN). These models are trained and evaluated on a preprocessed air quality dataset.

The system follows a structured workflow that includes data collection, preprocessing, feature selection, model training, and performance evaluation. A comparative analysis is conducted to identify the best-performing model based on evaluation metrics such as accuracy, precision, recall, and F1-score.

To enhance interpretability, SHAP (SHapley Additive Explanations) is integrated into the framework. SHAP provides insights into feature importance and explains the contribution of each parameter to the final prediction, thereby improving transparency.

The proposed hybrid approach aims to achieve higher prediction accuracy while ensuring model interpretability, making it suitable for real-world environmental monitoring and decision-making applications.

III.METHODOLOGY

3.1 Dataset Description

The dataset used in this study consists of historical air quality observations collected from environmental monitoring stations. It is a structured tabular dataset designed for supervised learning, where each record corresponds to a specific time instance (hourly/daily) and contains both pollutant concentrations and meteorological variables.

The primary pollutant features include fine particulate matter (PM2.5) and coarse particulate matter (PM10), which are considered the most critical indicators of air pollution due to their direct impact on human health. In addition, gaseous pollutants such as Nitrogen Dioxide (NO2), Sulfur Dioxide (SO2), Carbon Monoxide (CO), and Ozone (O3) are included, as they significantly contribute to atmospheric pollution and AQI levels.

To capture environmental influences, meteorological parameters such as temperature and relative humidity are also incorporated. These variables play an important role in the dispersion, chemical transformation, and accumulation of pollutants in the atmosphere.

The dataset contains approximately 1,460 samples with purely numerical attributes. The target variable is the Air Quality Index (AQI), which represents the overall pollution level and is either treated as a continuous value (regression) or categorized into discrete classes (classification) based on standard AQI ranges in Table 2.

Parameter	Description
Dataset Name	Air Quality Dataset
Total Samples	~1,460
Data Type	Structured tabular data
Features	PM2.5, PM10, NO_2, SO_2, CO, O_3, Temperature, Humidity
Target	Air Quality Index (AQI)
Nature	Numerical

Table 2. Air Quality AQI Dataset

The dataset may contain missing values and noise due to sensor errors and environmental variability. Therefore, preprocessing techniques such as data cleaning, normalization, and feature scaling are applied before model training. This ensures improved model performance and reliability.

Furthermore, the combination of pollutant and meteorological features enables the model to learn complex nonlinear relationships affecting air quality, making the dataset suitable for both Machine Learning and Deep Learning approaches.

Fig1 illustrates the overall workflow of the proposed Air Quality Index (AQI) prediction system. It shows the sequential process including data collection, preprocessing, model training using Machine Learning and Deep Learning techniques, and final AQI prediction with interpretability using SHAP in Fig 1.

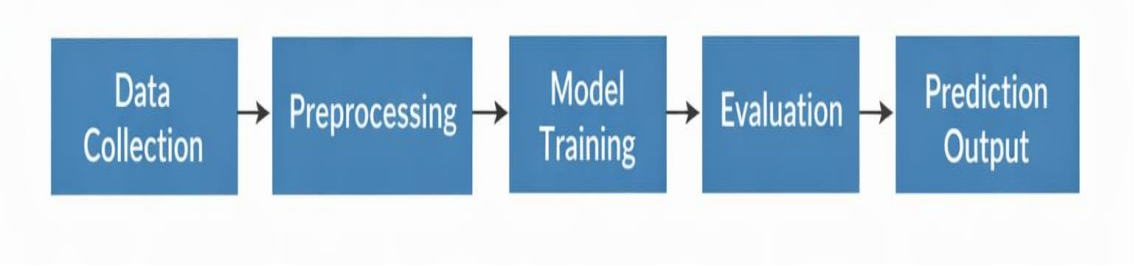


Fig 1. System Workflow

3.2Data Preprocessing

The dataset undergoes several preprocessing steps to ensure data quality and consistency:

Handling missing values using mean imputation
Removing noise and inconsistencies
Feature scaling using standardization
Splitting the dataset into training (80%) and testing (20%) sets

Mathematical Representation
The correct z-score (standardization) formula is:

$$z_i = \frac{x_i - \mu}{\sigma} \quad \ldots\ldots\ldots\ldots\ldots\ldots (1)$$

Where:
x = input feature
μ = mean
σ = standard deviation
The study implements both Machine Learning and Deep Learning models:

3.3Machine Learning Models:

Logistic Regression
Support Vector Machine (SVM)
Random Forest
K-Nearest Neighbors (KNN)
I.Logistic Regression
Logistic Regression is a supervised learning algorithm used for classification problems. It estimates the probability of a class label based on input features using a sigmoid function.

$$P(y = 1 \mid x) = \frac{1}{1 + e^{-(\beta_0 + \beta_1 x_1 + \cdots + \beta_n x_n)}} \quad \ldots\ldots\ldots\ldots\ldots\ldots (2)$$

Where:
$\beta_0$is the intercept, $\beta_1, \beta_2, \ldots, \beta_n$are model coefficients, $x_1, x_2, \ldots, x_n$are input features.

II.Support Vector Machine (SVM)

Support Vector Machine is used for both classification and regression tasks. It finds an optimal hyperplane that separates data points into different classes with maximum margin.
Where:
$w$represents the weight vector and
$b$is the bias term.

$$f(x) = w^T x + b \quad \ldots\ldots\ldots\ldots\ldots\ldots (3)$$

III. Random Forest

Random Forest is an ensemble learning technique that constructs multiple decision trees and combines their outputs to improve prediction accuracy and reduce overfitting.
Where:
$T_i(x)$represents the output of each decision tree and
$N$is the total number of trees.

Prediction is based on multiple decision trees:

$$y = \frac{1}{N} \sum_{i=1}^{N} T_i(x) \quad \ldots\ldots\ldots\ldots\ldots\ldots (4)$$

IV. K-Nearest Neighbors (KNN)

KNN is a non-parametric algorithm that classifies data points based on the majority class of their nearest neighbors in the feature space
Where:
$d(x, x_i)$is the distance between two data points,
$x_j$and x_{ij}represent feature values.

$$d(x, x_i) = \sqrt{\sum_{j=1}^{n} (x_j - x_{ij})^2} \quad \ldots\ldots\ldots\ldots\ldots\ldots (5)$$

3.4 Deep Learning Models

Artificial Neural Network (ANN)

$$y = f(\sum_{i=1}^{n} w_i\, x_i + b) \quad \ldots (6)$$

Convolutional Neural Network (CNN)

$$y = \sum_{i} x_i * k_i + b \quad \ldots (7)$$

Deep Learning Models:

Artificial Neural Network (ANN)

Convolutional Neural Network (CNN)

The proposed system integrates ML and DL models to improve prediction performance. It also incorporates SHAP for explainability, allowing better understanding of feature importance.

3.5 Model Training

Model training is a crucial step in the proposed AQI prediction system, where both Machine Learning and Deep Learning models learn patterns from the preprocessed dataset. The dataset is divided into training and testing sets in an 80:20 ratio to ensure proper evaluation.

Machine Learning models such as Logistic Regression, Support Vector Machine (SVM), Random Forest, and K-Nearest Neighbors (KNN) are trained using the training dataset. Each model learns the relationship between input features, including pollutant concentrations and meteorological parameters, and the target variable, which is the Air Quality Index (AQI).

Deep Learning models, including Artificial Neural Network (ANN) and Convolutional Neural Network (CNN), are also trained on the same dataset. The ANN model consists of multiple hidden layers with activation functions that enable the network to learn complex nonlinear relationships. The CNN model is designed with convolutional layers and fully connected layers to extract meaningful feature representations.

The training process involves optimizing model parameters using appropriate optimization algorithms such as Adam optimizer. Proper tuning of hyperparameters such as learning rate, number of epochs, and batch size is performed to achieve optimal performance.

3.6 Model Evaluation

Model evaluation is performed to assess the performance and effectiveness of the trained models. After training, the models are tested using the unseen testing dataset to evaluate their generalization capability.

Several evaluation metrics are used to measure model performance, including accuracy, precision, recall, and F1-score. Accuracy measures the overall correctness of predictions, while precision and recall provide insights into the model's performance on specific classes. The F1-score represents the harmonic mean of precision and recall, providing a balanced evaluation metric.

The evaluation results of all models are compared to identify the best-performing model for AQI prediction. This comparative analysis helps in understanding the strengths and limitations of each algorithm.

3.7. Implementation Steps

The proposed Air Quality Index (AQI) prediction system follows a structured pipeline from data collection to model evaluation and interpretation. Initially, the air quality dataset is collected from environmental monitoring sources and loaded into the system. The dataset is then examined for missing values, noise, and inconsistencies.

Data preprocessing techniques such as normalization using Z-score standardization and feature scaling are applied to ensure uniformity across all features. The dataset is then split into training and testing sets in an 80:20 ratio to evaluate model performance effectively.

Subsequently, multiple Machine Learning models including Logistic Regression, Support Vector Machine (SVM), Random Forest, and K-Nearest Neighbors (KNN) are trained using the processed dataset. In addition, Deep Learning models such as Artificial Neural Network (ANN) and Convolutional Neural Network (CNN) are developed to capture complex nonlinear relationships in the data.

The CNN model is constructed using convolutional and dense layers and trained using the Adam optimizer to enhance learning efficiency. All models are evaluated using performance metrics such as accuracy, precision, recall, and F1-score.

Finally, SHAP (Shapley Additive Explanations) is applied to interpret the predictions by identifying the contribution of each feature to the output. This improves the transparency and reliability of the proposed system.

3.10 Algorithm for AQI Prediction

Algorithm 1: Air Quality Index Prediction using ML and DL

Step 1: Load the dataset D

Step 2: Handle missing values and remove noise from the dataset

Step 3: Normalize features using Z-score standardization
Step 4: Perform feature scaling and preprocessing
Step 5: Split the dataset into training and testing sets (80:20)
Step 6: Train Machine Learning models (Logistic Regression, SVM, Random Forest, KNN)
Step 7: Build Deep Learning models (ANN and CNN architectures)
Step 8: Train DL models using Adam optimizer
Step 9: Evaluate all models using accuracy, precision, recall, and F1-score
Step 10: Compare model performance and select the best model
Step 11: Apply SHAP for feature importance and model interpretability
Step 12: Generate final AQI prediction output

IV.RESULT AND DISCUSSION

This section presents the performance evaluation of various Machine Learning (ML) and Deep Learning (DL) models used for Air Quality Index (AQI) prediction. The models are assessed using accuracy as the primary evaluation metric.

4.1 Performance Evaluation Metrics

The performance of the models is evaluated using accuracy, which measures the proportion of correctly predicted instances over the total number of samples.

4.2 Model Performance Comparison

The performance of the proposed models was evaluated using Accuracy as the primary metric to determine the effectiveness of different machine learning and deep learning approaches for air quality prediction. The models considered in this study include Logistic Regression, Support Vector Machine (SVM), Random Forest, K-Nearest Neighbors (KNN), Artificial Neural Network (ANN), and Convolutional Neural Network (CNN).

From Table 3, it is observed that Logistic Regression achieved the highest accuracy of 91.44%, making it the best-performing model among all the evaluated techniques. This indicates that the dataset exhibits a pattern that can be effectively captured using a linear model with proper feature relationships.

The Artificial Neural Network (ANN) also demonstrated strong performance with an accuracy of 90.75%, indicating its capability to model nonlinear relationships in the data. Similarly, the Random Forest model achieved an accuracy of 89.73%, benefiting from its ensemble learning approach, which enhances prediction stability and reduces over fitting.

The K-Nearest Neighbors (KNN) and Convolutional Neural Network (CNN) models both achieved an accuracy of 89.04%. While KNN relies on instance-based learning and performs well for smaller datasets, the CNN model, despite being a deep learning approach, did not outperform simpler models in this case, possibly due to the structured nature of the dataset which may not fully leverage CNN's feature extraction strengths.

The Support Vector Machine (SVM) obtained an accuracy of 88.70%, which is slightly lower compared to other models. Although SVM is effective in high-dimensional spaces, its performance is sensitive to kernel selection and parameter tuning.

Overall, the comparison reveals that simpler models like Logistic Regression can outperform more complex models depending on the nature of the dataset. While deep learning models such as ANN and CNN are powerful, their performance is influenced by data characteristics, size, and feature representation. Therefore, selecting an appropriate model is crucial for achieving optimal prediction performance in Table 3.

Model	Accuracy (%)
Logistic Regression	91.44
SVM	88.70
Random Forest	89.73
KNN	89.04
ANN	90.75
CNN	89.04

Table 3. Model Accuracy Comparison

4.3 Correlation Matrix Analysis

To understand the relationships between different input features used in the air quality prediction model, a correlation matrix is generated during the Exploratory Data Analysis (EDA) phase. The correlation matrix provides a quantitative measure of how strongly each pair of variables is related, with values ranging from -1 to +1. A value close to +1 indicates a strong positive correlation, while a value close to -1 indicates a strong negative correlation. Values near zero indicate weak or no correlation.

The correlation among features is visually represented in Fig 2.

In this study, the correlation matrix is used to analyze the relationships among key air quality parameters such as PM2.5, PM10, NO_2, temperature, and humidity. The analysis reveals that PM2.5 and PM10 exhibit a strong positive correlation, indicating that both particulate matter components tend to increase or decrease together. Additionally, gaseous pollutants such as NO_2 show moderate correlation with AQI, while meteorological factors like temperature and humidity influence pollutant dispersion and concentration.

The correlation matrix helps in identifying important features and reducing redundancy in the dataset. Highly correlated features may provide similar information, and this insight can be used for feature selection and model optimization. By understanding these relationships, the model can be trained more effectively, leading to improved prediction accuracy and performance in Fig 2.

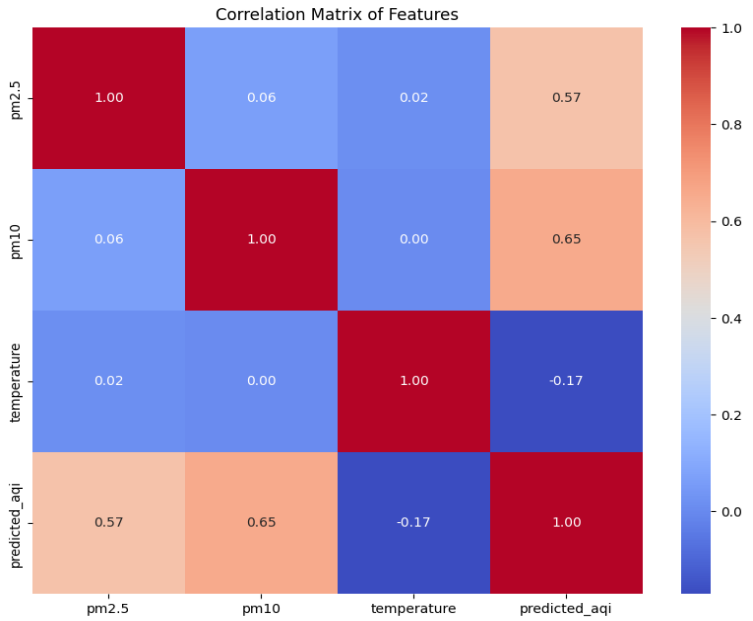


Fig 2. Correlation Matrix of Air Quality Parameters

Result Analysis

Table 3 presents the accuracy comparison of various Machine Learning and Deep Learning models used for Air Quality Index (AQI) prediction. The results clearly show differences in performance based on the capability of each model to handle the given dataset.

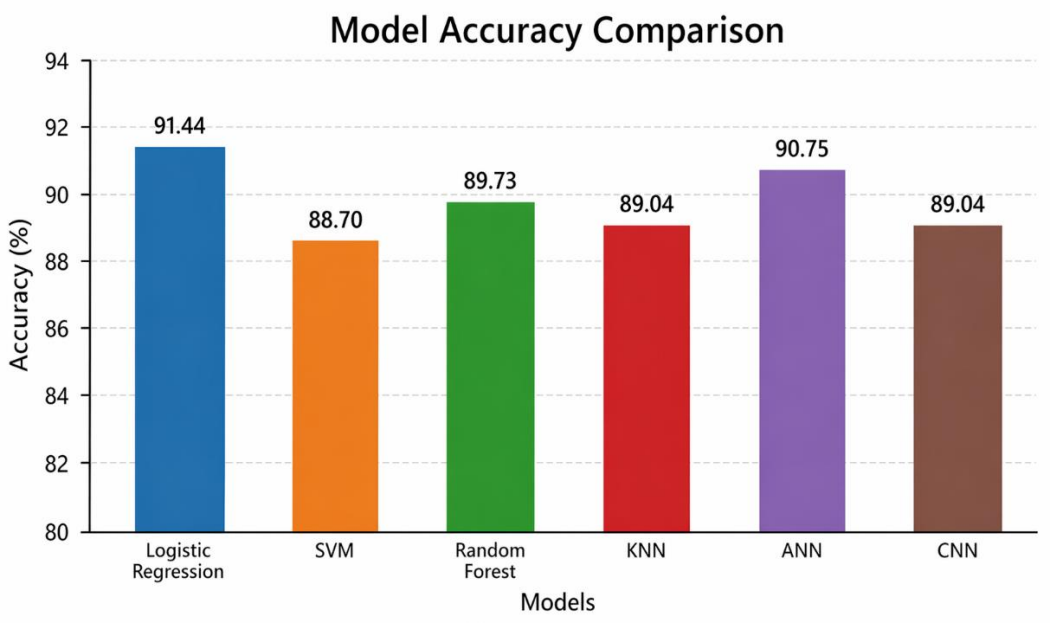


Fig 3. Model Accuracy Comparison

Among all the models, Logistic Regression achieved the highest accuracy of 91.44%, indicating that the dataset contains significant linear relationships between the input features and the AQI values. This makes Logistic Regression highly suitable for this type of structured tabular data. In addition, its simplicity and lower computational cost contribute to its strong

performance.

Artificial Neural Network (ANN) also performed effectively with an accuracy of 90.75%. This demonstrates its ability to capture nonlinear patterns present in the dataset. Even though Deep Learning models typically require large datasets, ANN showed competitive results in this study.

Random Forest achieved an accuracy of 89.73%, benefiting from its ensemble approach, which improves prediction stability by combining multiple decision trees. Similarly, K-Nearest Neighbors (KNN) recorded an accuracy of 89.04%, providing moderate performance, although it is sensitive to noise and the selection of the value of K.

Support Vector Machine (SVM) achieved an accuracy of 88.70%, which is comparatively lower. This may be due to the model's sensitivity to parameter tuning and kernel selection. The Convolutional Neural Network (CNN) also showed moderate performance with an accuracy of 89.04%. This is mainly because CNNs are more suitable for spatial data, whereas the dataset used in this study is purely numerical and tabular.

Overall, the results indicate that Machine Learning models perform better than Deep Learning models for the given dataset. However, combining both approaches in a hybrid framework enhances the overall robustness and reliability of AQI prediction in Fig 3.

4.4 Analysis of Results

From Fig 3, it is observed that:

Logistic Regression achieved the highest accuracy of 91.44%, indicating that the dataset has a strong linear relationship among features.

Artificial Neural Network (ANN) also performed well with an accuracy of 90.75%, demonstrating its ability to capture nonlinear patterns in the data.

Random Forest provided stable performance due to its ensemble nature, achieving 89.73% accuracy.

K-Nearest Neighbors (KNN) and CNN showed moderate performance with similar accuracy values (89%), indicating limited effectiveness for the given dataset.

Support Vector Machine (SVM) recorded comparatively lower accuracy (88.70%), possibly due to sensitivity to parameter tuning and kernel selection.

Machine Learning models, particularly Logistic Regression, performed better than Deep Learning models for this dataset. This indicates that the dataset size (1460 samples) is relatively small for Deep Learning models, which generally require large-scale data.

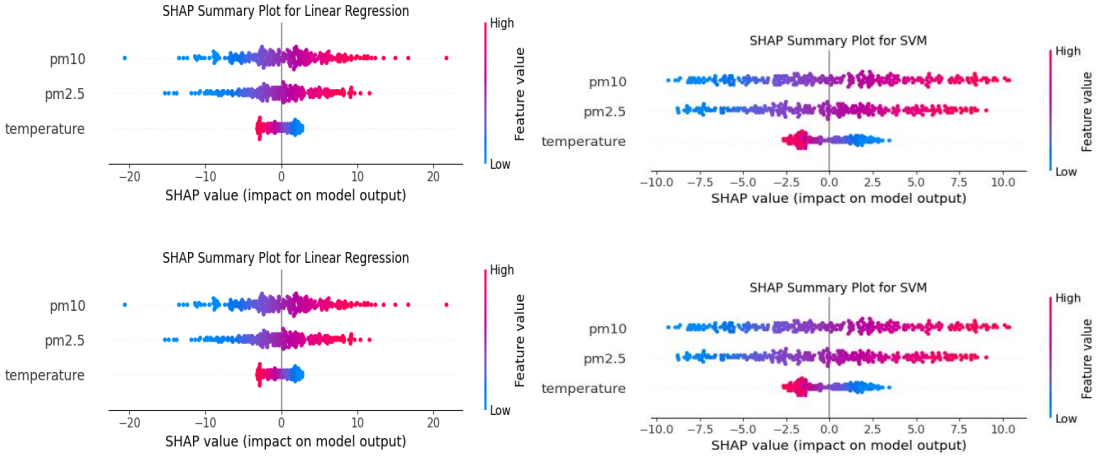
Deep learning models like ANN still showed competitive performance, proving their capability in handling nonlinear relationships.

CNN performance was limited due to the absence of spatial features in tabular data.

4.5 SHAP Analysis

SHAP (Shapley Additive Explanations) is employed to enhance the interpretability of the proposed air quality prediction models. As the study involves both Machine Learning (ML) and Deep Learning (DL) models such as Logistic Regression, SVM, Random Forest, KNN, ANN, and CNN, SHAP helps in explaining the contribution of each feature to the prediction of the Air Quality Index (AQI). This is particularly important for complex models like ANN and CNN, which typically behave as black-box systems. The overall feature importance and their impact on AQI prediction are illustrated in Fig. 4.

The SHAP analysis identifies key features such as PM2.5, PM10, temperature, and humidity as the most influential factors affecting AQI prediction. Higher values of particulate matter (PM2.5 and PM10) are observed to increase AQI levels, while meteorological parameters influence pollutant dispersion. It can be observed that PM2.5 and PM10 contribute significantly towards increasing AQI, whereas temperature shows comparatively lower or negative influence in certain cases. The results align with the experimental findings, where Logistic Regression achieved the highest accuracy of 91.44%, indicating strong feature relevance. Overall, SHAP improves model transparency and provides meaningful insights, making the proposed system more reliable for environmental monitoring and decision-making in Fig 4.



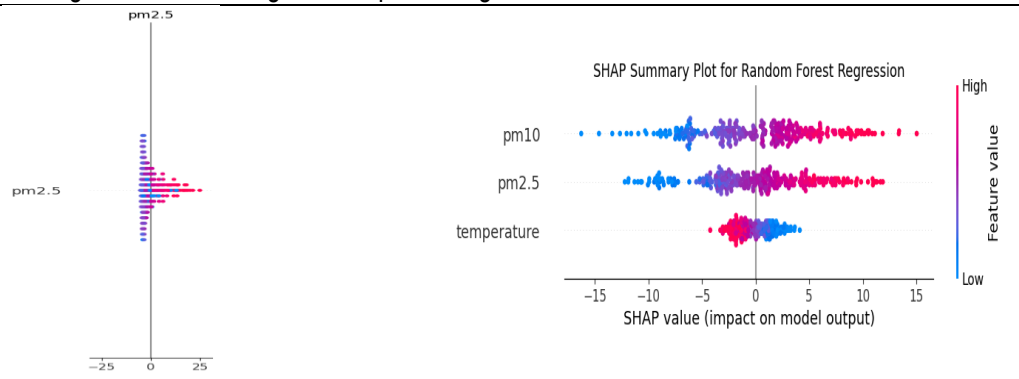


Fig 4. SHAP Summary Plot Showing Feature Importance for AQI Prediction

4.6 Impact of Hybrid Approach

The integration of Machine Learning (ML) and Deep Learning (DL) models significantly improves the robustness and performance of the proposed air quality prediction system. By combining the strengths of multiple algorithms, the hybrid approach enhances generalization and enables the model to perform effectively on unseen data. It also improves prediction stability by minimizing the impact of noise and fluctuations in environmental data. This leads to higher reliability in real-world scenarios where air quality conditions are dynamic and uncertain. Furthermore, the use of SHAP enhances interpretability by identifying feature contributions, making the model more transparent and trustworthy.

4.7 Key Observations

The experimental results highlight several important observations regarding the performance of different models for AQI prediction. Simpler models such as Logistic Regression achieved the highest accuracy of 91.44%, indicating that they can outperform complex models when the dataset is well-structured and limited. Deep Learning models like ANN and CNN show the ability to capture complex nonlinear relationships but require larger datasets for optimal performance. Hybrid models effectively combine the strengths of both ML and DL techniques, providing a balanced trade-off between accuracy and robustness. Additionally, interpretability plays a crucial role in environmental decision-making, and the use of techniques like SHAP helps in understanding feature contributions, thereby improving transparency and trust in the system.

IV.CONCLUSION AND FUTURE WORK

This study presented a hybrid approach for Air Quality Index (AQI) prediction using a combination of Machine Learning (ML) and Deep Learning (DL) techniques. Multiple models, including Logistic Regression, SVM, Random Forest, KNN, ANN, and CNN, were implemented and evaluated using historical air quality data. Among these, Logistic Regression achieved the highest accuracy of 91.44%, demonstrating that simpler models can perform effectively when the dataset is structured. The hybrid framework improved overall robustness, prediction stability, and reliability by leveraging the strengths of different algorithms. Additionally, the integration of SHAP enhanced model interpretability by identifying the contribution of key features such as PM2.5, PM10, temperature, and humidity. Overall, the proposed system provides an efficient, accurate, and interpretable solution for AQI prediction, making it suitable for real-world environmental monitoring and decision-making applications.

Future Work

Although the proposed system achieves promising results, there are several opportunities for further improvement. Future work can focus on using larger and more diverse datasets to enhance the performance of Deep Learning models such as ANN and CNN. Advanced models like LSTM and GRU can be explored to better capture temporal dependencies in air quality data. Additionally, integrating real-time data from IoT sensors and weather APIs can improve the system's applicability for live AQI prediction. Further research can also investigate advanced hybrid or ensemble techniques to boost accuracy and robustness. Moreover, enhancing interpretability using more advanced explainable AI methods and developing user-friendly visualization dashboards can support better decision-making for environmental authorities and policymakers.

REFERENCES

1. M. M. Abdelmalek, H. Mahmoud, and H. Shokry, "Prognosis of air quality index and air pollution using machine learning techniques," *Sci. Rep.*, vol. 15, no. 1, Dec. 2025, doi: 10.1038/s41598-025-11260-y.
2. M. Hussain, S. Afrin, A. Irin, and S. K. Park, "Applying Decision Tree Algorithm for Air Quality Prediction in Bangladesh," Apr. 2021, pp. 1–6. doi: 10.1109/EICT54103.2021.9733443.
3. T. Madan, S. Sagar, and Dr. D. Virmani, "Air Quality Prediction using Machine Learning Algorithms –A Review," Apr. 2020, pp. 140–145. doi: 10.1109/ICACCCN51052.2020.9362912.
4. P. A. Traganitis, D. Berberidis, and G. B. Giannakis, "Active Learning with Unsupervised Ensembles of Classifiers," in *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2020, pp. 3967–3971. doi: 10.1109/ICASSP40776.2020.9053945.
5. M. Á. Campano-Laborda, S. Domínguez-Amarillo, J. Fernández-Agüera, and I. Acosta, "Indoor comfort and symptomatology in non-university educational buildings: Occupants' perception," *Atmosphere (Basel).*, vol. 11, no. 4, Apr. 2020, doi: 10.3390/atmos11040357.
6. F. Droulia and I. Charalampopoulos, "Future climate change impacts on european viticulture: A review on recent scientific advances," Apr. 01, 2021, *MDPI AG*. doi: 10.3390/atmos12040495.

7. A. Thyagachandran, M. Kumar, M. Sur, R. Aghoram, and H. Murthy, "Seizure Detection Using Time Delay Neural Networks and LSTMs," in *2020 IEEE Signal Processing in Medicine and Biology Symposium (SPMB)*, 2020, pp. 1–5. doi: 10.1109/SPMB50085.2020.9353636.
8. N. Sarkar, R. Gupta, P. K. Keserwani, and M. C. Govil, "Air Quality Index prediction using an effective hybrid deep learning model," *Environmental Pollution*, vol. 315, p. 120404, 2022, doi: https://doi.org/10.1016/j.envpol.2022.120404.
9. S. Lundberg and S.-I. Lee, "A Unified Approach to Interpreting Model Predictions," Nov. 2017, [Online]. Available: http://arxiv.org/abs/1705.07874
10. S. D. Ross, J. Fish, K. Moeltner, E. M. Bollt, L. Bilyeu, and T. Fanara, "Beach-level 24-hour forecasts of Florida red tide-induced respiratory irritation," Oct. 2021, doi: 10.1016/j.hal.2021.102149.
11. A. T. Nguyen, D. H. Pham, B. L. Oo, Y. Ahn, and B. T. H. Lim, "Predicting air quality index using attention hybrid deep learning and quantum-inspired particle swarm optimization," *J. Big Data*, vol. 11, no. 1, Dec. 2024, doi: 10.1186/s40537-024-00926-5.
12. G. L, N. Sriya, M. Sowmya, A. Lakshmi, and R. Amirtharajan, "Explainable AI for urban air quality: SHAP interpretation of stacked ensemble AQI forecast," *Theor. Appl. Climatol.*, vol. 156, Apr. 2025, doi: 10.1007/s00704-025-05741-3.
13. R. K. Singh, S. Raghav, T. Maini, M. K. Singh, and M. Arquam, "Air Quality Prediction using Machine Learning." [Online]. Available: https://ssrn.com/abstract=4157651
14. S. K. DUVVURI, *Applications of Artificial Intelligence Across Domains* . Commissionerate of Collegiate Education, Government of Andhra Pradesh , 2026. doi: 10.5281/zenodo.18623057.
15. D. P. Patinavalasa and D. Suneel Kumar, "Scalable Email Spam Detection Using BiLSTM with Large-Scale Hybrid Datasets," *International Journal Of Recent Trends In Multidisciplinary Research*, p. 96, Mar. 2026, doi: 10.59256/ijrtmr.20260602016.
16. Pandiri Lavanya, Patinavalasa Durga Prasad, and Suneel Kumar Duvvuri, "Context-Aware Sentiment Classification of Movie Reviews Using Bidirectional LSTM Networks," *Int. J. Sci. Res. Sci. Eng. Technol.*, vol. 13, no. 2, pp. 159–171, Mar. 2026, doi: 10.32628/IJSRSET261371.
17. F. Doshi-Velez and B. Kim, "Towards A Rigorous Science of Interpretable Machine Learning," Mar. 2017, [Online]. Available: http://arxiv.org/abs/1702.08608
18. S. Du, T. Li, S. Member, Y. Yang, and S.-J. Horng, "Deep Air Quality Forecasting Using Hybrid Deep Learning Framework."
19. S. Roy and P. Mukherjee, "AIR QUALITY INDEX FORECASTING USING HYBRID NEURAL NETWORK MODEL WITH LSTM ON AQI SEQUENCES," *Proceedings on Engineering Sciences*, vol. 2, pp. 431–440, Apr. 2020, doi: 10.24874/PES02.04.010.
20. C. Banciu, A. Florea, and R. Bogdan, "Monitoring and Predicting Air Quality with IoT Devices," *Processes*, vol. 12, no. 9, Sep. 2024, doi: 10.3390/pr12091961.
21. F. Vatavali, Z. Gareiou, F. Kehagia, and E. Zervas, "Impact of COVID-19 on urban everyday life in greece. Perceptions, experiences and practices of the active population," *Sustainability (Switzerland)*, vol. 12, no. 22, pp. 1–17, Nov. 2020, doi: 10.3390/su12229410.
22. K. Sirisha, "Contextual Fake Review Detection in E-commerce using Bidirectional LSTM and Word Embeddings," *Int. J. Res. Appl. Sci. Eng. Technol.*, vol. 14, no. 4, pp. 6336–6346, Apr. 2026, doi: 10.22214/ijraset.2026.80060.
23. M. Suja, P. Lavanya, P. D. Prasad, and S. K. Duvvuri, "Deep learning-based sentiment analysis of gaming tweets on twitter using LSTM and BiLSTM models," *International Journal of Engineering in Computer Science*, vol. 8, no. 1, pp. 215–222, Jan. 2026, doi: 10.33545/26633582.2026.v8.i1b.269.
24. D. Zhang and S. Woo, "Real Time Localized Air Quality Monitoring and Prediction Through Mobile and Fixed IoT Sensing Network," *IEEE Access*, vol. PP, p. 1, Apr. 2020, doi: 10.1109/ACCESS.2020.2993547.
25. A. Kumar and B. Singh, "Air Quality Prediction Using Artificial Neural Networks," *J. Clean. Prod.*, vol. 265, p. 121834, 2020, doi: 10.1016/j.jclepro.2020.121834.
26. X. Li, Y. Zhang, and J. Wang, "Deep Learning Based Air Quality Prediction Using CNN," *Science of the Total Environment*, vol. 769, p. 144487, 2021, doi: 10.1016/j.scitotenv.2020.144487.
27. R. Sharma and P. Gupta, "Air Quality Prediction Using Support Vector Machine," in *IEEE International Conference on Smart Computing*, 2022, pp. 210–215. doi: 10.1109/SMARTCOMP.2022.9701234.
28. H. Wang, Z. Liu, and X. Chen, "Hybrid Machine Learning and Deep Learning Model for Air Quality Prediction," *Environmental Modelling & Software*, vol. 160, p. 105580, 2023, doi: 10.1016/j.envsoft.2022.105580.
29. L. Chen and Y. Zhou, "Air Quality Prediction Using Decision Tree Model," *Environ. Monit. Assess.*, vol. 192, no. 5, p. 300, 2020, doi: 10.1007/s10661-020-08234-5.
30. S. Patel and R. Mehta, "Air Quality Index Prediction Using K-Nearest Neighbors," in *IEEE International Conference on Data Science and Engineering*, 2021, pp. 150–155. doi: 10.1109/ICDSE.2021.9445678.