The University of Tehran Press

Pollution

Print ISSN: 2383-451X
Online ISSN: 2383-4501

https://jpoll.ut.ac.ir/

A Hybrid Machine Learning Model Based on Deep Learning for Air Quality Prediction

Mohammad Reza Mehregan[1] | Mohammad Taghi Taghavifard[2] | Amir Mohammad Khani[1] | Arman Rezasoltani[1]✉ | Mohammad Ali Nikkhah[1]

1. Department of Industrial Management, Faculty of Industrial Management and Technology, College of Management, University of Tehran, Tehran, Iran.
2. Department of Information Technology and Operations Management, Faculty of Management and Accounting, Allameh Tabataba'i University, Tehran, Iran.

Article Info

Article type:
Research Article

Article history:
Received: 18 January 2025
Revised: 29 May 2025
Accepted: 24 August 2025

Keywords:
Air Quality
Deep Learning
Ensemble Learning
Environmental Monitoring
Hybrid Model
Hyperparameter Optimization

ABSTRACT

Air pollution is a major global challenge, significantly and directly affecting public health, urban sustainability, and environmental policy. Accurate air quality prediction has increasingly become essential to address the challenges posed by environmental adversities. This study proposes a novel hybrid machine learning model that combines deep learning and advanced ensemble techniques to improve air quality prediction. This model combines Deep Neural Network (DNN), along with ensemble learning algorithms such as XGBoost, CatBoost, LightGBM, and Random Forest as a metamodel to aggregate the predictions. The model was tested on a dataset that included environmental aspects ranging from PM2.5, PM10, CO, and NO2 variables to socio-economic variables such as proximity to industrial areas and population density. Feature selection and data imbalance were handled using RFECV and SMOTE, respectively. The tuning of the hyperparameters in the model was done using both TPE implemented by Optuna and Bayesian optimization by Keras-Tuner. This model can achieve a remarkable accuracy of 97.34%, which is superior to conventional approaches. The results present a case for building hybrid machine learning techniques for air quality prediction as a basis for intelligent global environmental monitoring in an interpretable, accurate, and scalable manner. Future work can integrate the real-time incoming data from the Internet of Things (IOT) and extend the model concept for multi-prediction benchmarks to other environmental indices, thus broadening its horizon and applicability to upcoming global environmental challenges.

Cite this article: Mehregan, M. R., Taghavifard, M., T., Khani, A., M., Rezasoltani, A., & Nikkhah, M., A. (2025). A Hybrid Machine Learning Model Based on Deep Learning for Air Quality Prediction. *Pollution*, 11(4), 1199-1215. https://doi.org/10.22059/poll.2025.388743.2750

 Publisher: The University of Tehran Press.
DOI: https://doi.org/10.22059/poll.2025.388743.2750

INTRODUCTION

The increasing impacts of air pollution on human health, the environment, and global economies have made air quality prediction a critical research area. Air pollution has been forefront identified as a leading cause of respiratory and cardiovascular diseases, where particulate matter and certain harmful gases, including nitrogen dioxide (NO_2), sulfur dioxide (SO_2), and carbon monoxide (CO), play the first role (Mao et al., 2024; Wonderling et al., 2024). Machine learning and deep learning have enabled incorporating several prediction modeling techniques into monitoring and forecasting mechanisms. Usually, statistical models such as linear regression and autoregressive integrated moving average (ARIMA) were frequently applied in air quality forecasting. However, hybrid and deep learning apps are emerging as a new trend of study due to their incapacity to handle very complex and highly non-linear

*Corresponding Author Email: *rmanrezasoltani@ut.ac.ir*

relationships in environmental data (Ramadan et al., 2024; Sun et al., 2021; X. Wang et al., 2024). The accuracy and reliability of air quality prediction models have increased recently due to the latest advancements in hybrid machine learning techniques (Natarajan et al., 2024; Shankar & Arasu, 2023). Most of this focuses on learning deep learning architectures like convolutional neural networks (CNNs) and long short-term memories (LSTMs) and using ensemble learning methods like eXtreme Gradient Boosting (XGBoost) and Categorical Boosting (CatBoost). LSTMs are very good at processing temporal sequences, and CNNs are great at capturing spatial dependency and feature extraction from complex environmental datasets (Li et al., 2023). The prediction robustness and generalizability are improved, as powerful ensemble learning techniques combined with these approaches have allowed researchers to address the non-linearity and high dimensionality of air pollution data (Dong et al., 2024; Zhao & Ye, 2024). In terms of aggregating weak learners and reducing errors by iterative refinement processes, ensemble learning techniques like XGBoost and CatBoost are helpful contributions. These models reduce bias and variance, thereby enhancing predictive accuracy when processing structured environmental data (Ghosh et al., 2023). Moreover, the advanced optimization strategies are proposed, i.e., Bayesian hyperparameter tuning and feature selection techniques such as Recursive Feature Elimination with cross-validation for identifying the most contributing input variables (Awad & Fraihat, 2023). Furthermore, synthetic data augmentation methods, such as the Synthetic Minority Over-sampling Technique (SMOTE), integrate well to mitigate data imbalance, ensuring the model can effectively learn from the underrepresented pollution scenarios. The hybrid approaches introduced in this work are especially deemed useful in air quality prediction as they allow to model changes in pollutant dispersion caused by changing the processes of pollution source dispersion by different meteorological conditions, industrial emissions, and urban population density (Hettige et al., 2024; Hu et al., 2023). This synergy, therefore, offers an overall, scalable, and interpretable framework for air quality monitoring that takes advantage of both deep learning and ensemble learning techniques.

While the field has come a long way, today's prediction models on air quality are still fraught with several challenges. One of them is that most of these models rely on considerably limited and regional datasets - probably making generalization much more difficult concerning the different geographies (Hettige et al., 2024; Petrić et al., 2024). As previously mentioned, traditional models are very accurate in their results but involve significant complexity in computing, meaning that they cannot be used for real-time applications or edge devices. Also, most existing models are tailored towards standard pollutants like PM2.5 and PM10, while other - socioeconomic and environmental - factors that can contribute to a deep understanding of air quality dynamics are omitted. Additionally, evaluation metrics in quite a good number of studies are done around some basic measures like accuracy and RMSE, thus failing to capture a wide range of forecasting conditions, including short-term and long-term trends (Agbehadji & Obagbuwa, 2024; Khamlich et al., 2023). The high-impact consequence of air quality prediction would be on public health, urban sustainability, and environmental policy (Jayaraman & Abirami, 2025). Accurate and timely forecasting would enable governments and society to put in place proactive intervention against pollution risk (Hu et al., 2023) Moreover, integrating multi-source datasets into high-order machine learning techniques improves the dependability of the prediction and reveals the hidden influences on the air quality (Y. Wang et al., 2024). Efficient and interpretable hybrid models target earlier techniques' computational and generalizability problems, paving the way for scale-it solutions to satisfy real-time and geographical requirements (Rudin, 2019).

This research proposes a novel hybrid machine learning model for air quality prediction to increase accuracy and generalizability by integrating deep learning and ensemble learning techniques. This research proposes a novel hybrid machine learning model for air quality prediction to increase accuracy and generalizability by integrating deep learning and ensemble

learning techniques. This research brings forth one of the most promising innovations in the stacking-based hybrid model development, combining deep neural networks (DNN) and advanced ensemble learning algorithms like XGBoost, CatBoost, and LightGBM. This model combines deep learning features, capable of learning complex relationships, and ensemble learning to reduce prediction error in environmental data analysis. This study enhances air quality prediction using RFECV for feature selection, Bayesian optimization (Optuna, Keras Tuner) for tuning, and SMOTE to handle data imbalance. Unlike previous works, the current research includes socioeconomic and environmental factors, aiming for a scalable and IoT-compatible hybrid model with high accuracy and efficiency.

Theoretical Foundations

Air Quality Prediction

The atmospheric quality prediction uses computational techniques that model and forecast the environmental data trends. This is mainly a complicated scenario due to the multiplicity of factors that impinge on or result in the concentrations of pollutants such as PM2.5 and PM10, conditional phenomena such as temperatures and humidities, and socioeconomic factors. Therefore, advanced machine learning or deep learning frameworks have been used to capture the multivariate non-linear relationships and high-dimensional patterns hidden in the above data sets (Agbehadji & Obagbuwa, 2024; Shankar & Arasu, 2023). Historical statistical methods for air quality forecasting have included linear regression and autoregressive integrated moving average (ARIMA). These models can be satisfactory for smaller, linear datasets, but most prove inadequate when used to register large, non-linear, and multi-sourced data. Furthermore, static models could never adapt to the speedily changing environment, limiting their predictive accuracy (Chaturvedi, 2024).

Advancements such as inherited algorithms have greatly improved the accuracy of air quality forecasting since machine learning and deep learning techniques were introduced. For example, Support Vector Machines (SVM) and Random Forests are well-suited to be applied to the handling of structured environmental data, while Convolutional Neural Networks and Long Short-Term Memory network work together to extract both spatial and temporal data for a better estimate of pollutant concentrations (Ma et al., 2023). Additional developments in predictive power have been made through hybrid models with strength combinations of such algorithms. It has been shown that combining CNNs for spatial pattern extraction with LSTMs for temporal sequence analysis outperforms any standalone model (Gilik et al., 2022; Li et al., 2023). Integrating various socioeconomic variables (e.g., proximity to industrial zones or population density) and real-time data streams has gained traction for better generalization of models. Edge-compatible models in IoT-based systems offer real-time air quality monitoring and actionable insights. Future developments in hardware, cloud computing, and ensemble learning may offer better impetus for newer innovations in the field (Sharifi et al., 2024). This lucid introduction provides the springboard on which an innovative hybrid machine learning model stands for the challenges and enhancements in air quality forecast.

Machine Learning Methods in Air Quality Prediction

Machine Learning has transformed air quality predictions by allowing for complex relationship and pattern identification within high-dimensional environmental data (Rahman et al., 2024). Among the elementary techniques, decision trees and random forests are often applied for their simple structure and ability to classify a dataset based on feature importance (Scornet, 2023). While decision trees tend to overfit in most instances, random forests solve it by combining the outputs of several trees, thus increasing accuracy and robustness, especially when missing data is at play (Beaulac & Rosenthal, 2020). Support Vector Machines (SVM) are another potent tool for classification and regression, proving remarkably useful for structured

datasets. They find a hyperplane that separates the data points with the maximum possible margin , thus excelling in tasks such as pollutant threshold detection with a very high precision (Djeziri et al., 2022) Similarly, Gradient boosting machines, including advanced variants such as XGBoost, CatBoost, and LightGBM, are becoming popular for their efficiency in structured data. These models minimize the error of weak learners iteratively (Bentéjac et al., 2021; Jafarnejad Chaghoshi et al., 2024). These models iteratively minimize the error of weak learners. LightGBM has excellent speed and accuracy in large air-quality data sets because it grows trees leaf-wise.

Neural networks have further advanced predictive ability by mimicking the human brain's interconnected layers of neurons, thus being most adaptive for difficult, non-linear problems. DNNs are proficient in extracting, through Multidimensional air quality measurements, any non-linear pattern from the data, while CNNs are tailored for feature extraction from the spatial distribution of pollutants. For sequential data, RNNs and their advanced variant, LSTMs, are more efficient in analyzing time-series data related to pollutant levels and meteorological conditions (Gilik et al., 2022; Sun et al., 2021). Hybrid models generated a compelling approach by combining the powerful features of multiple algorithms to enhance their prediction performance. For example, CNNs can capture spatial features while LSTMs can work on temporal sequences, making for a perfect complement regarding the spatial-temporal modeling. Furthermore, stacking ensemble models would refer to having the output of algorithms like XGBoost, CatBoost, or Random Forest integrated using a meta-model, thus improving considerably the reliability and generalizability of the prediction (Tsokov et al., 2022).

Feature selection and optimization help build better machine learning models. Techniques such as Recursive Feature Elimination with Cross-Validation (RFECV) make the model simpler by addressing the varied importance of the variables to reduce redundancy while preserving the predictive power. Optimizers such as Optuna and Bayesian hyperparameter tuning will help have motors working better by working systematically to find the best stride for sometimes hard datasets (Awad & Fraihat, 2023).

Research background

Table 1 presents an investigation into various studies on air quality prediction, detailing the issue of authorship, research objective, models applied, datasets, findings, and performances as measured in terms of accuracy and RMSE. The table captures the latest developments in different deep learning architectures, hybrid models, and optimization methods toward accurate predictions and improvements in air quality indices by reducing error margins.

While previous studies have advanced air quality prediction, they often rely on limited datasets, complex models unsuitable for real-time use, and a priori evaluation criteria. To address these gaps, we developed a novel hybrid model combining DNN, XGBoost, CatBoost, and LightGBM with a Random Forest meta-model to increase stability and accuracy. Key innovations include using multi-source data (e.g., environmental, socioeconomic), hyperparameter tuning via Optuna and Keras-Tuner, feature selection with RFECV, and data balancing using SMOTE. Comprehensive evaluation criteria confirmed the superior performance and generalizability of our model.

Research Methodology

A Stacking-based Deep Ensemble Machine Learning Model was designed as the main feature of this study. The model is made up of a deep neural network along with other advanced ensemble learning models. The DNN is designed to extract complex and nonlinear relationships between features and is especially good at learning nonlinear patterns contained in data. The advanced methods in ensembles include XGBoost, CatBoost, and LightGBM, which were chosen for their high ability on structured data. A Random Forest classifier is then used as the metamodel

Table 1. Research background

Authors	Article title	Goals	Model used	Dataset	Conclusion	Accuracy /Precision
Du et al. (2019)	Sample-Evaluation-Enhanced Machine Learning Approach for Fault Diagnosis of Hybrid Systems	Forecasting PM2.5 air quality with high accuracy	1D-CNN + Bi-LSTM	Real-world datasets (2 datasets evaluated)	Combined spatial-temporal feature learning significantly improves PM2.5 prediction accuracy.	High prediction accuracy
Chang et al. (2020)	An Ensemble Learning Based Hybrid Model for Air Pollution Forecasting	Combine multiple forecasting methods to improve performance	Stacked ensemble learning	Spark + TensorFlow frameworks	Improved performance compared to individual models for hourly air pollution forecasting.	Outperformed GBTR, LSTM models
Wardana et al. (2021)	Optimising Deep Learning at the Edge for Accurate Hourly Air Quality Prediction	Design an edge-compatible hybrid model for air quality prediction	1D-CNN + LSTM	8272 hourly samples	Optimized model for edge devices with lower latency and high accuracy.	RMSE, MAE reduced significantly
Bhanja and Das (2021)	A Hybrid Deep Learning Model for Air Quality Time Series Prediction	Enhance feature representation and temporal order for PM2.5 prediction	CNN + BiLSTM	Air quality time-series data	Hybrid framework outperforms state-of-the-art models in prediction accuracy.	Best accuracy compared to baseline
Gilik et al. (2022)	Air Quality Prediction Using CNN+LSTM Hybrid Architecture	Predict pollutants like ozone, nitrogen oxides with spatial-temporal relationships	CNN + LSTM	Public air quality data from multiple cities	CNN-LSTM improves prediction by 11-53% over traditional LSTMs for various pollutants.	RMSE reduced 11-53%
Zhang et al. (2021)	A hybrid deep learning technology for PM2.5 air quality forecasting	Address PM2.5 volatility using frequency-domain decomposition	VMD + BiLSTM	Chinese city datasets	Hybrid model shows improved stability and forecasting performance over EMD-based models.	RMSE lower than EMD-based methods
Mengara Mengara et al. (2022)	Attention-Based Distributed Deep Learning Model for Air Quality Forecasting	Leverage attention-based BiLSTM for PM2.5 and PM10 prediction	Attention-based CNN + BiLSTM	South Korea traffic and air data	Attention mechanisms boost model accuracy; high performance across short/long-term predictions.	MAE: 5.02 (short-term); 22.59 (long)
Quynh et al. (2023)	Enhancing Air Quality Prediction Accuracy Using Hybrid Deep Learning	Predict PM2.5 and PM10 levels with hybrid models	Encoder-STM, BiLSTM	Hanoi air pollution dataset	Hybrid models with extended features improve prediction accuracy and error metrics.	MAE, RMSE low; significant accuracy
Xu et al. (2023)	A Hybrid Deep Learning Model for Air Quality Prediction Based on the Time–Frequency Domain Relationship	Predict PM2.5/PM10 using time-frequency data decomposition	Wavelet Transform + Transformer	Guilin air quality data (2018-2021)	Superior prediction performance compared to MLP, LSTM, and Transformer models.	Best across RMSE, MAE metrics
Rajagopal and Narayanan (2024)	A Novel Approach for Air Quality Index Prognostication	Forecast AQI using ensembled deep learning	CNN + BiLSTM + Autoencoder	Hybrid AQI datasets (various sources)	Hybrid optimization model outperforms traditional methods across all evaluation metrics.	R^2: 0.961; RMSE: 11.92; MAE: 10.29
Sigamani (2024)	Air quality index prediction with optimisation enabled deep learning model in IoT application	Develop IoT-based DL for air quality prediction	DFNN with TTSA	Time series data	DFNN optimized by TTSA provided superior RMSE (0.602), MSE, and other metrics.	RMSE: 0.602, R2: 0.598
Bhardwaj and Ragiri (2024)	A Deep Learning Approach to Enhance Air Quality Prediction: Comparative Analysis of LSTM, LSTM with Attention Mechanism and BiLSTM	Compare standard LSTM, Attention LSTM, BiLSTM for AQI prediction	LSTM variants	Comprehensive AQ data	Attention-based LSTM models excel in predicting AQI using 30-day sequences.	RMSE, MAE values show improved temporal modeling
Wang (2024)	Air Quality Prediction based on Neural Network	Introduce AirPhyNet integrating atmospheric physics for AQ forecasting	AirPhyNet	Beijing, Shenzhen datasets	Physical insights significantly improve accuracy; surpasses state-of-the-art deep learning models.	Better accuracy vs baselines
		Solve gradient	Quantum-		Quantum activation	Notable

Table 1. Research background

Authors	Article title	Goals	Model used	Dataset	Conclusion	Accuracy /Precision
Dong et al. (2024)	Quantum Optimized Hybrid Neural Network for AQ Prediction	issues in NN for AQ prediction	classical CNN	Various air quality datasets	optimized CNN showed superior prediction accuracy over conventional methods.	improvement in accuracy
Li and Dong (2024)	Quantum LSTM with Particle Swarm Optimization	Integrate QLSTM and ICEEMDAN for time-series AQ predictions	QLSTM with PSO optimization	Time-series AQ datasets	ICEEMDAN-QLSTM achieved highest predictive accuracy by reducing data complexity.	Enhanced accuracy using PSO
(Fathima et al., 2024)	Air Quality Prediction Using DL Models	Compare RNN, LSTM, GRU, BiLSTM for temporal air quality prediction	GRU, BiLSTM	Multiple pollutants datasets	GRU, LSTM captured temporal dependencies; GRU provided optimized predictions.	Performance metrics varied with architecture
Nguyen et al. (2024)	Hybrid Deep Learning for AQI Prediction	Design hybrid DL models integrating ARIMA, QPSO, CNN, and XGBoost	Attention-CNN, ARIMA, LSTM, XGBoost	Seoul AQ datasets	ARIMA-CNN-LSTM-XGBoost model excelled in accuracy with multi-station validation.	MSE reduced by 31.13%, R2 improved by 2%
Tejaswi (2024)	AIR MAP - DL for Smarter AQI Decisions	Develop a web-based system for air quality and weather forecasting	LSTM integrated with visual tools	Sensor-generated data	Real-time predictions with advanced visualizations enhance public and governmental decision-making.	Enhanced usability metrics

that combines the predictions of the base models and gives the final prediction output. This hybrid structure increases the accuracy and generalizability of the model. This model combines the predictive power of deep networks and the efficiency of advanced ensemble learning models to predict air quality levels accurately. This study has used Python programming language.

Study Area and Data

The data source used in this study is a dataset from the Kaggle website published by Mateen and includes information related to air quality and pollution (Mateen., 2024). This dataset is based on several real-world sources of air quality and related environmental factors, such as the World Health Organization (WHO) and World Bank Data. Table 2 presents the description of features in the air quality dataset.

Data Preprocessing

The data used in this study were completely clean and free of missing values, so there was no need to replace or manage missing values. StandardScaler was used to standardize the scale of numerical features. This process changed the numerical values so that their mean was zero and their standard deviation was one. Also, the target variable (Air Quality Levels), which was categorical (including the values “good”, “average”, “poor”, and “hazardous”), was converted to numerical (0, 1, 2, 3).

Figure 1 shows a correlation heatmap highlighting the relationships between environmental and socioeconomic factors affecting air quality (Kebriaeezadeh et al., 2022). To reduce redundancy from highly correlated features (such as PM2.5 and PM10), RFECV with random forest was used to select the most influential features through recursive elimination and cross-validation (Awad & Fraihat, 2023).

Figure 2 shows the feature importance scores from the random forest model, highlighting CO as the most influential factor in predicting air quality, followed by proximity to industrial areas (Hu et al., 2023). Using RFECV, eight key features—temperature, humidity, PM10, NO_2,

Table 2. Description of Features

Feature	Description	Unit	Data Type
Temperature	Average temperature of the region	Degrees Celsius (°C)	Numerical
Humidity	Relative humidity of the region	Percentage (%)	Numerical
PM2.5	Concentration of fine particulate matter less than 2.5 micrometers	μg/m³	Numerical
PM10	Concentration of particulate matter less than 10 micrometers	μg/m³	Numerical
NO_2	Nitrogen dioxide concentration	Parts per billion (ppb)	Numerical
SO_2	Sulfur dioxide concentration	Parts per billion (ppb)	Numerical
CO	Carbon monoxide concentration	Parts per million (ppm)	Numerical
Proximity to Industrial Areas	Distance to the nearest industrial area	Kilometers (km)	Numerical
Population Density	Number of people per square kilometer in the region	People per square kilometer	Numerical
Air Quality Levels (Target)	Air quality classification into Good, Moderate, Poor, and Hazardous	Good, Moderate, Poor, and Hazardous	Categorical

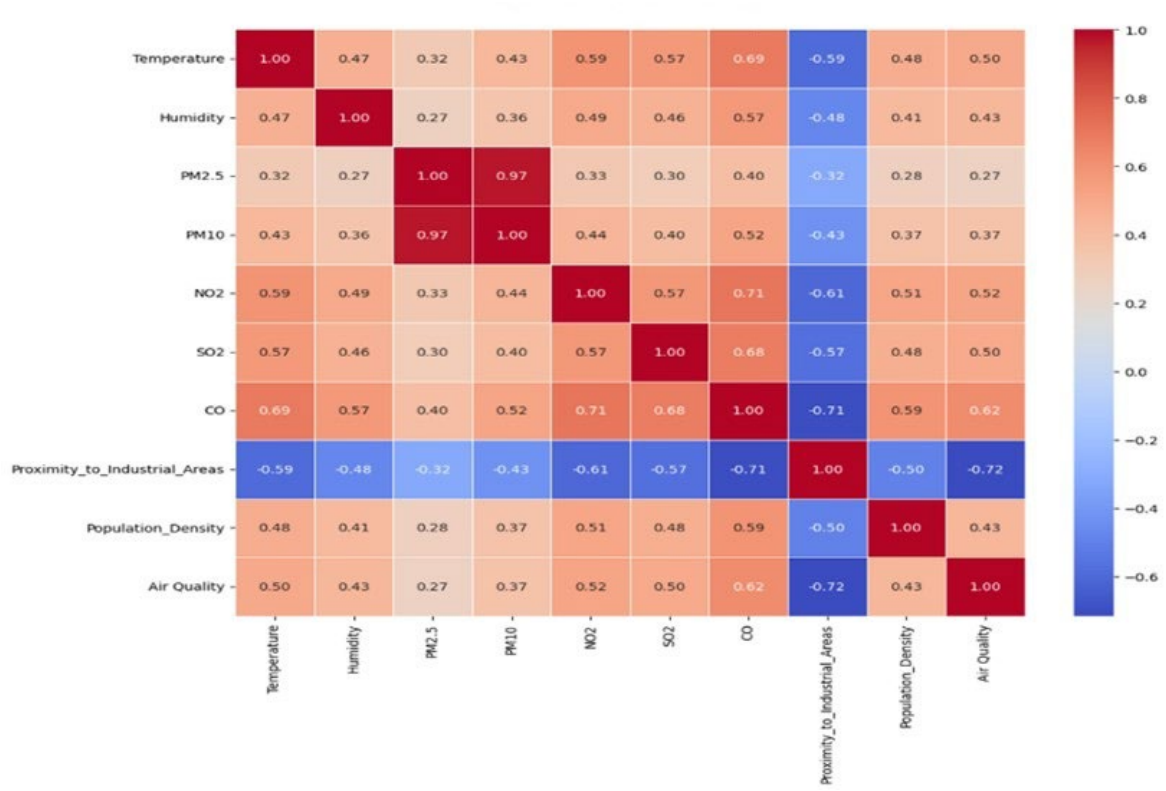

Fig. 1. Correlation heatmap of environmental variables, air quality and influencing factors

SO_2, CO, proximity to industrial areas, and population density—were selected for optimal model performance. As shown in Figure 3, the dataset is unbalanced, with the "good" class overrepresented and the "hazardous" class underrepresented, potentially affecting the model's accuracy in minority classes. The SMOTE method was used to deal with this problem. SMOTE is a synthetic data generation method for under-sampled classes (Zhao, 2025).

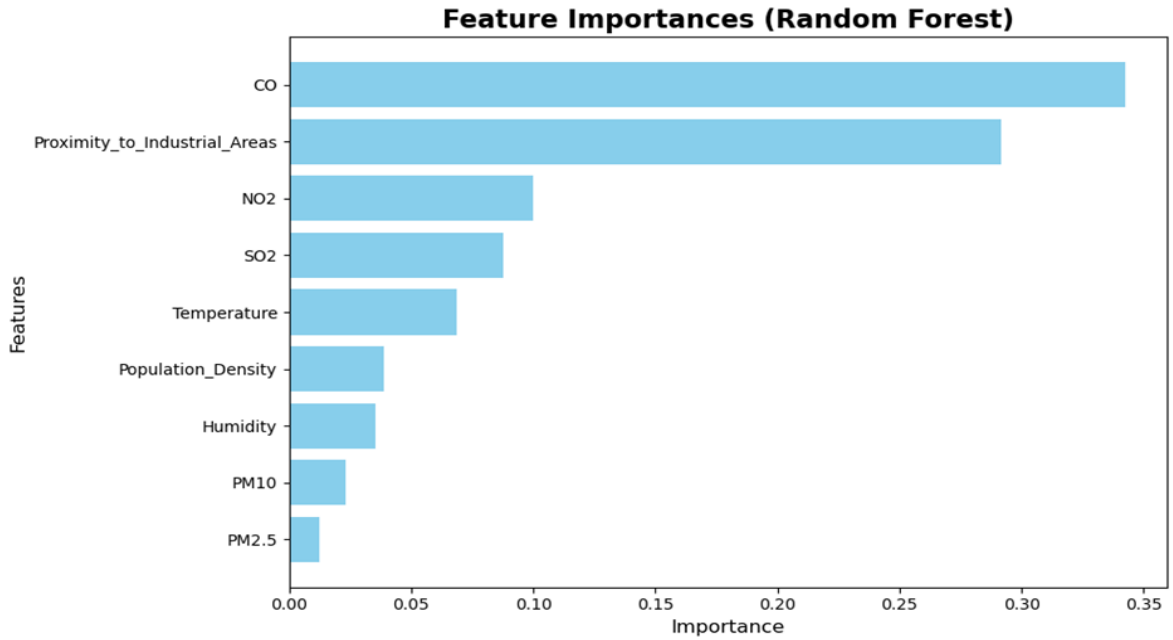

Fig. 2. Importance of features in air quality prediction using the random forest model

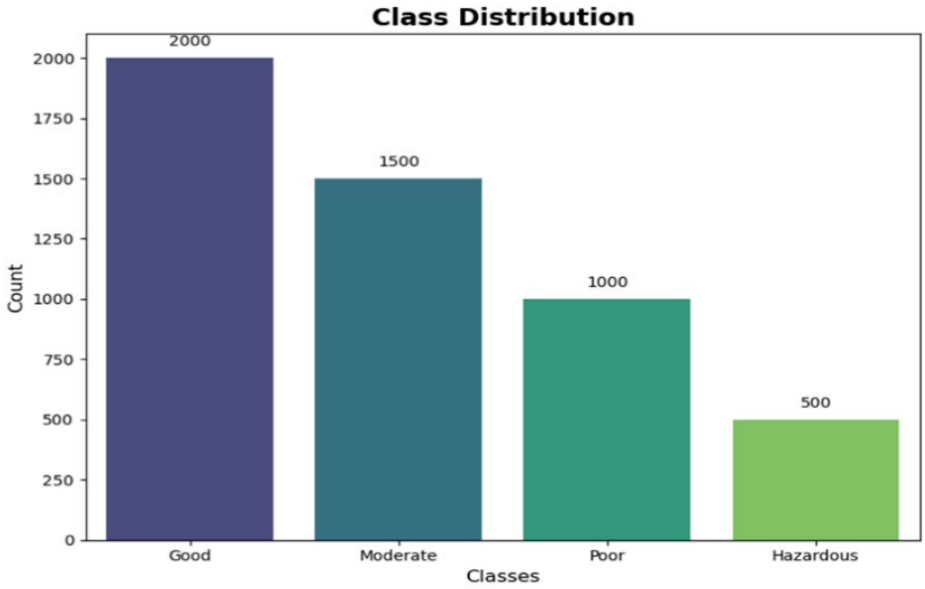

Fig. 3. Distribution of air quality classes in the dataset

Modeling

Stacking Ensemble Learning is one of the most powerful algorithms in machine learning that enhances accuracy and generalizability by combining the predictions of several base models (Dey & Mathur, 2023). In this approach, various models act as submodels (base models), and each creates independent predictions, serving as input to a higher-level model known as a Meta-Model (Nukui & Onogi, 2023). The meta-model then uses all these predictions to provide the final output. The basic intention behind stacking is to harness the strength of each base model in identifying certain patterns in the data and, thus, reduce individual errors. In this study, we constructed a Deep Ensemble Learning model combining a DNN model with ensemble

Table 3. Initial Settings for Hyperparameter Optimization

Parameter	Range or Values	Description
Number of Hidden Layers	3 to 6	The number of hidden layers to explore during optimization.
Neurons per Layer	64 to 512	The number of neurons in each layer, selected from this range.
Dropout Rate	0.1 to 0.5	The percentage of neurons randomly dropped for regularization.
Learning Rate	0.0001 to 0.01	Learning rate for the Adam optimizer.
Activation Function	ReLU	Non-linear activation function for hidden layers.
Epochs	50	The number of full passes through the dataset during training.
Batch Size	32	The number of samples processed before the model updates its weights.

Table 4. Specifications and optimal values of neural network architecture

Parameter	Optimized Value	Description
Number of Hidden Layers	5	The best number of hidden layers identified.
Neurons in Layer 1	128	Optimized number of neurons in the first hidden layer.
Neurons in Layer 2	448	Optimized number of neurons in the second hidden layer.
Neurons in Layer 3	448	Optimized number of neurons in the third hidden layer.
Neurons in Layer 4	256	Optimized number of neurons in the fourth hidden layer.
Neurons in Layer 5	192	Optimized number of neurons in the fifth hidden layer.
Dropout Rate (Layer 1)	0/4	Optimized dropout rate for the first hidden layer.
Dropout Rate (Layer 2)	0/3	Optimized dropout rate for the second hidden layer.
Dropout Rate (Layer 3)	0/3	Optimized dropout rate for the third hidden layer.
Dropout Rate (Layer 4)	0/1	Optimized dropout rate for the fourth hidden layer.
Dropout Rate (Layer 5)	0/4	Optimized dropout rate for the fifth hidden layer.
Learning Rate	0/001	Optimized learning rate for the Adam optimizer.

learning models. The model is built on a deep neural network tuned by Keras-Tuner and Bayesian optimization to capture complex and non-linear relationships between features. At the same time, XGBoost, CatBoost, and LightGBM ensemble learning models suitable for working with structured data were tuned using Optuna and Bayesian optimization. Meta-modeling was performed by stacking forecasts made by the models with the help of the Random Forest classifier. The deep ensemble learning proposed here, by harnessing the power of each model, enabled improved accuracy and reliability in air quality prediction.

Deep Neural Network

Deep neural networks are machine-learning models inspired by the human brain and are robust enough to extract highly complex and nonlinear relationships between data (Arifuzzaman et al., 2023). A DNN was developed, which works with input features and can establish complex and nonlinear relationships between them for air quality prediction (Liu et al., 2024). The model was built using Keras, and its parameters were tuned using Keras-Tuner and Bayesian optimization (Roy et al., 2023). While developing and optimizing the DNN, the ranges for parameters to carry out an optimal search were first listed (Chowdhury et al., 2022). The initial hyperparameter search space is outlined in Table 3. The purpose of fine-tuning all these settings was to create a sufficient search space into which the architecture could potentially fit.

A search domain is defined with these settings to identify the best combination of hyperparameters using Bayesian optimization in Keras-Tuner (Victoria & Maragatham, 2021). Table 4 summarizes the architecture and optimal hyperparameter values of the neural network.

Table 5. Optimal values of hyperparameters for ensemble learning models in the base

Model	n_ estimators	max_ depth	learning_ rate	subsample	colsample _bytree	iterations	depth	l2_leaf_reg	num_leaves
XGBoost	367	7	0/217084	0/731128	0/847117	-	-	-	-
CatBoost	-	-	0/299629	-	-	826	10	1/805096	-
LightGBM	276	0	0/060231	0/532938	0/960502	-	-	-	68

The proposed architecture consists of an input layer with 8 standard features, 5 optimized hidden layers, and an output layer with 4 neurons using Softmax for air quality classification. The model was tuned for neurons, dropout (0.1-0.4), and learning rate (0.001) using Adam optimizer and sparse cross-entropy loss. Early Stopping and Stratified K-fold cross-validation ensured robustness.

Other Base Models (XGBoost, CatBoost, LightGBM)

XGBoost (Extreme Gradient Boosting) is a boosting algorithm in machine learning based on decision trees that draws users' attention because of its efficiency and wide applicability (Mitchell & Frank, 2017). It sequentially boosts a weak decision tree model (Can et al., 2021). The XGBoost algorithm aims to minimize the model's overall error rate by increasing the importance of incorrectly predicted examples at each step (Bentéjac et al., 2021). This algorithm stands out for its extraordinary speed and capability of handling large amounts of structured data (Tang, 2024). The CatBoost algorithm is a variation of boosting with decision trees designed for handling data with any classification feature (Hancock & Khoshgoftaar, 2020). The Ordered Boosting algorithm is used to reduce the risk of overfitting and give a better level of performance than any other algorithm (Prokhorenkova et al., 2018). It automatically processes and encodes the classified data, relieving the need for manual preprocessing. This study has employed CatBoost to leverage the strength of error reduction with superior accurate predictions (Hancock & Khoshgoftaar, 2020). Due to its high speed, accuracy, and efficiency in handling large structured data using Leaf-Wise tree partitioning, LightGBM was chosen as the base model (Qiuqian et al., 2025). The optimized metaparameters of the ensemble models are shown in Table 5.

Random Forest as a Meta-Model

In this study, Random Forest was used as a meta-model due to its robustness, low underfitting, and high generalizability, effectively integrating DNN, XGBoost, CatBoost, and LightGBM predictions (Emeç & Yurtsever, 2025; Mirzadeh & Omranpour, 2024).

Comparison of the proposed model within the field of machine learning methods

To benchmark the performance evaluation of the proposed hybrid model, a comparison of its results against 12 popular machine learning algorithms is provided. Grid Search was used to search the set of optimal hyperparameter values. The algorithms that were used in this comparison are presented in Table 6.

Four key metrics in Table 7— accuracy, precision, recall, and F1 score—were calculated based on TP, TN, FP, and FN for each class to evaluate the model's performance in predicting air quality levels. These metrics provide a comprehensive view of the classification performance across classes. In addition, 5-fold cross-validation was used to ensure the model's generalizability and robustness (Ramadan et al., 2024).

Table 6. Comparison of methods and optimal parameters

Model	Description	References	Hyperparameters
Gradient Boosting	A combination of decision trees with gradual error updates to improve predictions.	(Saravani et al., 2025)	{'learning_rate': 0.2, 'max_depth': 7, 'n_estimators': 300}
AdaBoost	Boosting small decision trees by giving more weight to examples that are not classified correctly.	(Ding et al., 2022)	{'learning_rate': 1.0, 'n_estimators': 50}
Random Forest	A set of independent decision trees with random selection of features and samples to reduce overfitting.	(Kim, Kim, Mahdian, et al., 2024)	{'max_depth': 20, 'min_samples_split': 2, 'n_estimators': 200}
K-Nearest Neighbors	Predicting the new sample category based on the class of nearby samples in the feature space.	(Kramer, 2013)	{'n_neighbors': 3, 'weights': 'distance'}
Support Vector Machine	Data separation by finding hyperplane boundaries in feature space.	(Kim, Kim, Salamattalab, et al., 2024)	{'C': 10, 'kernel': 'rbf'}
Decision Tree	Classifying data based on the most information obtained in each division.	(Charbuty & Abdulazeez, 2021)	{'max_depth': 10, 'min_samples_split': 5}
Naive Bayes	Classification based on the conditional probability of features and the assumption of their independence.	(Hosein & Baboolal, 2024)	Default Parameters
Logistic Regression	A simple linear model for predicting category probabilities.	(Pal, 2021)	{'C': 1}
Ridge Classifier	A version of logistic regression that prevents overfitting by adding an L2 penalty.	(Hastie, 2020)	{'alpha': 0.1}
Linear Discriminant Analysis (LDA)	Data classification using linear combination of features and assuming Gaussian distribution of data.	(Zhao et al., 2024)	Default Parameters
Quadratic Discriminant Analysis (QDA)	Similar to LDA but using nonlinear boundaries to separate classes.	(Araveeporn, 2022)	Default Parameters
MLP	Multilayer neural network with the ability to learn nonlinear relationships.	(Noori et al., 2010)	{'hidden_layer_sizes': (100, 100), 'activation': 'tanh', 'solver': 'adam', 'learning_rate_init': 0.001}

Table 7. Evaluation metrics for machine learning models

Index	Definition	Formula
Accuracy	The ratio of correct predictions to the total number of samples.	$\frac{TP + TN}{TP + FP + FN + TN}$
Precision	The ratio of correctly predicted instances of a class to the total instances predicted as that class.	$\frac{TP}{TP + FP}$
Recall	The ratio of correctly predicted instances of a class to the total actual instances of that class.	$\frac{TP}{TP + FN}$
F1 Score	The harmonic mean of Precision and Recall to balance the trade-off between them.	$\frac{2 \times Precision \times Recall}{Precision + Recall}$

Table 8. Comparison of results

Model	Accuracy	Precision	Recall	F1 Score
Gradient Boosting	0/967625	0/967637	0/967625	0/967610
AdaBoost	0/714500	0/719707	0/714500	0/679580
Random Forest	0/965250	0/965251	0/965250	0/965228
K-Nearest Neighbors (KNN)	0/967125	0/967100	0/967125	0/966957
Support Vector Machine (SVM)	0/951125	0/951189	0/951125	0/951105
Decision Tree	0/926000	0/926882	0/926000	0/926248
Naive Bayes	0/926750	0/926644	0/926750	0/926635
Logistic Regression	0/928500	0/928194	0/928500	0/928283
Ridge Classifier	0/718625	0/748913	0/718625	0/669643
Linear Discriminant Analysis (LDA)	0/920625	0/920363	0/920625	0/920206
Quadratic Discriminant Analysis (QDA)	0/926625	0/926446	0/926625	0/926479
MLP	0/965250	0/965236	0/965250	0/965162
Proposed Deep Hybrid Model	**0.973375**	**0.972875**	**0.972875**	**0.972217**

RESULTS AND DISCUSSION

In this section, results obtained from the hybrid model have been analyzed and compared to other machine learning models. The results of model performance are presented in Table 8.

The study examined how various machine learning models performed in air quality level prediction and subsequently concluded that the suggested hybrid model (Proposed Deep Hybrid Model) excelled in all the evaluation metrics, including Accuracy, Precision, Recall, and F1 Score, for any comparison. The accuracy of the proposed hybrid model reached 0.973375, representing the peak among all models. The metric states that the model could predict the first maximum number of correct judgments for different air quality categories. The models closest to this were Gradient Boosting and KNN, offering values of 0.967625 and 0.967125, respectively, but unable to match the accuracy of the proposed model. The hybrid model again proved its worth in Precision by obtaining a value of 0.972875. This index is particularly critical for applications where reduced false positives are desirable. The Gradient Boosting and KNN models had accuracies of 0.967637 and 0.967100, respectively, indicating values less than the suggested model. The hybrid model had the highest recall index score at 0.972875. The index is significant for accurate air quality analyses by providing information about the model's ability to identify positive clusters correctly. Other trees, like the Random Forest and MLP, performed well, having a recall score of 0.965250 but still lagging behind the proposed hybrid model in this index. The F1 Score considered here indicates that the proposed hybrid model had the highest score of 0.972217 compared to the other models, indicating the balance of precision and recall for the model, further highlightingthe model's ability to provide predictions with a good balance between reducing false predictions and correctly identifying positive samples.

CONCLUSIONS

This study introduces a novel hybrid machine learning model that combines deep learning and ensemble techniques for air quality prediction, demonstrating exceptional prediction capabilities. The hybrid model beats other traditional and advanced machine learning models, on all four important evaluation metrics considered for this prediction. This study carries similar findings and extensions from prior studies, where CNN-LSTM models and those examined by Zhang

et al. (2021) showcased enhanced prediction accuracies (e.g., RMSE drop by 11–53%). In this study, the introduced hybrid model bears even added merit in terms of diminished error against most previous analyses. Unlike other studies that focused only on basic metrics (Bhanja & Das, 2021; Mengara Mengara et al., 2022), accuracy, precision, recall, and F1 score metrics have been introduced in this study aimed at getting a more nuanced view of the model's performance. Using advanced tools like Optuna and Bayesian tuning, the model has surpassed parameter optimization for other traditional models CNN-LSTM or BiLSTM in other studies like Gilik et al. (2022). The key advantages of the model include consideration of socioeconomic variables (for example, proximity to industrial zones and population density), widening the purview of prediction, feature selection through RFECV to ensure model simplicity and interpretability, and balancing data with SMOTE that will enable better predictions for underrepresented classes.

The hybrid model significantly enhances air quality prediction by ameliorating the shortcomings associated with traditional and deep learning approaches, including different data set integrations and complex optimization techniques to achieve better accuracy and robustness. The model's scalability makes it more suited to real-time air quality monitoring and decision-making. This model has a significant edge over studies conducted earlier. For example, Zhang et al. (2021) used a CNN-LSTM model for cities in China, improved the modeling of spatial and temporal aspects, but had limited consideration for socio-economic factors. Mengara Mengara et al. (2022) used an Attention-CNN-BiLSTM model for Korea with improved prediction but not with strong optimization of the relevant parameters. Gilik et al. (2022) implemented CNN-LSTM in multiple cities, realizing significant increases in accuracy but still facing computational intensity.

Future research is suggested to extend this prediction model by adding real-time data from Internet of Things (IOT) sensors to enhance the prediction accuracy and to improve the model applicability. In addition, transfer and federated learning can improve the model's generalizability across regions. Furthermore, integrating climate and industrial indicators, like temperature, humidity, and greenhouse gas emissions, into the model will ensure a more complete air quality analysis. As an optimization method, metaheuristic optimization techniques can decrease computational costs and increase the model's accuracy. Last, new hybrid learning methods and self-explanatory AI models will assist in developing more transparent and practical models for environmental decision-making.

GRANT SUPPORT DETAILS

The present research did not receive any financial support.

CONFLICT OF INTEREST

The authors declare that there is no conflict of interest regarding the publication of this manuscript. In addition, authors have completely observed the ethical issues, including plagiarism, informed consent, misconduct, data fabrication and/ or falsification, double publication and/or submission, and redundancy.

LIFE SCIENCE REPORTING

No life science threat was practiced in this research.

REFERENCES

Agbehadji, I. E., & Obagbuwa, I. C. (2024). Systematic Review of Machine Learning and Deep Learning

Techniques for Spatiotemporal Air Quality Prediction. *Atmosphere*, *15*(11), 1352. https://doi.org/https://doi.org/10.3390/atmos15111352
Araveeporn, A. (2022). Comparing the linear and quadratic discriminant analysis of diabetes disease classification based on data multicollinearity. *International Journal of Mathematics and Mathematical Sciences*, *2022*(1), 1-12. https://doi.org/https://doi.org/10.1155/2022/7829795
Arifuzzaman, M., Hasan, M. R., Toma, T. J., Hassan, S. B., & Paul, A. K. (2023). An advanced decision tree-based deep neural network in nonlinear data classification. *Technologies*, *11*(1), 1-24. https://doi.org/https://doi.org/10.3390/technologies11010024
Awad, M., & Fraihat, S. (2023). Recursive feature elimination with cross-validation with decision tree: Feature selection method for machine learning-based intrusion detection systems. *Journal of Sensor and Actuator Networks*, *12*(5), 67. https://doi.org/https://doi.org/10.3390/jsan12050067
Beaulac, C., & Rosenthal, J. S. (2020). BEST: A decision tree algorithm that handles missing values. *Computational Statistics*, *35*(3), 1001-1026. https://doi.org/https://doi.org/10.1007/s00180-020-00987-z
Bentéjac, C., Csörgő, A., & Martínez-Muñoz, G. (2021). A comparative analysis of gradient boosting algorithms. *Artificial Intelligence Review*, *54*, 1937-1967. https://doi.org/https://doi.org/10.1007/s10462-020-09896-5
Bhanja, S., & Das, A. (2021). A hybrid deep learning model for air quality time series prediction. *Indonesian Journal of Electrical Engineering and Computer Science*, *22*(3), 1611-1618. https://doi.org/https://doi.org/10.11591/ijeecs.v22.i3.pp1611-1618
Bhardwaj, D., & Ragiri, P. R. (2024). A Deep Learning Approach to Enhance Air Quality Prediction: Comparative Analysis of LSTM, LSTM with Attention Mechanism and BiLSTM. 2024 IEEE Region 10 Symposium (TENSYMP),
Can, R., Kocaman, S., & Gokceoglu, C. (2021). A comprehensive assessment of XGBoost algorithm for landslide susceptibility mapping in the upper basin of Ataturk dam, Turkey. *Applied Sciences*, *11*(11), 4993. https://doi.org/https://doi.org/10.3390/app11114993
Chang, Y.-S., Abimannan, S., Chiao, H.-T., Lin, C.-Y., & Huang, Y.-P. (2020). An ensemble learning based hybrid model and framework for air pollution forecasting. *Environmental Science and Pollution Research*, *27*, 38155-38168. https://doi.org/https://doi.org/10.1007/s11356-020-09855-1
Charbuty, B., & Abdulazeez, A. (2021). Classification based on decision tree algorithm for machine learning. *Journal of applied science and technology trends*, *2*(01), 20-28. https://doi.org/https://doi.org/10.38094/jastt20165
Chaturvedi, P. (2024). Air Quality Prediction System Using Machine Learning Models. *Water, Air, & Soil Pollution*, *235*(9), 578. https://doi.org/https://doi.org/10.1007/s11270-024-07390-0
Chowdhury, A. A., Das, A., Hoque, K. K. S., & Karmaker, D. (2022). A comparative study of hyperparameter optimization techniques for deep learning. Proceedings of International Joint Conference on Advances in Computational Intelligence: IJCACI 2021,
Dey, R., & Mathur, R. (2023). Ensemble learning method using stacking with base learner, a comparison. International Conference on Data Analytics and Insights,
Ding, Y., Zhu, H., Chen, R., & Li, R. (2022). An efficient AdaBoost algorithm with the multiple thresholds classification. *Applied Sciences*, *12*(12), 5872. https://doi.org/https://doi.org/10.3390/app12125872
Djeziri, M. A., Djedidi, O., Morati, N., Seguin, J.-L., Bendahan, M., & Contaret, T. (2022). A temporal-based SVM approach for the detection and identification of pollutant gases in a gas mixture. *Applied Intelligence*, *52*(6), 6065-6078. https://doi.org/https://doi.org/10.1007/s10489-021-02761-0
Dong, Y., Li, F., Zhu, T., & Yan, R. (2024). Air quality prediction based on quantum activation function optimized hybrid quantum classical neural network. *Frontiers in Physics*, *12*, 1412664. https://doi.org/https://doi.org/10.3389/fphy.2024.1412664
Du, S., Li, T., Yang, Y., & Horng, S.-J. (2019). Deep air quality forecasting using hybrid deep learning framework. *IEEE Transactions on Knowledge and Data Engineering*, *33*(6), 2412-2424. https://doi.org/https://doi.org/10.1109/tkde.2019.2954510
Emeç, M., & Yurtsever, M. (2025). A novel ensemble machine learning method for accurate air quality prediction. *International Journal of Environmental Science and Technology*, *22*(1), 459-476. https://doi.org/https://doi.org/10.1007/s13762-024-05671-z
Fathima, M. D., Donavalli, S., & Kambham, H. (2024). Air Quality Prediction using Deep Learning models. 2024 International Conference on Advancements in Power, Communication and Intelligent

Systems (APCI),
Ghosh, S., Gourisaria, M. K., Sahoo, B., & Das, H. (2023). A pragmatic ensemble learning approach for rainfall prediction. *Discover Internet of Things*, *3*(1), 13. https://doi.org/https://doi.org/10.1007/s43926-023-00044-3
Gilik, A., Ogrenci, A. S., & Ozmen, A. (2022). Air quality prediction using CNN+ LSTM-based hybrid deep learning architecture. *Environmental Science and Pollution Research*(29), 1-19. https://doi.org/https://doi.org/10.1007/s11356-021-16227-w
Hancock, J. T., & Khoshgoftaar, T. M. (2020). CatBoost for big data: an interdisciplinary review. *Journal of big data*, *7*(1), 94. https://doi.org/https://doi.org/10.1186/s40537-020-00369-8
Hastie, T. (2020). Ridge regularization: An essential concept in data science. *Technometrics*, *62*(4), 426-433. https://doi.org/https://doi.org/10.1080/00401706.2020.1791959
Hettige, K. H., Ji, J., Xiang, S., Long, C., Cong, G., & Wang, J. (2024). Airphynet: Harnessing physics-guided neural networks for air quality prediction. *arXiv preprint arXiv:2402.03784*, *2*, 1-16. https://doi.org/https://doi.org/10.48550/arxiv.2402.03784
Hosein, P., & Baboolal, K. (2024). Bayes Classification using an approximation to the Joint Probability Distribution of the Attributes. International Conference on Deep Learning Theory and Applications,
Hu, Y., Li, Q., Shi, X., Yan, J., & Chen, Y. (2023). Multi-spatial Multi-temporal Air Quality Forecasting with Integrated Monitoring and Reanalysis Data. *arXiv preprint arXiv:2401.00521*, *1*. https://doi.org/https://doi.org/10.48550/arxiv.2401.00521
Jafarnejad Chaghoshi, A., Rezasoltani, A., & Khani, A. M. (2024). Unleashing the Power of Ensemble Learning: Predicting National Ranks in Iran's University Entrance Examination. *Industrial Management Journal*, *16*(3), 457-481. https://doi.org/https://doi.org/10.22059/imj.2024.381521.1008178
Jayaraman, S., & Abirami, S. (2025). Enhancing urban air quality prediction using time-based-spatial forecasting framework. *Scientific Reports*, *15*(1), 4139. https://doi.org/https://doi.org/10.1038/s41598-024-83248-z
Kebriaeezadeh, S., Ghodduosi, J., Alesheikh, A. A., Arjmandi, R., & Mirzahosseini, S. A. (2022). Analyzing trend and factors affecting air quality in urban areas: a case study in Isfahan-metropolis, Iran. *Environmental Sciences*, *20*(2), 171-184.
Khamlich, M., Stabile, G., Rozza, G., Környei, L., & Horváth, Z. (2023). A physics-based reduced order model for urban air pollution prediction. *Computer Methods in Applied Mechanics and Engineering*, *417*, 116416. https://doi.org/https://doi.org/10.48550/arxiv.2305.04575
Kim, H. I., Kim, D., Mahdian, M., Salamattalab, M. M., Bateni, S. M., & Noori, R. (2024). Incorporation of water quality index models with machine learning-based techniques for real-time assessment of aquatic ecosystems. *Environmental Pollution*, *355*, 124242. https://doi.org/https://doi.org/10.1016/j.envpol.2024.124242
Kim, H. I., Kim, D., Salamattalab, M. M., Mahdian, M., Bateni, S. M., & Noori, R. (2024). Machine learning-based modeling of surface water temperature dynamics in arctic lakes. *Environmental Science and Pollution Research*, *31*(49), 59642-59655. https://doi.org/https://doi.org/10.1007/s11356-024-35173-x
Kramer, O. (2013). *Dimensionality reduction with unsupervised nearest neighbors* (Vol. 51). Springer. https://doi.org/https://doi.org/10.1007/978-3-642-38652-7_2
Li, F., & Dong, Y. (2024). Air quality prediction based on improved quantum long short-term memory neural networks. *Physica Scripta*, *99*(8), 085035. https://doi.org/https://doi.org/10.1088/1402-4896/ad619a
Li, Y., Jiang, T., Gu, H., Lu, W., Wu, Q., & Yu, Y. (2023). Air Quality Index Prediction Based on CNN-LSTM-Attention Hybrid Modeling. 2023 International Conference on the Cognitive Computing and Complex Data (ICCD),
Liu, H., Cheng, J., & Liao, W. (2024). Deep neural networks are adaptive to function regularity and data distribution in approximation and estimation. *arXiv preprint arXiv:2406.05320*, *1*. https://doi.org/https://doi.org/10.48550/arxiv.2406.05320
Ma, X., Chen, T., Ge, R., Xv, F., Cui, C., & Li, J. (2023). Prediction of PM2. 5 concentration using spatiotemporal data with machine learning models. *Atmosphere*, *14*(10), 1517. https://doi.org/https://doi.org/10.3390/atmos14101517
Mao, Q., Zhu, X., Zhang, X., & Kong, Y. (2024). Effect of air pollution on the global burden of cardiovascular diseases and forecasting future trends of the related metrics: a systematic analysis from the Global Burden of Disease Study 2021. *Frontiers in Medicine*, *11*, 1472996. https://doi.org/

https://doi.org/10.3389/fmed.2024.1472996
Mateen., M. (2024). *Air Quality and Pollution Assessment [Data set]* (https://doi.org/https://doi.org/10.34740/KAGGLE/DS/6197184
Mengara Mengara, A. G., Park, E., Jang, J., & Yoo, Y. (2022). Attention-based distributed deep learning model for air quality forecasting. *Sustainability*, *14*(6), 3269. https://doi.org/https://doi.org/10.3390/su14063269
Mirzadeh, H., & Omranpour, H. (2024). Extended Random Forest for multivariate air quality forecasting. *International Journal of Machine Learning and Cybernetics*, *16*, 1-25. https://doi.org/https://doi.org/10.1007/s13042-024-02329-7
Mitchell, R., & Frank, E. (2017). Accelerating the XGBoost algorithm using GPU computing. *PeerJ Computer Science*, *3*, e127. https://doi.org/https://doi.org/10.7717/peerj-cs.127
Natarajan, S. K., Shanmurthy, P., Arockiam, D., Balusamy, B., & Selvarajan, S. (2024). Optimized machine learning model for air quality index prediction in major cities in India. *Scientific Reports*, *14*(1), 6795. https://doi.org/https://doi.org/10.1038/s41598-024-54807-1
Nguyen, A. T., Pham, D. H., Oo, B. L., Ahn, Y., & Lim, B. T. (2024). Predicting air quality index using attention hybrid deep learning and quantum-inspired particle swarm optimization. *Journal of big data*, *11*(1), 71. https://doi.org/https://doi.org/10.1186/s40537-024-00926-5
Noori, R., Hoshyaripour, G., Ashrafi, K., & Araabi, B. N. (2010). Uncertainty analysis of developed ANN and ANFIS models in prediction of carbon monoxide daily concentration. *Atmospheric Environment*, *44*(4), 476-482. https://doi.org/https://doi.org/10.1016/j.atmosenv.2009.11.005
Nukui, T., & Onogi, A. (2023). An R package for ensemble learning stacking. *Bioinformatics Advances*, *3*(1), vbad139. https://doi.org/https://doi.org/10.1093/bioadv/vbad139
Pal, A. (2021). Logistic regression: A simple primer. *Cancer Research, Statistics, and Treatment*, *4*(3), 551-554. https://doi.org/https://doi.org/10.4103/crst.crst_164_21
Petrić, V., Hussain, H., Časni, K., Vuckovic, M., Schopper, A., Andrijić, Ž. U., Kecorius, S., Madueno, L., Kern, R., & Lovrić, M. (2024). Ensemble Machine Learning, Deep Learning, and Time Series Forecasting: Improving Prediction Accuracy for Hourly Concentrations of Ambient Air Pollutants. *Aerosol and Air Quality Research*, *24*(12), 230317. https://doi.org/https://doi.org/10.4209/aaqr.230317
Prokhorenkova, L., Gusev, G., Vorobev, A., Dorogush, A. V., & Gulin, A. (2018). CatBoost: unbiased boosting with categorical features. *Advances in neural information processing systems*, *31*, 1-11. https://doi.org/https://doi.org/10.48550/arxiv.1706.09516
Qiuqian, W., GaoMin, KeZhu, Z., & Chenchen. (2025). A light gradient boosting machine learning-based approach for predicting clinical data breast cancer. *Multiscale and Multidisciplinary Modeling, Experiments and Design*, *8*(1), 75. https://doi.org/https://doi.org/10.1007/s41939-024-00662-6
Quynh, T. P. T., Viet, T. N., Thi, H. D., & Manh, K. H. (2023). Enhancing air quality prediction accuracy using hybrid deep learning. *Int J Environ Sci Dev*, *14*(2), 155-159. https://doi.org/https://doi.org/10.18178/ijesd.2023.14.2.1428
Rahman, M. M., Nayeem, M. E. H., Ahmed, M. S., Tanha, K. A., Sakib, M. S. A., Uddin, K. M. M., & Babu, H. M. H. (2024). AirNet: predictive machine learning model for air quality forecasting using web interface. *Environmental Systems Research*, *13*(1), 44. https://doi.org/https://doi.org/10.1186/s40068-024-00378-z
Rajagopal, K., & Narayanan, K. (2024). A Novel Approach for Air Quality Index Prognostication using Hybrid Optimization Techniques. *International Research Journal of Multidisciplinary Technovation*, *6*(2), 84-99. https://doi.org/https://doi.org/10.54392/irjmt2427
Ramadan, M. S., Abuelgasim, A., & Al Hosani, N. (2024). Advancing air quality forecasting in Abu Dhabi, UAE using time series models. *Frontiers in Environmental Science*, *12*, 1393878. https://doi.org/https://doi.org/10.3389/fenvs.2024.1393878
Roy, S., Mehera, R., Pal, R. K., & Bandyopadhyay, S. K. (2023). Hyperparameter optimization for deep neural network models: a comprehensive study on methods and techniques. *Innovations in Systems and Software Engineering*, 1-12. https://doi.org/https://doi.org/10.1007/s11334-023-00540-3
Rudin, C. (2019). Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature machine intelligence*, *1*(5), 206-215. https://doi.org/https://doi.org/10.48550/arxiv.1811.10154
Saravani, M. J., Noori, R., Jun, C., Kim, D., Bateni, S. M., Kianmehr, P., & Woolway, R. I. (2025). Predicting chlorophyll-a concentrations in the world's largest lakes using Kolmogorov-Arnold

networks. *Environmental Science & Technology*, *59*(3), 1801-1810. https://doi.org/https://doi.org/10.1021/acs.est.4c11113
Scornet, E. (2023). Trees, forests, and impurity-based variable importance in regression. Annales de l'Institut Henri Poincare (B) Probabilites et statistiques,
Shankar, L., & Arasu, K. (2023). Deep Learning Techniques for Air Quality Prediction: A Focus on PM2. 5 and Periodicity. *Migration Letters*, *20*(S13), 468-484. https://doi.org/https://doi.org/10.59670/ml.v20is13.6477
Sharifi, M. S., Aslami, A., Zaheb, H., Abed, I., Shokoori, A. W., & Yona, A. (2024). Modeling the Impact of Socio-Economic and Environmental Factors on Air Quality in the City of Kabul. *Sustainability*, *16*(24), 10969. https://doi.org/https://doi.org/10.3390/su162410969
Sigamani, S. (2024). Air quality index prediction with optimisation enabled deep learning model in IoT application. *Environmental Technology*, *46*(11), 1892–1908. https://doi.org/https://doi.org/10.1080/09593330.2024.2409993
Sun, Q., Zhu, Y., Chen, X., Xu, A., & Peng, X. (2021). A hybrid deep learning model with multi-source data for PM 2.5 concentration forecast. *Air Quality, Atmosphere & Health*, *14*, 503-513. https://doi.org/https://doi.org/10.1007/s11869-020-00954-z
Tang, S. (2024). The box office prediction model based on the optimized XGBoost algorithm in the context of film marketing and distribution. *Plos one*, *19*(10), e0309227. https://doi.org/https://doi.org/10.1371/journal.pone.0309227
Tejaswi, M. (2024). AIR MAP- Deep Learning Prediction in Air Quality for Smarter Decisions. *Interantional Journal of Scientific Research in Engineering and Management*, *08*(05), 1-5. https://doi.org/https://doi.org/10.55041/ijsrem35317
Tsokov, S., Lazarova, M., & Aleksieva-Petrova, A. (2022). A hybrid spatiotemporal deep model based on CNN and LSTM for air pollution prediction. *Sustainability*, *14*(9), 5104. https://doi.org/https://doi.org/10.3390/su14095104
Victoria, A. H., & Maragatham, G. (2021). Automatic tuning of hyperparameters using Bayesian optimization. *Evolving Systems*, *12*(1), 217-223. https://doi.org/https://doi.org/10.1007/s12530-020-09345-2
Wang, T. (2024). Air Quality Prediction based on Neural Network. *Highlights in Science, Engineering and Technology*, *105*, 37-43. https://doi.org/https://doi.org/10.54097/2fsfav47
Wang, X., Zhang, S., Chen, Y., He, L., Ren, Y., Zhang, Z., Li, J., & Zhang, S. (2024). Air quality forecasting using a spatiotemporal hybrid deep learning model based on VMD–GAT–BiLSTM. *Scientific Reports*, *14*(1), 17841. https://doi.org/https://doi.org/10.54097/2fsfav47
Wang, Y., Liu, K., He, Y., Wang, P., Chen, Y., Xue, H., Huang, C., & Li, L. (2024). Enhancing air quality forecasting: a novel spatio-temporal model integrating graph convolution and multi-head attention mechanism. *Atmosphere*, *15*(4), 418. https://doi.org/https://doi.org/10.1038/s41598-024-68874-x
Wardana, I. N. K., Gardner, J. W., & Fahmy, S. A. (2021). Optimising deep learning at the edge for accurate hourly air quality prediction. *Sensors*, *21*(4), 1064. https://doi.org/https://doi.org/10.3390/s21041064
Wonderling, D., Mariani, A., Samarasekera, E. J., Wilkinson, C., Patel, R. S., & Mills, J. (2024). Secondary prevention of cardiovascular disease, including cholesterol targets: summary of updated NICE guidance. *bmj*, *384*, 1-4. https://doi.org/https://doi.org/10.1136/bmj.q637
Xu, R., Wang, D., Li, J., Wan, H., Shen, S., & Guo, X. (2023). A hybrid deep learning model for air quality prediction based on the time–frequency domain relationship. *Atmosphere*, *14*(2), 405. https://doi.org/https://doi.org/10.3390/atmos14020405
Zhang, Z., Zeng, Y., & Yan, K. (2021). A hybrid deep learning technology for PM 2.5 air quality forecasting. *Environmental Science and Pollution Research*, *28*, 39409-39422. https://doi.org/https://doi.org/10.1007/s11356-021-12657-8
Zhao, M. (2025). Synthetic minority oversampling technique based on natural neighborhood graph with subgraph cores for class-imbalanced classification. *The Journal of Supercomputing*, *81*(1), 1-35. https://doi.org/https://doi.org/10.1007/s11227-024-06655-z
Zhao, M., & Ye, N. (2024). High-Dimensional Ensemble Learning Classification: An Ensemble Learning Classification Algorithm Based on High-Dimensional Feature Space Reconstruction. *Applied Sciences*, *14*(5), 1956. https://doi.org/https://doi.org/10.3390/app14051956
Zhao, S., Zhang, B., Yang, J., Zhou, J., & Xu, Y. (2024). Linear discriminant analysis. *Nature Reviews Methods Primers*, *4*(1), 70. https://doi.org/https://doi.org/10.1038/s43586-024-00346-y