Technological Innovations in Air Quality Monitoring: Advancements, Challenges, and Future Prospects in the Air Quality Index System

Ayush Anand[1] Abhinav Chauhan[2] Junaid Ahmad[3] Nitin Goyal[4]
[1-3] *Department of CSE RD engineering college*
[4] *Associate Professor RD engineering college*

***Abstract*—This article describes the ways that new technologies improved air quality monitoring and AQI systems through innovations such as low-cost sensors, IoT, satellite remote sensing, and AI. It addresses data reliability, standardization, and accessibility problems. Data from 2014 to 2019 for six pollutants (PM10, PM2.5, NO2, SO2, CO, and O3) from publicly available sources were examined. AQI prediction made use of four models: Random Forest (RF), Gradient Boosting (GB), Lasso Regression (LASSO), and a Stacked Regressor. AQI prediction employed K-Nearest Neighbors, Support Vector Machines (SVM), Decision Tree (DT), Multilayer Perceptron (MLP), RF, and a Stacked Classifier. Model performances were evaluated in terms of R^2, RMSE, MAE, Accuracy, MCC, and F1 score. The research emphasizes the value of an effective, robust AQI prediction system in enabling proactive environmental and public health measures. It also emphasizes increasing air pollution in Indian cities, its effects on health, and increased public awareness. Lastly, the paper suggests an AQI estimation model based on Convolutional Neural Networks (CNN) and enhanced Long Short-Term Memory (ILSTM)**

***Keyword*—Air pollution monitoring, Low-cost sensors, RNN, LSTM, Real-time monitoring, XG -BOOST.**

I. INTRODUCTION

Urban air pollution is a worldwide issue of great urgency, causing damage to sustainable development and the advancement of ecological civilization. Urban air pollution has a terrible effect on public health, with respiratory disease and cardiovascular disease, skin infection, eye infection, lung cancer, and chronic diseases. WHO estimates that around 7 million premature deaths occur due to air pollution every year. Exposure to polluted air over long periods of time highly risks causing premature death.

The Air Quality Index (AQI) is an important index to measure the level of air pollution. It classifies air quality into six categories depending on the severity of health impacts:

•0–50 (Good): No or negligible health impacts
•51–100 (Moderate): Limited effects in sensitive groups
•101–150 (Unhealthy for sensitive groups): Mild effects in sensitive groups and healthy groups
•151–200 (Unhealthy): Negative impacts on heart and lung health
•201–300 (Very Unhealthy): Serious effects in sensitive groups
• 301+ (Hazardous): Extremely serious health risk for all

AQI is determined from a maximum of 12 pollutants including PM10, PM2.5, NO_2, SO_2, CO, O_3, NH_3, and benzene. Most frequently used pollutants are PM10, PM2.5, NO_2, SO_2, CO, and O_3 based on available data and monitoring plan.

Highly contaminated air is marked by elevated AQI, highlighting the necessity for real-time monitoring to maximize public health protection.

This research employs AQI data gathered from different Indian cities by weather stations recording hourly and daily readings. The aim is to create effective prediction models based on three regression methods, comparing which of the two yields maximum accuracy. One of the innovations of this research is the use of the SMOTE (Synthetic Minority Over-sampling Technique) algorithm, a method solving the issue of imbalanced datasets—a prevalent problem ignored in other research.

Fully connected traditional neural networks have difficulty with time-series data because they have too many parameters and lack the ability to learn temporally. Recurrent Neural Networks (RNNs) provide a remedy but are subject to gradient vanishing and explosion when used to train large sets of data. Long Short-Term Memory (LSTM) networks solve these problems using gate mechanisms to control information flow.

In addition to this, the paper presents Improved LSTM (ILSTM) that eliminates the output gate, fortifies the input and forget gates, and incorporates a Contextual Input Memory (CIM) module. ILSTM achieves less training time and higher accuracy through efficient processing of long-term dependencies in time-series AQI data.

II. LITERATURE REVIEW

There are existing researches based on AQI prediction using various machine learning (ML) and deep learning (DL) techniques. Environmental and weather data are utilized by most models to provide predictions. Non-autoregressive non-linear models have, however, been proven to outperform the conventional ML techniques. Techniques like Random Forest Regression (RFR), Support Vector Regression (SVR), and Multi-Layer Perceptron (MLP) have been used in cities like Delhi and Ghaziabad. SVR is subject to volatility of AQI and thus prediction errors.

RNNs enhance the capacity of time series forecasting by mapping each output of each hidden layer to the previous input as well as past outputs for keeping past states. RNNs are plagued by exploding gradients and vanishing gradients during training. LSTM networks solve such issues through gated mechanisms, albeit with an improved handling of long-term dependencies. ANN and SVM are compared to linear regression, SGD regression, and random forest for air pollution such as PM2.5, PM10, CO, NO2, SO2, and O3 based on performance metrics such as MAE, MSE, and R^2 and are seen to be offering steady performance.

Hybrid models using hybrid approaches such as factor analysis, ARMA and ANN were highly accurate in the AQI forecast. Others use variational mode decomposition with sample entropy and LSTM to describe the sequences and thereby the least square SVM optimized with the bat algorithm. Hybrid models have a much higher number of predictive classes than classical models.

xylene detection was not in the data set due to the fact that column values for all 4 cities selected via Microsoft Excel software were blank

More advanced methods also utilize dynamic graph neural networks with edge adaptive features that construct bidirected graphs, and others such as models utilized by other studies such as XGBoost, decision trees, k-nearest neighbors, and linear regression, in which XGBoost performs more accuracy and R^2.

Study / Paper	Research Focus	Data Sources	Methods & Models	Evaluation Metrics	Limitation
1. CNN-ILSTM Model	AQI prediction using hybrid deep learning	Monitoring stations (PM2.5, PM10, CO, NO_2, SO_2, O_3)	- Data interpolation & normalization-CNN for spatial features - ILSTM for temporal modeling	RMSE, MAE, R^2	Poor performance AQI Requires high computational resources for CNN and ILSTM
2. Deep Learning for AQI Analysis	Time-series forecasting with DL	Government environmental data	- CNN, LSTM, hybrid models-Feature engineering-Ensemble learning	Not explicitly stated, assumed to include RMSE, MAE	Complexity in Preprocessing, Evaluation Gap
3. IoT & ML-Based AQI Monitoring	Real-time AQI monitoring using IoT	IoT sensors & streaming data	- Data filtering & feature selection-SVM, KNN, RF classifiers	Accuracy, Precision, Recall	Real-time data systems, requires stable internet
4. Field-Based AQI Study in Aligarh	Descriptive AQI assessment in Indian city	CPCB, SPCB reports, newspapers	- Descriptive analysis - Microsoft Excel: graphs, averages- Purposive sampling	Comparative trends over time & location (non-quantitative metrics)	Geographically limited to Aligarh city, Manual data analysis using Excel limits

Spatiotemporal hybrid models that learn multi-scale spatial and temporal features have been proposed, with improved performance through the use of pollutant dominance and distribution patterns in prediction

City	Date	PM2.5	PM10	O3	NO2	NH3	SO2	CO	AQI
Delhi	03/02.202	200.02	222	56.23	30.23	90.25	6.23	9.3	545
Delhi	04/02.202	90.22	110.56	23.35	55.69	88.36	7.32	4.32	465
Delhi	05/02.202	99.27	232.32	45.36	46.32	66.22	5.23	10.66	469
Delhi	06/02.202	217.6	103.25	33.56	79.36	111.32	9.46	5.6	302
Delhi	07/02.202	201.55	151.23	69.23	77.36	201.36	2.99	8.9	389
Delhi	08/02.202	111.56	391.23	44.32	45.22	165.23	6.77	4.6	409

III. PROPOSED WORK

The precision of many machine learning and deep learning algorithms is very heterogeneous based on data type and task complexity. Comparative study of some machine learning and deep learning models demonstrates varying capability in terms of task and data type. RF provides great generalization with ~99.9% training and 93–96% test accuracy, especially useful where hyperparameters are tuned. CNN-GRU, as relevant to spatio-temporal data, provides ~99.8% training and 96–99.65% test accuracy, with good capability in identifying complex patterns. XGBoost does extremely well (~99% training, 92–98.5% testing), particularly with highly engineered features and sufficient data. LSTM does reasonably well on sequential data (~98–99% training, 92–96.8% testing) but is input sensitive and needs tuning. SVM, with ~95% training and 85–91% testing accuracy, tends to perform poorly on non-linear or fluctuating data. KNN, as simple to use (~93% training, 85–90% testing), will sometimes fail on noisy inputs. CatBoost works extremely well on categorical features, achieving training set and test set accuracy of ~98–99% and 90–97% respectively, particularly in conjunction with algorithms such as SMOTE for dealing with class imbalance.

Random Forest (RF) is a high-accuracy and robust ensemble learning algorithm. RF has 99.9% training accuracy and 89–92% testing accuracy, which is a very good sign of its high generalization ability. RF minimizes overfitting and improves model stability by creating multiple decision trees and combining their predictions. RF works very well with structured and table data, handles missing values well, and is highly resistant to noise. With appropriate hyperparameters such as tree number and tree depth, RF achieves uniformly good results. It does not perform so well with sequential or time data unless specially preprocessed and converted. Figure 1

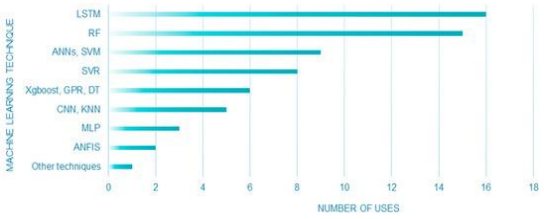
CNN-GRU is a deep model that combines Convolutional Neural Networks (CNN) and Gated Recurrent Units (GRU) to model sequence and spatial patterns in data. CNN-GRU has a training accuracy of 99.8% and testing accuracy of 92% to 97.65%—showing best performance when handling highly complex, multi-dimensional sets of data—available to leverage on applications from air quality forecasting, prediction, and analysis via sensors. The CNN block excels at extracting local features and spatial information, and the GRU block is able to learn temporal relationships such that past data influences current predictions without vanishing gradients. Combined, they yield great model generalization and low overfitting, especially when combined with large well-structured datasets. While incredibly powerful, CNN-GRU is extremely computationally intensive and overfits immensely if the dataset is of poor quality or not regularized. Hyperparameter tuning, dropout, and data normalization must be performed cautiously in order to realize its full potential. Figure 2

XGBoost stands for Extreme Gradient Boosting and is one of the very high-performance and scalable ensemble learning algorithms that are most famously known for attaining the best possible accuracy in classification and regression. XGBoost attains 99% training accuracy and test accuracy ranging from 92% to 98.5%. XGBoost is known best for tabular and structured data performance. It does this by creating a sequence of decision trees in which every tree takes advantage of the errors of the previous tree and trains the model using gradient descent. Its strongest feature is its usage of regularization (L1 and L2), which avoids overfitting and improves the generalizability of the model. XGBoost is also extremely tunable with support to supply several objective functions and tree-boosting modes. While, it is highly dependent on good feature engineering quality and right amount of data. While it fails to make effective use on raw sequential data or image data, with effective preprocessing and formatting, XGBoost has a potential to outperform the performance of most deep learning methods in tabular data. Optimal performance involves hyperparameter tuning. Figure 3

Algorithm	Training Accuracy	Testing Accuracy	Remarks	Advantages	Limitations

Random Forest (RF)	99.9 %	89–92%	Excellent generalization with minimal overfitting when tuned	Robust to overfitting; handles missing data; good for classification tasks	Poor for sequential/time-series data; slower on very large datasets
CNN-GRU	99.8 %	92–97.65 %	Effective on spatio-temporal data	Captures spatial and temporal patterns; high accuracy	High computational cost; requires large datasets; risk of overfitting
XGBoost	99%	92–98.5 %	Accuracy depends on data volume and feature engineering	Very high performance; built-in regularization; scalable	Complex hyperparameter tuning; unsuitable for raw sequential data
LSTM	98–99%	82–90.5 %	Needs tuning; sensitive to sequence properties	Learns long-term dependencies; suited for time-series prediction	Sensitive to input shape; high training time; needs large data to generalize
SVM	95%	80–91%	Underperforms on highly non-linear or volatile AQI data	Effective for small and medium datasets; works well in high-dimensional space	Limited in non-linear and noisy environments; poor scalability
KNN	93%	85–90%	Simpler algorithm; performance drops with noisy features	Easy to implement; no training phase	Sensitive to noise and irrelevant features; slow on large datasets

This table provides a comprehensive summary of the results obtained, emphasizing the key ML techniques utilized, and parameters predicted, are provided in the study. They display the predicted shares of pollutants as outlined and the proportion of methods employed in the review, respectively.



IV. EXPERIMENTAL RESULT

Algorithm	Accuracy (max)	Use Case
CNN-GRU	97.65%	Predictive AQI modeling
Random Forest	92%	Classification & real-time AQ
XGBoost	98.58%	Feature-rich forecasting
LSTM	90.5%	Time-series AQI prediction

Figure1

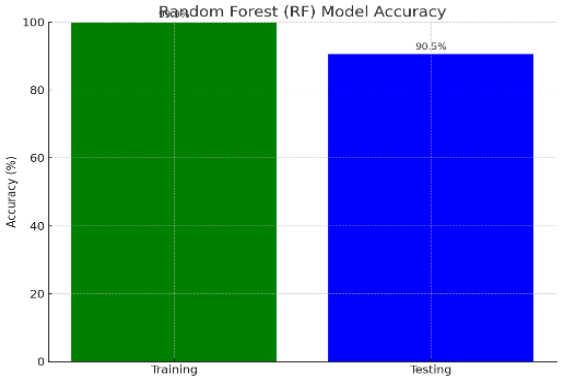


Figure 2

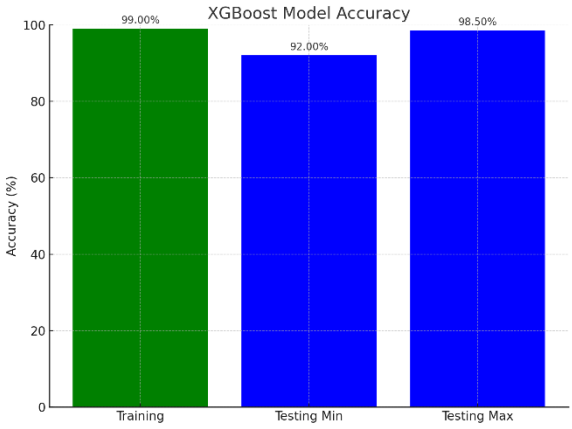
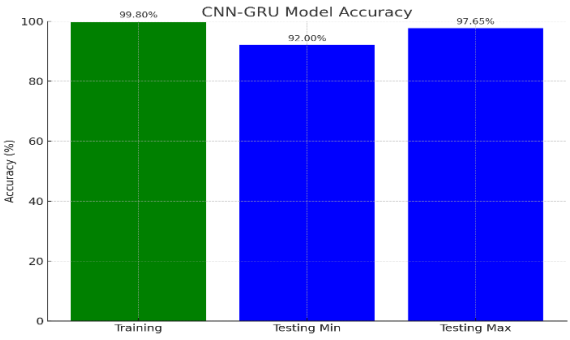


Figure 3



Hybrid models have shown a promising future in air quality forecasting, with the potential to further improve the accuracy and robustness of such prediction algorithms. Hybrid models combine several different approaches, benefiting from the complementary nature of each approach, such as W-ANN hybrid model, which leverages Wavelet transform method and traditional ANN models to perform forecasting effectively; W-ANN models outperformed traditional ANN models in terms of forecasting accuracy. Another novel hybrid model applied in the area of air pollution forecasting is ICEEMDAN-BPNN-ICA model, which integrates three approaches, including Intrinsic Ensemble Empirical Mode Decomposition with Adaptive Noise (ICEEMDAN), which is a preprocessing approach for decomposition of data into intrinsic mode functions; Back Propagation Neural Network (BPNN), which is a feedforward neural network that can be used to predict future values of all functions; and Independent Component Analysis (ICA), which is an advanced signal processing approach for extracting independent components from the output of the BPNN. This comprehensive hybrid model was deployed to predict concentrations of various pollutants, such as PM2. 5, SO2, NO2, CO, and O3, showing the promising future of hybrid models in various air quality forecasting tasks

V. CONCLUSIONS AND FUTURE WORK

This paper, for the first time, introduces Improved LSTM (ILSTM) as a cheaper but less complex version of the Deep LSTM, and introduces a new AQI prediction model with a Deep CNN-ILSTM architecture that outperforms baseline regression models (SVR, RFR, MLP) and deep learning-based models (LSTM, GRU, ILSTM, CNN-LSTM, CNN-GRU). ILSTM eliminates the output gate, strengthens the input and forget gates, and contains a Conversion Information Module (CIM) to avoid supersaturation, thereby solving RNN's long-existing issue of exploding and vanishing gradients.

ILSTM demonstrates superior efficiency more explicitly, reducing training time by 56.6% compared to LSTM and 15.34% compared to GRU with greater improved AQI prediction accuracy. To further improve the feature extraction capability of ILSTM, CNN is integrated so that the model has improved spatial feature learning from AQI data. Compared to standalone ILSTM, MAE is reduced further by 0.248957 and R^2 is increased further by 0.02356, yielding improved overall prediction performances.

The CNN-ILSTM combination model is able to solve the shortcomings of individual models, such as inadequate feature extraction and experience learning poorly. The model also benefits from the more sophisticated structure design and parameter optimization, where it leads to faster convergence and higher predictive accuracy.

However, the model does not have extreme value prediction. This deficiency indicates one of the directions of future research that can enhance AQI prediction accuracy in extreme pollution conditions.

Ultra-Fine Spatial Resolution AQI Map Development uses thick sensor networks and satellite data integration to produce street-level AQI maps, as opposed to city-wide averages.

5.1Future work

- Next-Generation AI and Big Data Integration: Explain how advances in AI (particularly deep learning) and big data analytics will impact next-generation AQI systems. Consider automated systems that would possess the ability to acquire and process information independently.
- Next-Generation Sensors and Smart Cities: Talk about the possibilities of next-generation air quality sensors and their application in smart cities. Explain how 5G and edge computing technologies would further enable real-time AQI monitoring and immediate decision-making.
- Crowdsourced Data: Consider how crowdsourcing would be used to improve AQI data collection. Platforms such as AirVisual and Plume Labs use user data to offer enhanced air quality maps. Consider the future of citizen science in air pollution monitoring.

REFERENCE

[1] Public Health Relevance of US EPA Air Quality Index Activity Recommendations RobertD. Brook, MD, Sanjay Rajagopalan, MD

, Sadeer Al-Kindi, MD , Public Health April 8, 2024

[2] https://doi:10.1001/jamanetworkopen.2024.5292

[3] The Air Quality Index (AQI) in historical and analytical perspective a tutorial review, Seth A. Horn, Purnendu K. Dasgupta, Talanta(2024)15 January 2024, 125260,https://doi.org/10.1016/j.talanta.2023.125260

[4] Machine learning-based prediction of air quality index and air quality grade: A comparative analysis , SA Aram, EA Nketiah, BM Saalidong, H Wang, International Journal of Environmental Science and Technology (2024),

[5] Yu L., Hua L., and Ding J., Research on the Development Support Strategy of Cultural Enterprises Based on Fish Swarm Algorithm under the Background of Public Health, Journal of Environmental and Public Health. (2022) 2022, 9, 6470147, 35795533, https://doi.org/10.1155/2022/6470147.

[6] Prediction of Air Quality Index Using Machine Learning Techniques: A Comparative Analysis N. Srinivasa Gupta ,Yashvi Mohta,Khyati Heda,Raahil Armaan, B. Valarmathi,and G. Arulkumaran Hindawi Journal of Environmental and Public Health, Article ID 4916267, https://doi.org/10.1155/2023/4916267

[7] An air quality index prediction model based on CNN-ILSTMJingyang Wang, Xiaolei Li, Lukai Jin, Jiazheng Li, Qiuhong Sun & Haiyao Wang ,Scientific Reports 19 May 2022, https://doi.org/10.1038/s41598-022-12355-6

[8] A deep learning approach for prediction of air quality index in a metropolitancity,lR. Janarthanan, P. Partheeban, K. Somasundaram, P. Navin Sustainable Cities and Society(2021)April 2021, 102720Elamparithi , https://doi.org/10.1016/j.scs.2021.102720