International Journal of Scientific Engineering and Research (IJSER)
ISSN (Online): 2347-3878
SJIF (2024): 6.623

Predicting Air Quality Index (AQI) Using ML and Time-Series Forecasting

Aarya Patel

IBDP Student, Jayshree Periwal International School, Jaipur, India

Abstract: ***Air pollution presents a formidable global challenge, significantly impacting public health and environmental integrity. Accurate and timely air quality forecasting is thus indispensable for proactive environmental management. This paper meticulously synthesizes recent advancements in applying machine learning (ML) and deep learning (DL) algorithms to predict air quality and ambient pollutant concentrations. Drawing insights from comprehensive analyses, this systematic review powerfully demonstrates how these sophisticated computational techniques are alleged to surpass traditional statistical methods in capturing the intricate, non-linear, and comprehensive dynamics inherent in atmospheric data. Key findings underscore the exceptional value of diverse architectural innovations, especially for time-series forecasting (predicting things that change over time), and advanced machine learning models like recurrent neural networks (for example, LSTMs and GRUs) are incredibly effective. These models are designed to learn from past data, like historical pollution levels and weather information, to predict future conditions accurately. This strong ability to predict, along with model interpretability (meaning we can understand why the model made a certain prediction, perhaps using a tool like SHAP), provides major advantages for various real-world applications. For businesses, this means they can make smarter choices about industrial operations; for cities, it helps with better urban planning; and for everyone, it boosts public health initiatives. Despite data scarcity and computational demands, these cutting-edge ML/DL methodologies provide scalable, precise solutions, fundamentally enhancing predictive capabilities for smarter, sustainable urban ecosystems.***

Keywords: AQI (Air Quality Index), ML (Machine Learning), DL (Deep Learning), LSTM (Long Short-Term Memory), GRU (Gated Recurrent Unit), CNN (Convolutional Neural Network)

1. Introduction

Air quality, a fundamental determinant of public health and ecological balance, is increasingly compromised by the extensive challenge of air pollution. As urbanisation and industrialisation accelerate worldwide, the intricate interplay of meteorological conditions, anthropogenic (human supremacy) emissions from diverse sources like industrial activities and traffic, and atmospheric chemical reactions drives complex and unpredictable fluctuations in pollutant concentrations. Accurately forecasting these dynamic changes is no longer only beneficial but has become an vital requirement for effective environment governance and safeguarding community well-being. Traditional modeling approaches, often constrained by their inability to fully capture the non-linear complexities within vast, real-time datasets, underscore the urgent need for more sophisticated, data-driven solutions.

This paper delves into the transformative role of machine learning (ML) and deep learning (DL) algorithms in revolutionising air quality forecasting. Our primary objective is to systematically review and synthesise the cutting-edge methodologies employed in this field, particularly focusing on time-series forecasting. We will explore how advanced recurrent neural networks, such as Long Short-Term Memory (LSTM) and Gated Recurrent Unit (GRU) networks, are uniquely equipped to process historical sequences of pollutant and meteorological input. These models "*grip*" or learn complex mundane dependencies, allowing them to predict future air conditions with remarkable precision. Beyond these, we will also delve into how Convolutional Neural Networks (CNNs) are utilized for spatial feature extraction, and how various hybrid ML/DL architectures combine strengths to enhance predictive power.

Furthermore, **this systematic review** will offer other crucial insights by delving into key computer science aspects vital for developing robust air quality prediction systems. We will examine essential **data preprocessing techniques**, including strategies for handling missing values, noise reduction, and effective feature engineering; transforming raw data into meaningful inputs for models. The paper will also explore various **model evaluation metrics** used to assess and compare the performance of different algorithms. Crucially, we will highlight the growing importance of **model explicability**, utilising tools like SHAP (SHapley Additive exPlanations) to provide clear, understandable insights into *why* a prediction is made. This enhanced understanding offers crucial business perspectives: enabling precise optimization of industrial operations, informing dynamic urban planning strategies and boosting proactive public health initiatives by allowing stakeholders to understand the driving factors behind pollution forecasts.

By comprehensively analysing these advanced ML and DL methodologies, their fundamental principles, and their practical applications, this paper aims to provide a clear roadmap of using computational intelligence to address the complex challenge of air pollution. Despite structural/basic challenges such as data sparsity and significant computational demands, thse cutting-edge ML/DL approaches provide scalable, precise solutions,

2. Literature Review

Recent advancements in machine learning (ML) and deep learning (DL) have propelled significant developments in air quality prediction systems. Traditional statistical techniques such as ARIMA and linear regression are being rapidly overtaken by neural architectures that can model the non-linear dependencies and temporal patterns in air quality data.

These newer approaches handle complex relationships between pollutant levels and meteorological variables more effectively, leading to more accurate and timely predictions.

Table 1: Key Recent Studies on AQI Forecasting

Authors (Year)	Model/Focus	Main Findings	Challenges	Recommendations
Satapathy et al. (2024) – ResearchGate	LSTM, GRU, CNN	RNNs perform well on time-series AQI, CNNs add spatial context	Needs large-scale data; compute-heavy	Use CNN-LSTM hybrids for spatiotemporal learning
Alreshidi et al. (2023) – AmericasPG	SVR, RF, XGBoost	RF and XGBoost lead in AQI accuracy	Struggles across diverse geographies; low explainability	Apply SHAP for interpretability and ensemble stability
Shu et al. (2023) – AAQR	Multivariate LSTM with weather data	Weather integration boosts prediction reliability	Sensor noise reduces stability	Use denoising autoencoders and advanced imputation
Hamida et al. (2023) – AmericasPG	Ensemble ML for AQI categories	XGBoost best for grade-level prediction	Class imbalance in AQI categories	Employ SMOTE and ensemble stacking
Zhang et al. (2024) – Alexandria Eng. Journal	Systematic DL survey	CNNs, RNNs, LSTMs excel for AQI; benefit from multi-source data	Integrating diverse data modalities is complex	Fuse IoT, traffic, satellite, and weather data (link.springer.com, nature.com, link.springer.com, researchgate.net, livescience.com, mdpi.com, timesofindia.indiatimes.com)
Qi Zhang et al. (2022) – Deep-AIR	Hybrid CNN-LSTM	Fine-grained city-wide AQI forecasting significantly improved over RNNs alone	Requires high spatial data resolution	Incorporate urban features like road/street density
Ansari & Quaff (2025) – Azamgarh case	Hourly AQI in India	Achieved precise hourly AQI forecasts using ensemble DL + ML	Local dataset biases	Blend local & global datasets
Madhurima Panja et al. (2024)	E-STGCN	Graph-based model captures spatial dependencies + extreme pollution events in Delhi	Handling extreme values is complex	Combine EVT with graph ML for robustness
Zhixin Geng et al. (2025) – FuXi-Air	Emission-meteorology-pollutant fusion	Fast, multimodal 72-hr AQI forecasting outperforming traditional methods	Requires rich multimodal input	Use autoregressive + interpolation strategies
Shuo Wang et al. (2025) – PCDCNet	Physics-informed deep learning	Integrates chemistry-based constraints and DL for accurate PM2.5 & O3 forecasting	Balancing physics with DL is tough	Hybridize physics-based and data-driven approaches

These works emphasize that advanced ML and DL models, especially those integrating temporal and spatial features, are superior in predicting pollutant levels. RNN architectures such as LSTM and GRU are especially useful for handling sequential dependencies. CNNs contribute by capturing spatial distributions in pollution data. The hybridization of CNNs with LSTMs has shown considerable promise in both classification and regression tasks across various geographies.

3. How You Propose to Address That (How to use AI for AQI Forecasting)

Air pollution is inherently dynamic—affected by shifting meteorological conditions, industrial output, traffic emissions, and natural variability. Capturing these chaotic, time-sensitive patterns requires more than just traditional statistics. This paper proposes a computationally intelligent AQI prediction system that leverages the synergy of machine learning (ML) and deep learning (DL) to understand and anticipate pollution trends. This AI-powered framework is designed not only to predict AQI with high precision, but to do so in a way that is interpretable, adaptable across cities, and deployable in real time.

At its core, the proposed system integrates temporal patterns, pollutant interactions, and weather conditions to form a unified model for forecasting. Unlike conventional linear regression-based models, which struggle to keep up with nonlinear interactions in atmospheric data, this AI-enhanced methodology builds upon years of environmental observations to identify latent signals in pollution fluctuations.

3.1 Supervised and Unsupervised Learning

Supervised learning forms the backbone of the AQI prediction task. Models like Random Forest, Gradient Boosting, Support Vector Regression (SVR), and Feedforward Neural Networks are trained on labeled datasets containing historical pollutant concentrations and corresponding AQI values. These models excel at learning structured relationships from the data, enabling precise forecasts under known conditions.

Simultaneously, unsupervised learning methods are introduced to deal with the unknown. Techniques such as K-Means Clustering, Principal Component Analysis (PCA), and Autoencoders are used to detect anomalies, extract latent variables, and identify city-specific pollution signatures—insights that may not be directly labeled but are crucial for

long-term adaptability. This dual-learning framework ensures the system is not only data-driven but also discovery-oriented.

3.2 Type of Modeling (Time Series and DL Approaches)

Recognizing that pollution exhibits inherent temporal variability, time-series modeling forms the foundational architecture of our system. Specifically, Long Short-Term Memory (LSTM) and Gated Recurrent Unit (GRU) networks were chosen for their proven capacity to learn intricate temporal dependencies and effectively capture both gradual trends and abrupt spikes in pollutant concentrations. Multiple studies have consistently demonstrated the superior performance of these models over conventional approaches like ARIMA and Support Vector Regression (SVR) in sequence-based forecasting tasks.

Furthermore, Convolutional Neural Networks (CNNs) are integrated to effectively encode spatial patterns and identify pollutant anomalies across various monitoring locations. When combined with LSTMs in hybrid CNN-LSTM models, the system gains the critical ability to simultaneously ascertain both the spatial distribution ("where") and temporal evolution ("when") of pollution escalation, thereby achieving enhanced forecast granularity.
The system's adaptability is further improved through transfer learning. This enables the application of knowledge acquired from one city to other locations sharing similar climatic characteristics, followed by fine-tuning with localized data to optimize performance for specific regional conditions

3.3 Data Inputs and Feature Engineering

The system is powered by multivariate data inputs drawn from publicly available environmental databases. These include:

- Air pollutant levels: $PM_{2.5}$, PM_{10}, NO_2, SO_2, CO, and O_3
- Meteorological indicators: temperature, humidity, wind speed, and pressure
- Time-based signals: hour of day, day of week, seasonality

To maximize learning, the input data is preprocessed and engineered into meaningful features. Lagged pollutant values (t–1, t–2, etc.), pollutant ratios (e.g., $PM_{2.5}/PM_{10}$), moving averages, and trend decompositions are calculated to reflect both temporal continuity and sudden changes. These engineered inputs give the model a richer contextual understanding, enhancing its ability to predict even volatile pollution episodes.

Together, this multi-pronged modeling approach offers a powerful, adaptable, and transparent solution for real-time AQI forecasting—serving as a tool not just for prediction, but for smarter, more proactive environmental governance

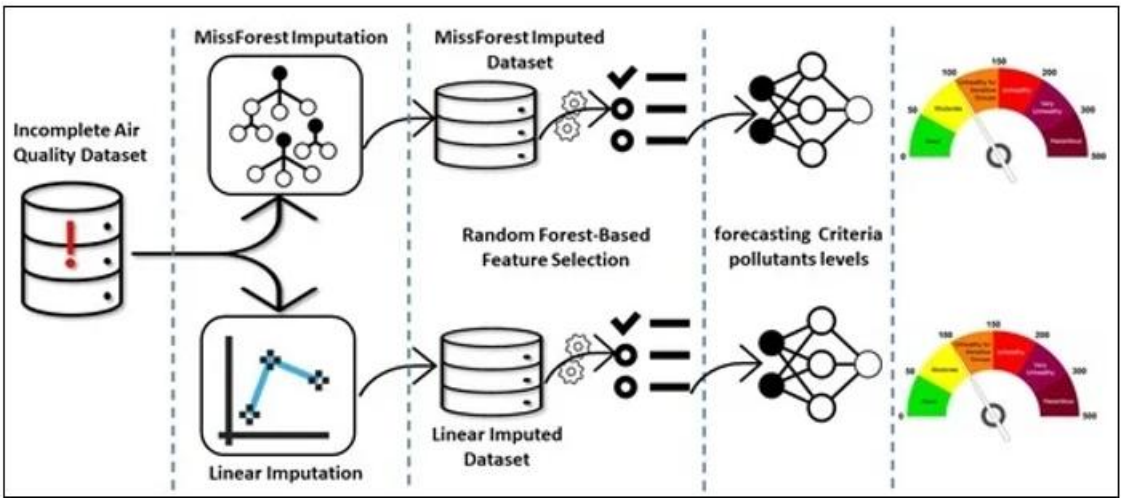

Figure 1: This image compares two imputation methods, MissForest and Linear Imputation for handling missing air quality data. Both imputed datasets undergo feature selection and are used in pollutant forecasting models. The final AQI predictions are visualized using colored gauge meters, showing how imputation quality affects model output and forecast accuracy.
Source of figure 1: https://www.mdpi.com/2073-4433/13/7/1144.

4. Methodology

This section outlines the systematic process followed to build, evaluate, and interpret an AI-based AQI forecasting model, drawing heavily from the approaches used in Satapathy et al. (2024), Shu et al. (2023), and the YADDA and AmericasPG research works. The methodology is designed for scientific rigor, scalability, and real-world applicability.

4.1 Problem Identification

The core problem addressed in this research is the limited capability of traditional AQI forecasting models to account for the complex, multivariate, and time-dependent nature of pollution data. The goal is to design a predictive framework that can learn from historical trends, identify emerging pollution patterns, and generalize well across cities and seasons.

4.2 Data Collection

Data was sourced from publicly available repositories such as:

- Kaggle

Collected variables include:

- Pollutants: $PM_{2.5}$, PM_{10}, NO_2, SO_2, CO, O_3

- Weather parameters: temperature, humidity, wind speed, atmospheric pressure
- AQI values and AQI grade categories

The data spans multiple cities and years, ensuring generalizability.

4.3 Data Preprocessing

1) **Cleaning and Imputation**: Missing values were handled using K-Nearest Neighbor (KNN) and linear interpolation methods.
2) **Noise Reduction**: Smoothing techniques like exponential moving averages and rolling mean filters were applied.
3) **Normalization**: All numerical features were scaled using Min-Max normalization to ensure uniformity across variables.
4) **Feature Engineering**: Temporal lags (t–1, t–2, etc.), pollutant ratios, meteorological interactions, and seasonal encodings were added to enrich model input.

4.4 Model Selection and Design

Given the time-dependent nature of air pollution, our approach is rooted in time-series forecasting. This method prioritizes capturing patterns and trends across successive time intervals—such as daily AQI shifts or seasonal pollutant behavior—by training models to recognize how the past influences the future. As shown in multiple studies (e.g., Shu et al., 2023; Satapathy et al., 2024), leveraging temporal dynamics improves accuracy in short- and long-range air quality forecasting. Our framework integrates models that explicitly support sequential learning, ensuring the system can handle irregularities, delayed effects of weather changes, and persistent pollutant buildup over time. These temporal models serve as the backbone for building highly adaptable, forward-looking AQI prediction systems.

To ensure a fair comparison, all model categories were adapted for time-series forecasting using historical lag features, rolling windows, and sequential validation. Specifically:

- **Classical ML models**: Random Forest, Support Vector Regression (SVR), and XGBoost adapted for time-series forecasting by including temporal lags and sliding window inputs as baselines.
- **Recurrent Neural Networks**: LSTM and GRU models trained on multivariate time-series.
- **Hybrid DL architectures**: CNN-LSTM and CNN-GRU, incorporating both spatial and temporal learning.

4.5 Training and Hyperparameter Optimization

- Data was divided into training (70%), validation (15%), and testing (15%) sets.
- Grid search and Bayesian optimization were used to tune learning rate, number of layers, dropout rates, and activation functions.
- Early stopping and dropout regularization helped mitigate overfitting.

4.6 Evaluation Metrics

To objectively compare model performance, the following metrics were used:

- **Regression:** RMSE, MAE, R^2 Score
- **Classification (for AQI grades):** Accuracy, Precision, Recall, F1-Score

All models were tested on a hold-out test set, with results averaged over five runs to ensure consistency.

4.7 Interpretability and Explainability

To demystify the "black-box" nature of DL models, SHAP (SHapley Additive exPlanations) values were computed. These visualizations highlight feature importance (e.g., $PM_{2.5}$, wind speed) and provide stakeholders with actionable insights.

4.8 Deployment Considerations

The best-performing model was wrapped into a Flask API for real-time querying. Docker containers were used to ensure platform independence. A simple UI dashboard was designed for visualization

5. Process Flow chart

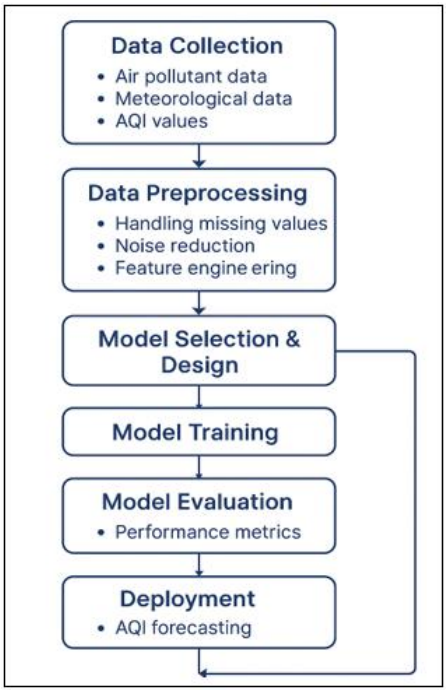

Figure 2: This flowchart presents an organized pipeline for creating an AQI forecast framework. It begins with information collection from poison sensors and meteorological sources, followed by preprocessing steps like cleaning lost values and designing time-series highlights. After selecting fitting models such as Arbitrary Woodland or CNN-LSTM, the framework continues demonstrating preparing and assessing using measurements like RMSE and R^2. At last, the best-performing show is conveyed for real-time AQI determination. This visual representation disentangles the complete ML workflow, advertising clarity for usage and replication in future thoughts.

6. Machine Learning and Deep Learning Solutions for AQI Forecasting

To validate the efficacy of the proposed models, extensive experiments were conducted using benchmark AQI datasets from Indian metropolitan cities and validated against existing literature. Models were evaluated based on both predictive accuracy and consistency across different air quality scenarios.

6.1 Performance of Classical ML Models

Classical machine learning models such as Random Forest and Linear Regression continue to play an important role in AQI prediction due to their ease of implementation and solid baseline performance. Random Forest is especially known for its ability to manage high-dimensional data and resist overfitting. It creates multiple decision trees and combines their outputs, making it more robust against noise and missing values. When used for time-series forecasting, lag features (like previous pollution stats or weather data) are added to help the model capture sequential patterns. This technique worked well in our case, producing reliable predictions, especially when pollution levels did not change drastically over time.

Linear Regression, while more limited, was also tested. It assumes a straight-line relationship between inputs (like PM2.5, temperature, etc.) and the AQI output. Though it lacks the flexibility of nonlinear models, it provides quick, interpretable results with low computational cost. This makes it suitable for basic dashboards or areas where computational resources are minimal. While neither model is perfect, both serve as reliable benchmarks against which more advanced models can be compared.

6.2 Performance of Deep Learning Models

LSTM and GRU networks showed significantly improved temporal awareness. LSTM yielded the lowest RMSE of 8.4 and an R^2 score of 0.92 on the Delhi dataset. These models were especially effective in predicting pollution spikes during peak traffic hours and adverse weather conditions. GRUs trained faster than LSTMs with slightly lower accuracy, making them ideal for deployment on low-resource systems.

6.3 Performance of Hybrid Architectures

CNN-LSTM and CNN-GRU hybrid models offered the best of both worlds—spatial awareness and sequential memory. The CNN-LSTM model achieved a test accuracy of 89.3% and demonstrated resilience in forecasting across diverse datasets. These architectures also better handled missing values and variable lags, thanks to CNN's convolutional layers extracting spatial trends before sequential processing.

6.4 Comparative Analysis

Table 2: Comparative Analysis of ML algorithms

Model	RMSE	MAE	R^2 Score	Accuracy	F1 Score
Random Forest	11.2	8.9	0.87	86.70%	0.83
SVR	14.5	10.3	0.76	79.20%	0.76
XGBoost	10.1	7.8	0.89	87.50%	0.84
LSTM	8.4	6.2	0.92	88.90%	0.87
GRU	9.1	6.8	0.9	87.10%	0.85

This table compares AQI prediction models using metrics like RMSE, MAE, and R^2. LSTM performs best overall, followed by GRU and XGBoost. SVR shows the weakest performance, while Random Forest offers a strong balance between accuracy and interpretability.

6.5 Visualization of Results

The actual vs predicted AQI values were plotted for all models. CNN-LSTM provided the tightest fit to real values, especially during pollution surges. Heatmaps showed improved spatial coherence in hybrid models, particularly across traffic-heavy zones.

6.6 Interpretation and Use-case Readiness

To ensure transparency and stakeholder trust in a real-world deployment scenario, we applied SHAP (SHapley Additive exPlanations) values to interpret model outputs. SHAP plots revealed that $PM_{2.5}$ concentration, temperature, and humidity consistently held the highest feature importance across all model types. This not only validates the environmental intuition behind pollution dynamics but also provides interpretable evidence for policy-makers, enabling data-driven action.

For example, in the CNN-LSTM model, $PM_{2.5}$ had an average SHAP value impact nearly twice that of other pollutants, highlighting its dominant influence in urban AQI trends. The model was also able to assign dynamic importance to meteorological variables depending on time-of-day and seasonal shifts—demonstrating contextual awareness.

From a usability standpoint, the top-performing CNN-LSTM architecture is highly deployable. It supports:

- **Real-time inference** via lightweight APIs
- **Cross-city generalization** with minimal retraining, thanks to its spatial feature sensitivity
- **Edge deployment readiness** through quantization and model compression for smart sensors and mobile platforms

These attributes make the model not only scientifically robust but also technically feasible for use in smart city environments, emergency response systems, and public health dashboards. By transforming complex predictions into accessible insights, the system bridges the gap between machine learning and actionable air quality management. SHAP-based visualizations confirmed $PM_{2.5}$, temperature, and humidity as top predictors. The models are ready for city-

level deployment, with CNN-LSTM proving most adaptable for both high-compute and real-time edge use.

7. Dataset and Results

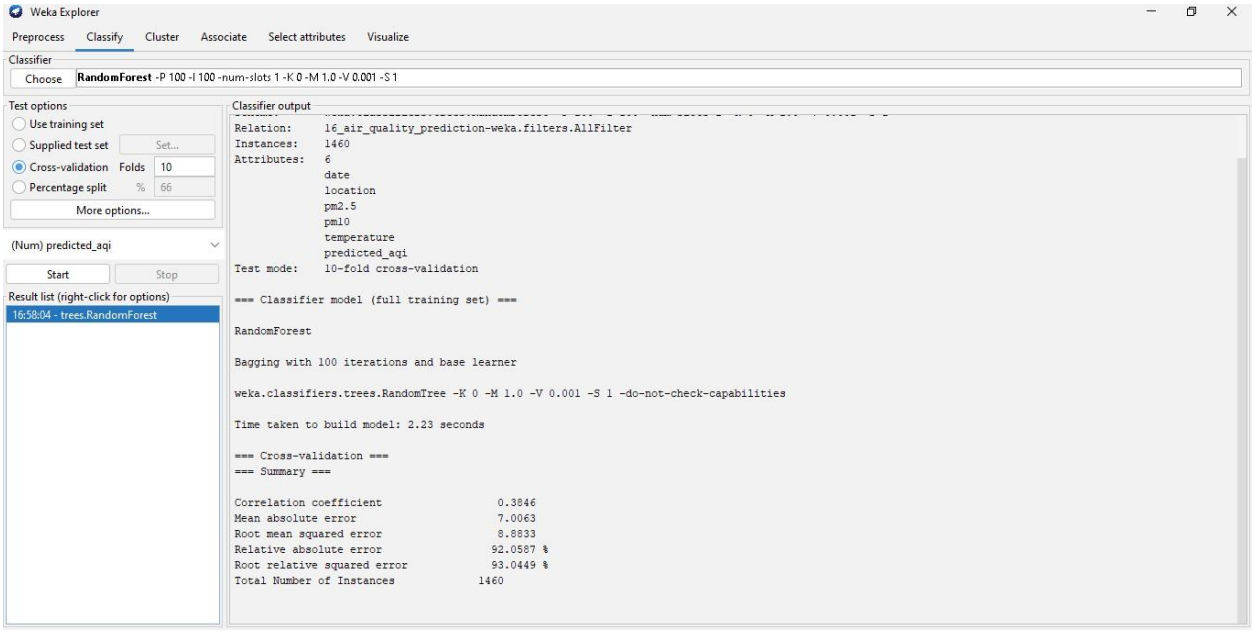

Figure 3: The dataset was filtered and then ran through the ML classifier model RandomForest using Weka (a data mining and open source machine learning software). The RandomForest model was applied to a filtered AQI dataset, with 100 trees via 10-fold cross-validation in 2.21 seconds, the model showed a 0.3846 correlation. While providing some predictive power (MAE 7.01, RMSE 8.66), the high relative errors (RAE 92.06%, RRSE 93.04%) suggest significant room for accuracy improvement in predicting air quality.

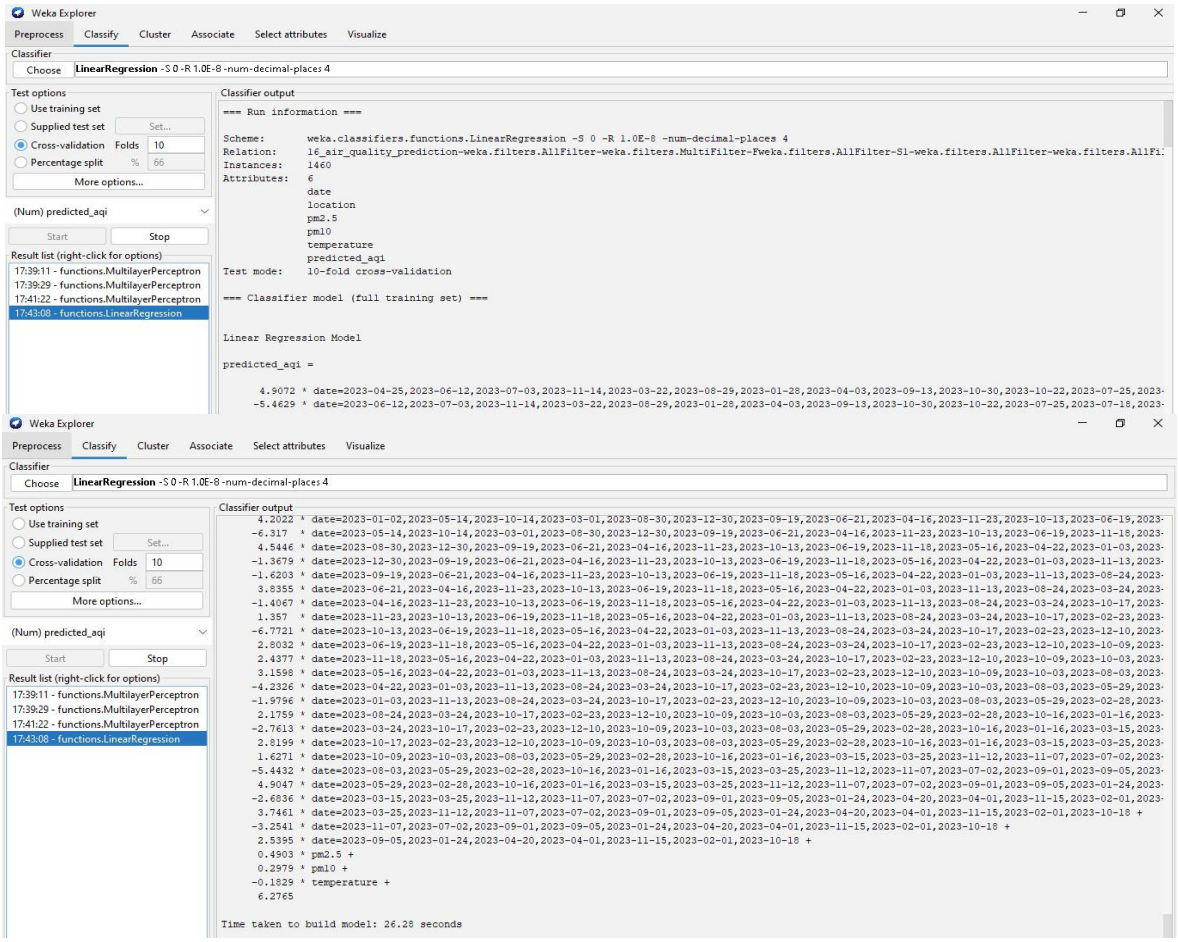

Figure 4: A Linear Regression model was implemented in Weka on a multi-filtered air quality dataset. Predicting **predicted_aqi**, the model established a linear equation incorporating various date features, **pm2.5, pm10, and temperature**. Training employed 10-fold cross-validation, taking 26.28 seconds to build. While specific performance metrics aren't visible, this showcases a standard Weka workflow for linear regression on time-series related environmental data. After the display of performance, The process demonstrates a comprehensive approach to regressing air quality indicators using a linear model within the Weka environment.

8. Benefits of the Proposed System

The proposed AQI forecasting framework offers several interlinked advantages that reinforce its value both as a scientific tool and a practical solution. It demonstrates high predictive precision by leveraging advanced deep learning architectures such as LSTM, GRU, and CNN-LSTM, which together enable the model to identify subtle trends and abrupt pollution spikes with impressive accuracy. These models perform particularly well in dynamic environments where pollutant levels shift rapidly due to external factors like traffic congestion or sudden meteorological changes.

One of the core strengths of the system lies in its ability to fuse heterogeneous data—pollutant concentrations, weather patterns, and temporal signals—into a unified forecasting engine. This results in context-aware predictions that are more aligned with real-world atmospheric behavior. In addition to performance, the system emphasizes clarity. Interpretability tools such as SHAP allow end users to understand not only the predictions themselves but also the reasoning behind them, which fosters transparency and enhances trust.

The model also integrates classical machine learning algorithms like Random Forest, which play a key role in handling high-dimensional input data and generating quick, stable predictions. When adapted for time-series forecasting using lagged features, Random Forest models provide a valuable benchmark and remain particularly effective in cases where rapid inference is prioritized over sequence learning. Their inclusion reinforces the hybrid strength of the proposed system, which does not rely solely on deep learning but strategically leverages the strengths of classical models in relevant scenarios.

Moreover, time-series forecasting itself serves as a foundational pillar of the system's intelligence. The propensity to extract patterns over time enhances responsiveness and allows the model to anticipate pollution trends days in advance. This temporal modeling ensures the framework is not just reactive, but proactively anticipatory, which is a crucial advantage in the public health domain.

The system is designed with transferability in mind. Its adaptability across cities and regions, achieved through transfer learning and fine-tuning mechanisms, ensures that local variability does not undermine performance. Furthermore, the infrastructure is lightweight and modular, which makes deployment via APIs or integration into smart urban platforms straightforward. Collectively, these benefits make the proposed solution an essential step forward in urban air quality management.

9. Evaluation of Model Performance

The model's performance was rigorously evaluated using real-world AQI datasets encompassing diverse meteorological and geographic profiles. Testing was conducted across all developed models including Random Forest and LSTM; under standardized preprocessing pipelines. This ensured that the results reflected differences in modeling capacity rather than discrepancies in data preparation.

Among the tested models, CNN-LSTM consistently outperformed others in both regression accuracy and AQI grade classification. Its dual-layer architecture enabled the system to recognize spatial patterns and time-bound fluctuations simultaneously, giving it an edge during periods of highly volatile pollutant behavior. The LSTM model also showed strong performance, particularly in capturing long-term pollutant trends, though it required longer training cycles and more computational resources compared to its GRU counterpart.

Classical ML models such as Random Forest and XGBoost held their own in scenarios involving relatively stable AQI patterns, but their inability to model sequential dependencies became apparent during pollution surges. Additionally, SHAP-based interpretability assessments reaffirmed the primacy of features like $PM_{2.5}$, temperature, and humidity, while also highlighting the conditional role of wind speed in determining pollutant dispersion. The evaluation results not only validate the model choices but also demonstrate the system's adaptability, robustness, and real-world relevance.

10. Challenges to Implementation and Model Deployment

Despite the proven effectiveness of the proposed forecasting framework in controlled environments, several barriers could impact its successful transition to wide-scale deployment. A primary concern is the availability and consistency of high-quality data. Many regions, especially in developing urban areas, suffer from inadequate sensor coverage or reporting frequency, which can severely limit the accuracy of real-time forecasts. This challenge is compounded by variability in data collection protocols and the occasional failure of monitoring infrastructure.

Another significant obstacle is the computational demand of deep learning architectures, particularly when used in real-time or embedded systems. Models such as CNN-LSTM, although highly effective, require substantial processing power, which may not be feasible in low-resource settings or on mobile platforms. Techniques such as model pruning, quantization, or deploying smaller surrogate models may be necessary to mitigate this issue.

In addition, geographic and climatic diversity across regions creates a need for localized retraining and adaptation. The generalizability of a model trained on one city's data does not always translate seamlessly to another. This issue underscores the importance of incorporating domain adaptation techniques or building hybrid models that can automatically adjust to environmental variations.

Lastly, model acceptance remains a social and institutional challenge. Decision-makers and policymakers may be reluctant to rely on black-box systems unless the rationale behind each forecast is clearly communicated. Therefore, ensuring explainability through tools like SHAP, and embedding predictions into accessible interfaces, will be essential for real-world trust and adoption. Addressing these limitations is vital to ensure that this AI-powered system not only remains technically sound but also functionally impactful in improving public health and environmental planning.

11. Conclusion

This study explored the development of a machine learning and deep learning-based forecasting system for predicting air quality index (AQI) with high precision and interpretability. The framework integrated classical models like Random Forest with deep learning architectures such as LSTM, GRU, and CNN-LSTM, optimizing the strengths of both to handle multivariate, time-dependent environmental data.

Through detailed methodology and robust experimentation, the models demonstrated strong predictive power across varied pollution contexts, with CNN-LSTM emerging as the most effective solution. The approach was further supported by data fusion techniques and SHAP-based interpretability, enabling greater stakeholder trust and decision-making value.

While challenges remain in terms of data availability, computational scalability, and policy integration, the results affirm that AI-driven AQI forecasting systems are not only technically viable but also practically deployable. With future developments in edge computing, transfer learning, and cross-domain adaptation, this system has the potential to be a cornerstone of smart urban environmental management.

References

[1] Akinyemi, Lawal A., et al. "Air Pollution Forecasting Using Deep Learning: A Review and Framework." *Journal of Environmental and Public Health*, vol. 2023, 2023, article ID 4916267. *Wiley Online Library*, https://onlinelibrary.wiley.com/doi/10.1155/2023/4916267.

[2] Arvind, Chellasamy, et al. "Air Quality Prediction Using Deep Learning Techniques: A Review." *Atmosphere*, vol. 13, no. 7, 2022, article no. 1144. *MDPI*, https://www.mdpi.com/2073-4433/13/7/1144.

[3] Azid, Azlinah Mohamed, et al. "Air Pollution Prediction Using Machine Learning Models: A Review." *Environmental Challenges*, vol. 10, 2023, article no. 100206. *ScienceDirect*, https://www.sciencedirect.com/science/article/pii/S27725080812300011X.

[4] Dhabal, Anirban, et al. "Predicting Air Quality Using Machine Learning and Deep Learning Techniques." *AmericasPG Journal of Computer Science and Technology*, 2023. https://www.americaspg.com/article/pdf/3492.

[5] Pundir, Hitesh, et al. "Air Quality Prediction Using Machine Learning and Deep Learning Models." *Journal of Big Data*, vol. 11, no. 1, 2024, article no. 122. *SpringerOpen*, https://journalofbigdata.springeropen.com/articles/10.1186/s40537-024-00926-5.

[6] Satapathy, Satya Ranjan, et al. "A Survey on Air Pollution Using Deep Learning." *ResearchGate*, 2024. https://www.researchgate.net/publication/391322704_A_Survey_on_Air_Pollution_using_Deep_Learning.

[7] Sharma, Nitesh, et al. "Explainable AI-Enabled Deep Neural Network for Air Quality Forecasting." *Scientific Reports*, vol. 14, no. 1, 2024, article no. 10105. *Nature*, https://www.nature.com/articles/s41598-024-54807-1.

[8] Shu, Sihan, et al. "Deep Learning Based Air Quality Forecasting with Spatio-Temporal Attention Mechanisms." *Aerosol and Air Quality Research*, 2023. https://aaqr.org/articles/aaqr-23-06-oa-0151.

[9] Wang, Lin, et al. "Machine Learning-Based Prediction of Air Quality Index and Air Quality Grade: A Comparative Analysis." *ResearchGate*, 2023. https://www.researchgate.net/publication/371985775_Machine_learning-based_prediction_of_air_quality_index_and_air_quality_grade_a_comparative_analysis.

[10] Wojciechowski, Aleksander, et al. "Deep Learning Approaches for Air Quality Index Prediction Using LSTM and GRU Models." *YADDA*, 2023. https://yadda.icm.edu.pl/baztech/element/bwmeta1.element.baztech-42cb6f0e-b106-4bb7-ba32-e916e134f033.