Evaluation of Machine Learning Models for Air Quality Index Prediction

1. I.S. Priya ,2. J. Uma Mahesh ,3. K. Lakshmi Prasanna ,4. K. Gnanasri ,5. M. Madhan, 6. K. Sai Charan

1.Assistant Professor, Department of Electronics and Communication Engineering, Sri Venkatesa Perumal college of Engineering and Technology, Andhra Pradesh, India

2.UG scholar, Department of Electronics and Communication Engineering, Sri Venkatesa Perumal college of Engineering and Technology, Andhra Pradesh, India

3.UG scholar, Department of Electronics and Communication Engineering, Sri Venkatesa Perumal college of Engineering and Technology, Andhra Pradesh, India

4. UG scholar, Department of Electronics and Communication Engineering, Sri Venkatesa Perumal college of Engineering and Technology, Andhra Pradesh, India

5. UG scholar, Department of Electronics and Communication Engineering, Sri Venkatesa Perumal college of Engineering and Technology, Andhra Pradesh, India

6. UG scholar, Department of Electronics and Communication Engineering, Sri Venkatesa Perumal college of Engineering and Technology, Andhra Pradesh, India

ABSTRACT

Air pollution is a hotspot of wide concern in Indian cities. With the worsening of air pollution, urban agglomerations face an increasingly complex environment for air quality monitoring, hindering sustainable and high-quality development in India. More effective methods for predicting air quality are urgently needed. The imperative lies in maintaining pristine air quality and comprehending diverse air pollutants to effectively manage and model air pollution. Given the capricious and variably spatiotemporal nature of pollution, predicting air quality emerges as a notably intricate endeavour.

We used Linear Regression, SGD Regression, LASSO Regression, Random Forest Regression, Ada Boost Regression, Gradient Boost Regression, ANN Regression and LSTM Regression to build regression models for predicting the Air Quality Index (AQI) in India. The root-mean-square error (RMSE), correlation coefficient (r), and coefficient of determination (R2) were used to evaluate the performance of the regression models.

Keywords: *Linear Regression, SGD Regression, LASSO Regression, Random Forest Regression, Ada Boost Regression, Gradient Boost Regression, ANN Regression, LSTM Regression, Root-mean- square error (RMSE), correlation coefficient (r), and coefficient of determination (R^2).*

1.INTRODUCTION:

1.1. GENERAL SYSTEM:

Air pollution has become one of the most pressing global environmental challenges, with widespread implications for public health, climate change, and ecosystem stability. Emissions from human activities such as industrial operations, vehicular traffic, construction, and fossil fuel combustion release harmful pollutants into the atmosphere. These include nitrogen dioxide (NO_2), sulphur dioxide (SO_2), ozone (O_3), carbon monoxide (CO), and fine particulate matter (PM2.5 and PM10), all of which contribute significantly to respiratory and cardiovascular illnesses.

According to the World Health Organization (WHO), air pollution caused approximately 3.7 million premature deaths globally in 2012. Particulate matter, especially PM2.5 and PM10, is considered particularly

dangerous due to its ability to penetrate deep into the lungs and even enter the bloodstream. Chronic exposure to these fine particles has been linked to conditions ranging from asthma and bronchitis to heart disease and cancer.

The **Air Quality Index (AQI)** is widely used to communicate air pollution levels to the public. It aggregates data from multiple pollutants into a single, interpretable number that categorizes air quality on a scale from "Good" to "Hazardous." Accurate and timely prediction of AQI is essential for minimizing the health impact of pollution by enabling preemptive actions and informed policy decisions.

Traditional deterministic modeling techniques for air quality forecasting are often limited by their complexity and dependence on hard-to-obtain physical parameters. As a result, **machine learning (ML)** and **deep learning (DL)** approaches have gained popularity for their ability to learn patterns from historical data, adapt to nonlinear behaviours, and produce reliable forecasts even in noisy environments.

This project aims to build an intelligent air quality prediction system using a combination of regression algorithms and deep learning models. The system leverages historical air quality and meteorological data to forecast pollutant concentrations and AQI with high accuracy. By integrating data preprocessing, advanced feature selection, and multiple prediction algorithms—including Random Forest Regression, Gradient Boosting, and Long Short-Term Memory (LSTM) networks—the proposed solution aims to support public health initiatives, environmental planning, and proactive pollution control strategies.

1.2. APPROACHES TO TUMOR DETECTION

Accurate prediction and monitoring of the Air Quality Index (AQI) require a multifaceted approach, combining advanced sensing technologies with robust computational models. Over the years, researchers and environmental agencies have adopted various techniques to estimate AQI, ranging from traditional deterministic models to modern machine learning and deep learning methods. Below are some of the widely used approaches:

a) Statistical and Deterministic Models

Traditional air quality forecasting systems are often based on physical and chemical models that simulate atmospheric dispersion, meteorological conditions, and pollutant behavior. These models require detailed environmental parameters such as wind speed, humidity, temperature, emission inventories, and chemical reaction rates. While scientifically grounded, these methods are limited by their dependency on comprehensive data collection and complex modeling of atmospheric processes.

b) Time Series Forecasting Models

These models rely on historical pollutant data to predict future AQI values. Techniques such as Autoregressive Integrated Moving Average (ARIMA), Seasonal ARIMA (SARIMA), and Exponential Smoothing are commonly employed. These models are relatively easy to interpret and work well with linear and stationary datasets. However, they struggle with non-linear patterns and multivariate dependencies.

c) Machine Learning Models

Machine learning models such as **Linear Regression**, **Support Vector Regression (SVR)**, **Random Forest**, **Gradient Boosting**, and **LASSO Regression** have proven to be highly effective for AQI prediction. These models can learn from large-scale data and capture complex, nonlinear relationships between pollutant levels and influencing factors such as weather conditions and industrial activity. Ensemble methods like Random Forest and Gradient Boosting often outperform traditional statistical models in both accuracy and robustness.

d) Deep Learning Approaches

Deep learning methods, particularly **Artificial Neural Networks (ANNs)** and **Recurrent Neural Networks (RNNs)** such as **Long Short-Term Memory (LSTM)** networks, are capable of modeling intricate temporal dependencies and high-dimensional data interactions. These models excel in capturing the sequential patterns of air pollution data and are particularly effective for long-term AQI forecasting. Their ability to handle multivariate time series data makes them ideal for real-world applications.

1.3. DOMAIN OVERVIEW

The domain of this project lies at the intersection of environmental monitoring and artificial intelligence, particularly focusing on the application of machine learning (ML) and deep learning (DL) techniques for the prediction of air quality. As air pollution continues to escalate across urban and industrial regions, there is an increasing demand for intelligent systems that can interpret complex environmental data and generate accurate forecasts to mitigate health and ecological impacts. Machine learning plays a pivotal role in building predictive

models that learn from historical data to make future predictions. Among these models, regression algorithms, decision trees, ensemble methods, and neural networks have shown significant success in AQI forecasting.

In parallel, deep learning—a subfield of machine learning—has revolutionized pattern recognition and time-series prediction tasks. Algorithms such as Artificial Neural Networks (ANNs), Convolutional Neural Networks (CNNs), and Long Short-Term Memory (LSTM) networks are capable of automatically extracting hierarchical features from complex and high-dimensional data, making them suitable for handling the temporal and spatial variability of air pollutant levels.

To support these techniques, modern programming tools and environments are employed. Python is widely used due to its simplicity and rich ecosystem of data processing and visualization libraries such as:

- **Pandas** for data manipulation,
- **NumPy** for numerical computations,
- **Matplotlib** and **Seaborn** for data visualization,
- **Scikit-learn** for classical ML algorithms, and
- **TensorFlow** or **PyTorch** for building deep learning models.

Development is often carried out in **Jupyter Notebooks**, which provide an interactive coding interface that supports step-by-step data exploration and model development. The use of **Anaconda**, a popular data science platform, simplifies package management and environment setup, making it easier for researchers to focus on experimentation and results.

In summary, this domain encompasses the use of intelligent computational tools and statistical techniques to address a real-world environmental problem—air pollution. The integration of AI with environmental science enables the development of robust and scalable solutions that can guide proactive decision-making and policy formulation.

2: METHODOLOGY

This research presents a fully automated, data-driven framework for predicting the Air Quality Index (AQI) using a combination of machine learning and deep learning models. The methodology is organized into the following sequential phases: data acquisition, preprocessing, feature engineering, model training, evaluation, and optimization. Each stage is tailored to enhance prediction accuracy, interpretability, and computational efficiency.

A. Data Acquisition

Air quality and pollutant concentration data were collected from publicly available datasets and monitoring stations, including measurements for key pollutants such as PM2.5, PM10, NO_2, SO_2, CO, and O_3. The dataset also includes meteorological variables such as temperature and humidity, essential for modeling pollutant behaviour under different atmospheric conditions.

B. Data Preprocessing

Before modeling, raw data undergoes a series of preprocessing operations to ensure consistency and reliability:

- **Missing Value Handling:** Missing or inconsistent entries are identified and corrected through interpolation or removal.
- **Outlier Detection:** Extreme values are detected using statistical thresholds and managed to reduce noise.
- **Normalization (Min-Max Scaling):** Ensures all features fall within the same range, improving model convergence.
- **Feature Encoding and Resampling:** Time-stamped data is resampled to consistent intervals, and categorical features are encoded appropriately.

C. Feature Engineering and Selection

The success of any predictive model depends heavily on the quality of input features. The process includes Forward Feature Selection which iteratively adds features that contribute most to model performance. This structured feature engineering enhances model accuracy while reducing overfitting and computational cost.

D. Model Development

Multiple regression-based models are developed to compare performance and ensure robustness. These include:

- **Linear Regression & LASSO Regression:** Baseline linear models for interpretability.
- **Random Forest Regression (RFR):** An ensemble-based model that handles feature interaction and noise robustly.
- **Gradient Boost Regression & AdaBoost Regression:** Boosting methods that correct previous model errors.
- **Artificial Neural Network (ANN) Regression:** Captures complex, nonlinear dependencies.
- **Long Short-Term Memory (LSTM):** A deep learning model designed for time-series forecasting, capable of learning long-term dependencies.

Each model is trained on a historical dataset and evaluated against a separate testing set.

E. Model Evaluation

To assess the effectiveness and generalizability of each model, the following metrics are used:

- **Root Mean Square Error (RMSE):** Measures prediction error magnitude.
- **Mean Absolute Error (MAE):** Evaluates average prediction error.
- **R-squared Score (R^2):** Indicates how well the model explains variance in the data.
- **Mean Square Error (MSE):** The average of squares of the differences between predicted and actual.
- **Correlation Coefficient (r):** Measures strength of relationship between actual and predicted values.

F. Computational Efficiency

The system demonstrates high computational efficiency by achieving rapid prediction performance across all regression models. The optimized LSTM and Random Forest models process large input datasets and deliver results within milliseconds to seconds per instance, making them suitable for real-time AQI forecasting applications.

3.EXPERIMENTAL DETAILS

This section highlights the experimental validation of the AQI prediction framework using various regression models, including both machine learning (ML) and deep learning (DL) approaches. The system is evaluated using standard error metrics and tested on real-world air quality datasets.

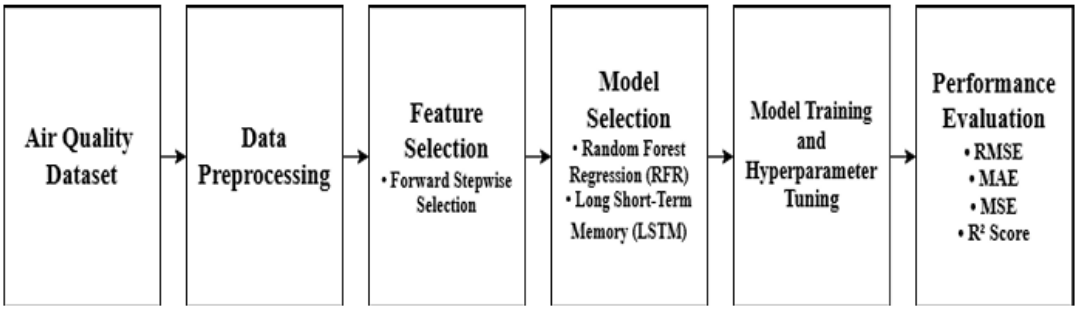

Fig 1: Proposed system architecture

1. **Air Quality Dataset**
   The system begins with the collection of historical air pollution and meteorological data from reliable sources.
2. **Data Preprocessing**
   Raw data is cleaned by handling missing values, removing outliers, and applying normalization to prepare it for analysis.
3. **Feature Selection**
   Important features are selected using forward stepwise selection to reduce redundancy and improve model efficiency.
4. **Model Selection**
   Random Forest Regression (RFR) and Long Short-Term Memory (LSTM) are selected for training based on their predictive strengths.
5. **Model Training and Hyperparameter Tuning**
   Selected models are trained on historical data, with parameters fine-tuned to optimize prediction accuracy.
6. **Performance Evaluation**
   The models are evaluated using RMSE, MAE, MSE, and R^2 Score to measure accuracy and robustness.

Model Performance Graph (LSTM):

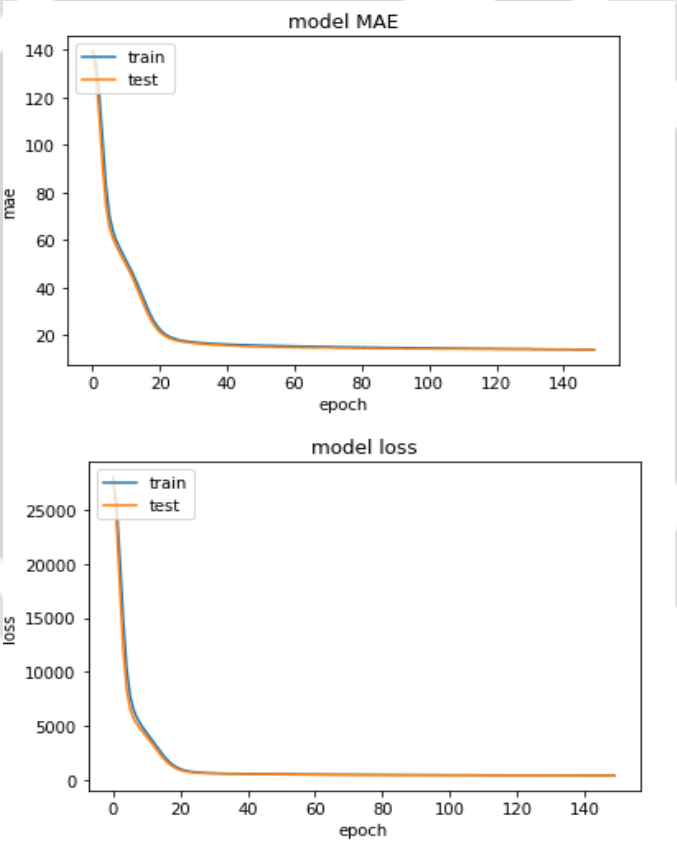

Fig 2: Model MAE and Loss Curves

- **MAE Curve:**
  - Training and test MAE sharply decrease and converge below 20 within 20 epochs.
  - Indicates effective error minimization and strong generalization.
- **Loss Curve :**
  - Both training and validation loss drop rapidly and stabilize before epoch 40.
  - Consistent alignment suggests no overfitting.

These trends confirm that the LSTM model effectively captures the temporal dependencies within the air quality data.

Model Performance Comparison:

The table below summarizes the performance of different regression models used in the experiment:

```
In [147…  results_df
```

Out[147]:

Model Type	Algorithm	Training Score List	r2 Score	MAE Score	MSE Score	RMSE Score
ML: Linear Regression	Linear Regression	0.910501	0.915147	18.142078	787.334723	28.059485
	SGD Regression	0.910373	0.915099	18.152498	791.740502	28.137884
	LASSO Regression	0.908913	0.913202	18.494064	778.996414	27.910507
ML: Ensemble Regression	Random Forest Regression	0.991107	0.948211	13.544907	456.219191	21.359288
	Ada Boost Regression	0.85551	0.818789	30.521939	1277.609149	35.743659
	Gradient Boost Regression	0.956611	0.945580	14.448218	473.809860	21.767174
DL: Deep Learning Regression	ANN Regression	NA	0.937481	15.854204	550.361280	23.459780
	LSTM Regression	NA	0.951494	13.972309	428.259734	20.694437

Fig 2: Model Performance Comparison

LSTM Regression's superior performance across all key metrics especially R^2 and RMSE demonstrates its ability to effectively model the time-series nature of AQI data. It captures the underlying trends and seasonality of air quality patterns better than other models, making it the best choice for accurate, real-time AQI prediction.

4.CONCLUSION

This research demonstrates the effectiveness of various machine learning and deep learning models for predicting the Air Quality Index (AQI) in Indian cities, with a focus on the comparison of several regression techniques. Among the models evaluated—Linear Regression, SGD Regression, LASSO Regression, Random Forest Regression, Ada Boost Regression, Gradient Boost Regression, ANN Regression, and LSTM Regression—the LSTM Regression model emerges as the most accurate and reliable for AQI prediction.

The LSTM model excels due to its ability to capture temporal dependencies and long-term patterns in the air quality data, which are critical for accurate forecasting. The model's ability to handle the time-series nature of AQI data allows it to predict future pollutant concentrations with high precision, minimizing prediction errors as evidenced by its superior performance across all key metrics, especially R^2 and RMSE. In contrast to traditional models like Linear Regression and Random Forest, the LSTM Regression model not only handles nonlinear relationships effectively but also adapts to the dynamic and seasonally fluctuating nature of air quality data. Its ability to generalize well to unseen data, as shown by stable validation losses, further underscores its robustness and suitability for real-world applications.

The findings highlight that while traditional machine learning models can offer useful insights, deep learning models like LSTM are better equipped to handle the complexities inherent in AQI prediction. This has significant implications for urban air quality management, as accurate and timely AQI forecasts can help policymakers and citizens take proactive measures to mitigate health risks associated with air pollution.

In conclusion, the LSTM Regression model is not only the best performer for AQI prediction in terms of accuracy but also offers a scalable solution for real-time, data-driven air quality forecasting in urban environments. Future research could explore further enhancements, such as integrating additional environmental and socio-economic factors, to improve the model's predictive power and applicability across different regions.

5.REFERENCES

[1] Fareena Naz,Muhammad Fahim,Adnan Ahmad Cheema,Nguyen Trung Viet,Tuan-Vu Cao,Ruth Hunter,Trung Q. Duong Two Stage Feature Engineering to Predict Air Pollutants in Urban Areas IEEE Access, 2024

[2] Ying Zhang,Yanhao Wang,Minghe Gao,Qunfei Ma,Jing Zhao,Rongrong Zhang,Qingqing Wang,Linyan Huang A Predictive Data Feature Exploration-Based Air Quality Prediction Approach IEEE Access, 2019

[3] Yuchao Zhou,Suparna De,Gideon Ewa,Charith Perera,Klaus Moessner Data Driven Air Quality Characterization for Urban Environments: A Case Study IEEE Access, 2018

[4] Dongming Qin,Jian Yu,Guojian Zou,Ruihan Yong,Qin Zhao,Bo Zhang A Novel Combined Prediction Scheme Based on CNN and LSTM for Urban PM2.5 Concentration IEEE Access, 2019

[5] Yu Cong,Ximeng Zhao,Ke Tang,Ge Wang,Yanfei Hu,Yingkui Jiao FA-LSTM: A Novel Toxic Gas Concentration Prediction Model in Pollutant Environment IEEE Access, 2021

[6] Xiaoxu Wei,Xiaokai Wang,Tao Zhu,Zhen Gong Fusion Prediction Model of Atmospheric Pollutant Based on Self-Organized Feature IEEE Access, 2021

[7] Md. Mokhlesur Rahman,Kamal Chandra Paul,Md. Amjad Hossain,G. G. Md. Nawaz Ali,Md. Shahinoor Rahman,Jean-Claude Thi l Machine Learning on the COVID-19 Pandemic, Human Mobility and Air Quality: A Review IEEE Access, 2021

[8] Uzair Aslam Bhatti,Yuhuan Yan,Mingquan Zhou,Sajid Ali,Aamir Hussain,Huo Qingsong,Zhaoyuan Yu,Linwang Yuan Time Series Analysis and Forecasting of Air Pollution Particulate Matter (PM2.5): An SARIMA and Factor Analysis Approach IEEE Access, 2021

[9] Abdelaziz El Fazziki,Djamal Benslimane,Abderrahmane Sadiq,Jamal Ouarzazi,Mohamed Sadgal An Agent Based Traffic Regulation System for the Roadside Air Quality Control IEEE Access, 2017

[10] Ichrak Mokhtari,Walid Bechkit,HervÃ© Rivano,Mouloud Riadh Yaici Uncertainty-Aware Deep Learning Architectures for Highly Dynamic Air Quality Prediction IEEE Access, 2021 DEPARTMENT OF ECE 53 SVPCET,PUTTUR Evaluation of Machine Learning Models for Air Quality Index Prediction

[11] Qilong Han,Peng Liu,Haitao Zhang,Zhipeng Cai A Wireless Sensor Network for Monitoring Environmental Quality in the Manufacturing Industry IEEE Access, 2019

[12] F. H. Dominski, J. H. L. Branco, G. Buonanno, L. Stabile, M. G. da Silva and A. Andrade, Effects of air pollution on health: A mapping review of systematic reviews and meta-analyses, Environ. Res., vol. 201, Oct. 2021. [13]. R. Thompson, R. B. Smith, Y. B. Karim, C. Shen, K. Drummond, C. Teng, et al., Air pollution and human cognition: A systematic review and meta-analysis, Sci. Total Environ., vol. 859, Feb. 2023.