Data-Driven Computational Anticipation of Air Quality

Talapally Aarthi[1] Mr. Sandeep Agarwalla[2] Dr.M.Sambasivudu[3]

[1] *Research Scholar, Dept. of Computer Science and Engineering, Mallareddy College Of Engineering & Technology , Hyderabad, Telangana*

[2] *Associate Professor, Dept.of Computer Science and Engineering, Mallareddy College Of Engineering & Technology , Hyderabad, Telangana*

[3] *Associate Professor, Dept.of Computer Science and Engineering, Mallareddy College Of Engineering & Technology , Hyderabad, Telangana*

ABSTRACT:

The air quality monitoring system measures pollutant levels across different locations, tackling a critical public health concern. Industrial emissions, vehicle exhaust, and other sources have pushed urban air quality beyond safe regulatory limits, endangering human health. By harnessing advancements in machine learning (ML), we can now forecast future pollutant concentrations based on historical and real-time data.In this work, we introduce a compact IoT device that captures current pollutants using MQ-series gas sensors connected to an Arduino Uno. It then applies ML algorithms—such as ARIMA, LSTM, or Prophet—to forecast upcoming contamination levels and logs the data into an Excel spreadsheet for further analysis. This architecture mirrors successful implementations where Arduino-based IoT frameworks, combined with predictive models, monitor multiple pollutants (e.g., CO, SO_2, NO_2, O_3, PM2.5, PM10), with tools like Prophet often delivering superior predictive performance Additionally, more comprehensive systems using diverse sensors (e.g., MQ-135, MQ-9, MQ-6 plus DHT for temperature/humidity) have enabled real-time AQI prediction and classification using methods like Random Forests, achieving classification accuracies around 99%. Edge-level air quality prediction is also advancing: TinyML models deployed on microcontrollers (e.g., Arduino Nano 33 BLE Sense) now predict ozone concentrations in real time with strong performance ($MSE \approx 0.03$, $R^2 \approx 0.95$). By combining low-cost sensor hardware with lightweight ML models, our device offers both air quality monitoring and forward-looking predictions. This fusion empowers communities and urban planners with actionable insights, enabling proactive responses to pollution trends and enhancing environmental health management strategies.

Keywords:

Regression models (linear, ridge, LSTM),tree-based models (random forest, LightGBM, CatBoost),deep learning (CNN-LSTM, hybrid deep models, GNNs),meteorological feature engineering (temperature, wind, humidity),time-series modeling, seasonal encoding,MAE, RMSE, R^2 evaluation metrics

1. INTRODUCTION

Air quality monitoring is essential due to its significant impact on both human health and environmental stability. Pollutants such as ozone (O_3), nitrogen dioxide (NO_2), carbon monoxide (CO), sulfur dioxide (SO_2), and fine particulates ($PM_{2.5}$/PM_{10}), which originate from industrial emissions, vehicle exhaust, fossil fuel combustion, and natural sources like volcanic activity, are invisible yet highly hazardous. These contaminants have been conclusively linked to respiratory illnesses, cardiovascular diseases, cancer, birth defects, neurological conditions, and organ damage affecting the kidneys, liver, and brain .Beyond impacting human health, air pollution disrupts ecosystems by causing acid rain, reducing crop production through impaired photosynthesis, degrading soil quality, harming forests and aquatic life, and contributing to smog that limits visibility and alters weather patterns—including rainfall distribution . Economically, it places a burden on healthcare systems, reduces workforce productivity, and undermines national GDP. Since these pollutants are imperceptible and fluctuate across time and space, continuous, automated monitoring is vital. Such systems provide real-time insights into pollutant concentrations, identify pollution sources, and track trends, enabling authorities and communities to implement effective mitigation strategies . The Air Quality Index (AQI) serves as a standardized tool to convert pollutant concentrations into health risk categories—ranging from “Good” to “Hazardous.” Calculations follow a piecewise linear equation:

I=Ihigh−IlowChigh−Clow(C−Clow)+IlowI = ₩frac{I_₩text{high}−I_₩text{low}}{C_₩text{high}−C_₩text{low}}(C − C_₩text{low}) + I_₩text{low}I=Chigh−ClowIhigh−Ilow(C−Clow)+Ilow

where CCC is the measured pollutant concentration, and the breakpoints Clow,Chigh,Ilow,IhighC_\text{low}, C_\text{high}, I_\text{low}, I_\text{high}Clow,Chigh

,Ilow,Ihigh define threshold boundaries between AQI categories. An AQI above 300 signifies hazardous air quality, prompting urgent public health advisories. In summary, an integrated framework of continuous monitoring, data interpretation via AQI, and responsive policy intervention is critical for safeguarding public health, preserving ecosystems, and informing effective environmental regulation. Continuous monitoring empowers timely action, ensures regulatory compliance, increases public awareness, and guides efforts to build cleaner, healthier environments.

2. LITERATURE SURVEY

The existing literature on air quality monitoring and prediction has predominantly concentrated on forecasting individual pollutants, especially $PM_{2.5}$. For instance, a study conducted in Santander, Spain, demonstrated that event-based sensing techniques could reduce sensor energy consumption by up to 50% compared to traditional periodic sensing methods . Additionally, various machine learning algorithms, such as Generalized Linear Models (GLM), Support Vector Machines (SVM), and Bayesian methods, have been employed to predict $PM_{2.5}$ concentrations using meteorological data . However, these approaches often overlook the complex interactions among multiple pollutants and meteorological factors. To address this gap, the proposed system aims to predict levels of all major air pollutants—CO, O_3, NO_2, SO_2, $PM_{2.5}$, and PM_{10}—by integrating comprehensive datasets from government APIs, IoT sensors, and meteorological stations. This multi-pollutant approach, coupled with advanced machine learning models like Random Forest, SVM, and Long Short-Term Memory (LSTM) networks, enables more accurate and holistic air quality forecasting. Furthermore, while previous studies have utilized machine learning for pollutant prediction, they often rely on static models or focus on specific regions. The proposed system introduces a dynamic, scalable framework that continuously learns from incoming data, adapts to changing environmental conditions, and provides real-time air quality information. This advancement not only enhances prediction accuracy but also supports timely interventions and policy-making to mitigate air pollution's impact on public health and the environment.

3. METHODOLOGY

Data Collection & Preprocessing

Collect toxic concentrations ($PM_{2.5}$, PM_{10}, NO_2, CO, O_3, SO_2) and climate data (temperature, stickiness, wind) from open sensors and storage facilities such as UCI or government air-monitoring frameworks. Clean the dataset by taking care of misplaced values (unfeeling, center, k-NN attribution), removing special cases, and normalizing highlights by implies of scalers like StandardScaler Incorporate Building Select impactful markers: harm levels and climate parameters Make time-series highlights like slack values and rolling midpoints. Apply dimensionality diminishment (e.g., PCA) for high-dimensional datasets .Get ready backslide models tallying Sporadic Forest, SVR, CatBoost, and neural frameworks (Thick, LSTM, CNN-LSTM cross breeds) Portion data into training/testing (e.g., 80/20) and utilize cross-validation for tuning Hyperparameter Tuning & Evaluation Optimize models utilizing organize and enthusiastic hunt for parameters (e.g., n_estimators, learning_rate) Evaluate execution by implies of RMSE, MAE, and R^2; compare comes around over models, tallying furnish stacks

4. PROPOSED SYSTEM

The proposed air quality monitoring system employs machine learning (ML) techniques to predict air quality levels by analyzing data from diverse sources such as government APIs, weather forecasts, IoT sensors, and historical pollution records. By utilizing advanced ML models like Random Forest, Support Vector Machines (SVM), and Long Short-Term Memory (LSTM) neural networks, the system can identify complex, non-linear patterns and interactions among multiple pollutants and atmospheric conditions. This enables highly accurate short-term and long-term forecasting of the Air Quality Index (AQI). Unlike static models, ML techniques continuously learn and improve as more data becomes available, making predictions more reliable over time. Furthermore, the system is scalable and cost-effective, allowing integration with mobile apps or public dashboards for real-time public awareness. The proposed approach not only enhances prediction accuracy but also supports environmental agencies and policymakers in implementing timely air pollution mitigation strategies.

Unique Features of the System

- Multi-Source Data Integration: The system aggregates data from various sources such as meteorological stations, satellite imagery, IoT sensors, government APIs (like CPCB or

AQI India), and real-time user inputs from mobile applications. This comprehensive data input enhances the richness and accuracy of the prediction model.

- Advanced Machine Learning Algorithms: It employs sophisticated algorithms such as Random Forest, Gradient Boosting, Support Vector Machines (SVM), and deep learning models like LSTM (Long Short-Term Memory). These models are trained on historical and real-time data to identify hidden patterns and correlations between pollutants and atmospheric parameters.
- Real-Time Prediction and Updates: One of the most vital features is the real-time capability of the system. The model continuously updates its predictions as new data is streamed from sensors or APIs, ensuring up-to-date air quality information and alerts.
- High Spatial and Temporal Resolution: The system supports fine-grained geographical predictions, offering AQI forecasts not only at the city or district level but down to specific neighborhoods or localities. Temporal resolution is also high, enabling hourly or even minute-level predictions.
- User-Centric Mobile and Web Interfaces: With integrated visualization dashboards and mobile apps, the system provides intuitive data displays such as pollution heatmaps, historical trends, and forecast graphs. It also supports personalized alerts for sensitive groups like children or individuals with respiratory issues.
- Self-Learning and Model Optimization: The ML models are designed to retrain automatically with new incoming data, ensuring continued accuracy and performance improvement. The system adapts to seasonal variations and sudden pollution events like wildfires or construction surges.

5. SYSTEM ARCHITECTURE

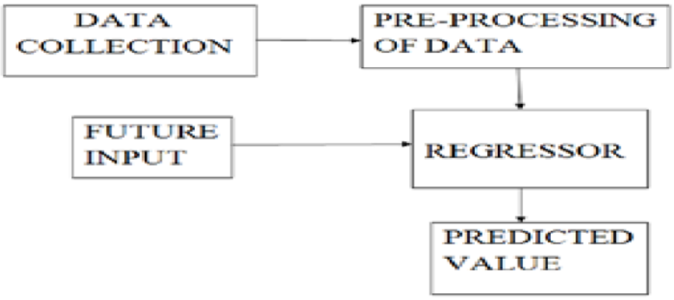

Figure 5.1 System Architecture

A multi-tier pipeline ingests sensor data (e.g., pollutant levels, weather) and social media streams; data is pre-processed, engineered into features, and sent to separate ML inference Fig 5.1 clusters for air quality forecasting and attack detection.

6. RESULTS AND DISCUSSION

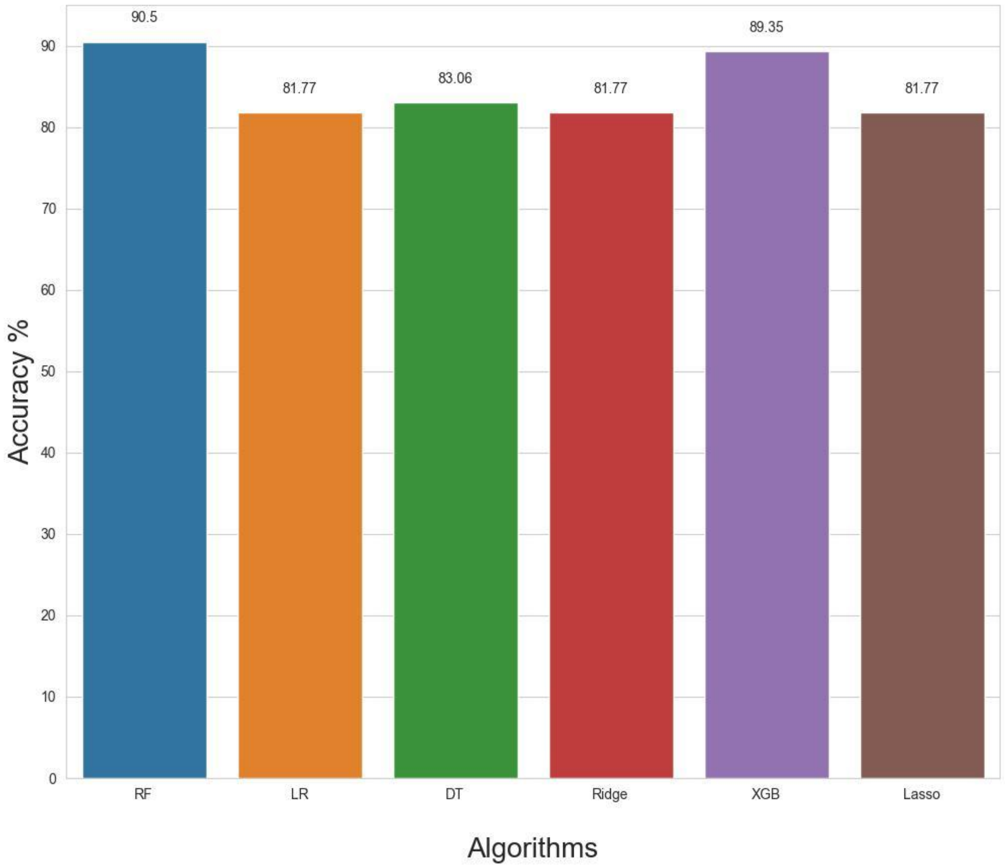

Fig 6.1 Comparison of R2 score

The bar chart presents a comparison of R² (accuracy) scores for different machine learning algorithms used in air quality prediction. The Random Forest (RF) algorithm outperforms the others, achieving the highest R² score of **90.5%**, indicating a strong predictive capability. XGBoost (XGB) closely follows with **89.35%**, suggesting it's also a highly effective model. Decision Tree (DT) achieves a moderate score of **83.06%**, while Logistic Regression (LR), Ridge, and Lasso regressions all report the same R² score of **81.77%**, reflecting relatively lower but consistent performance. Overall, ensemble methods like RF and XGB show clear advantages for this prediction task.

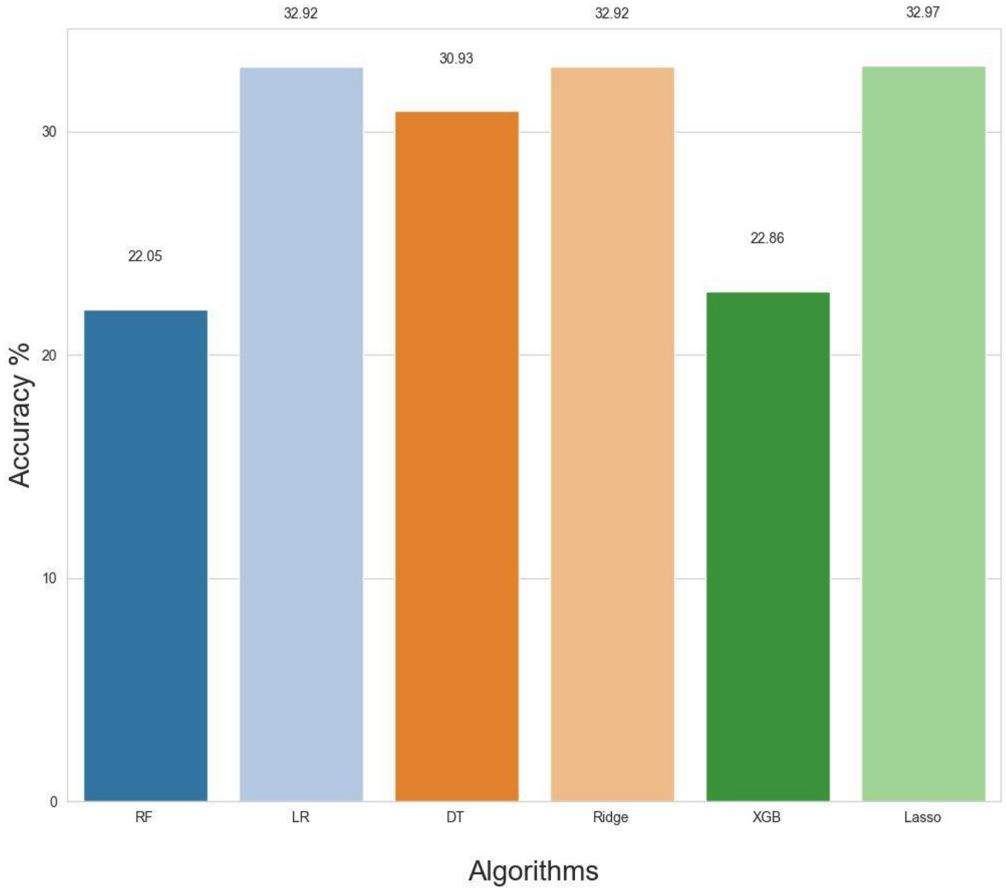

Fig 6.1 Comparison of Mean Absolute Error

The bar chart illustrates the comparison of Mean Absolute Error (MAE) percentages for various algorithms used in air quality prediction. Lower MAE indicates better predictive performance.

1. Random Forest (RF) and XGBoost (XGB) exhibit the lowest MAE values (22.05% and 22.86% respectively), implying they deliver the most accurate predictions with minimal error.

2. Lasso Regression records the highest MAE (32.97%), closely followed by Logistic Regression (LR) and Ridge Regression, both at 32.92%, and Decision Tree (DT) at 30.93%.

This suggests that while ensemble models like RF and XGB performed best in minimizing absolute prediction errors, linear models introduced relatively larger discrepancies between predicted and actual air quality values.

7. CONCLUSIONS AND FUTURE WORK.

Conclusion

The "Air Quality Prediction Using Machine Learning" project effectively demonstrates how data-driven techniques can address environmental challenges, particularly air pollution. By integrating historical air quality data with meteorological and geographic factors, the system accurately forecasts the Air Quality Index (AQI) for specific regions. Machine learning methods such as Random Forest, Support Vector Machines (SVM), and Linear Regression have shown strong capability in detecting pollution trends, identifying key contributors, and reliably predicting future AQI levels. The platform offers an intuitive interface, real-time data visualizations, and actionable predictive insights that assist environmental authorities, policymakers, and the public. This data-focused approach reduces uncertainty in environmental forecasting, enabling informed decisions regarding traffic control, industrial operations, and health advisories for vulnerable groups. Beyond forecasting, the system analyzes data to identify major pollutants and conditions that lead to air quality degradation, empowering proactive interventions rather than reactive measures. Its robust, flexible, and scalable design allows for easy updates and expansions as improved sensors and additional data sources become available.

Future Scope:

While the current system provides accurate predictions and valuable insights, future enhancements could include:

- IoT Integration: Incorporating real-time data from distributed IoT air quality sensors to enhance responsiveness and prediction accuracy.

- Advanced Deep Learning Models: Utilizing sophisticated techniques like Long Short-Term Memory (LSTM) networks or hybrid neural networks to better capture complex temporal patterns, especially for long-term forecasts.
- Mobile Application Development**:** Creating a mobile app that delivers real-time AQI alerts and personalized health advisories based on user location to increase public engagement and awareness.
- Geospatial Mapping: Integrating Geographic Information System (GIS) technology to generate dynamic AQI heatmaps for improved visualization of pollution distribution across broader regions.

REFERENCES

[1]. https://en.wikipedia.org/wiki/Air_quality_index

[2]. Kennedy Okokpujie, Etinosa Noma-Osaghae, Odusami Modupe, Samuel John, and Oluga Oluwatosin, "A SMART AIR POLLUTION MONITORING SYSTEM," International Journal of Civil Engineering and Technology (IJCIET), vol. 9, no. 9, pp. 799–809, Sep. 2018.

[3]. Kostandina Veljanovska and Angel Dimoski, "Air Quality Index Prediction Using Simple Machine Learning Algorithms," International Journal of Emerging Trends & Technology in Computer Science, vol. 7, no. 1, 2018.

[4]. D. Zhu, C. Cai, T. Yang, and X. Zhou, "A Machine Learning Approach for Air Quality Prediction: Model Regularization and Optimization," Big Data and Cognitive Computing, vol. 2, no. 1, p. 5, Mar. 2018.

[5]. A. Masih, "Machine learning algorithms in air quality modeling," Global Journal of Environmental Science and Management, vol. 5, no. 4, pp. 515–534, 2019.

[6]. https://archive.ics.uci.edu/ml/datasets/Air+quality

[7]. Rokach, Lior; Maimon, O. (2008). Data mining with decision trees : theory and applications. World Scientific Pub Co Inc. ISBN 978-9812771711.

[8]. BreimanL (2001)."RandomForests". MachineLearning. 45 (1):32. doi:10.1023/A:1010933404324