**Deep learning approaches for Air Quality Monitoring and Forecasting**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Submitted by | | |  |
| **Name:** | Arijit Ghosh | **Name:** | | Sajal Kumar Ghosh |
| **Roll No.:** | 11500122046 | **Roll No.:** | | 11500122036 |
|  |  | |
| **Name:** | Supratim Roy | **Name:** | | Swastik Tarafdar |
| **Roll No.:** | 11500122048 | **Roll No.:** | | 11500122050 |

Under the supervision of

Prof. Jahnabi Dutta

Academic Year: 2025 – 2026

**Bachelor of Technology in Computer Science and Engineering**



Department of Computer Science & Engineering

B. P. Poddar Institute of Management & Technology

137, V.I.P Road. Poddar Vihar Kolkata - 700 052.



*Maulana Abul Kalam Azad University of Technology*

**Project Synopsis**

**Title:** Deep learning approaches for Air Quality Monitoring and Forecasting

# Abstract / Project Definition:

This project proposes to leverage advanced deep learning approaches to enhance the accuracy and reliability of air quality monitoring and forecasting. We will explore various deep learning architectures, including Recurrent Neural Networks (RNNs) like Long Short-Term Memory (LSTM) and Gated Recurrent Units (GRU), as well as Convolutional Neural Networks (CNNs) for capturing both temporal and spatial features. Hybrid models, combining the strengths of different architectures (e.g., CNN-LSTM), will also be investigated to address the intricate relationships within multi-source environmental data. The project will involve comprehensive data collection from monitoring stations, incorporating diverse parameters such as PM2.5, PM10, NO2, SO2, CO, O3, and meteorological variables (temperature, humidity, wind speed/direction, pressure).

# Literature Review:

**Paper1:**

The application of machine learning (ML) techniques to air pollution prediction has been extensively explored in recent years due to the limitations of traditional monitoring systems, which are costly and spatially restricted. Multiple studies have investigated the viability of regression-based and ensemble ML models to estimate pollutant concentrations using diverse data sources, such as meteorological parameters, traffic data, and prior pollutant levels.Saithanu and Mekparyup (2014) utilized multiple linear regression in Thailand, incorporating meteorological and pollution data to estimate PM10. Similarly, Rybarczyk and Zalakeviciute (2017) enhanced PM2.5 predictions in Quito by progressively integrating traffic and meteorological parameters, eventually reaching an R² of 0.8. Support Vector Regression (SVR) has been employed in various contexts, such as CO estimation in Australia and O₃ forecasting in India, highlighting its flexibility in capturing nonlinear relationships (Chelani, 2009; Hu et al., 2016).Ensemble models like Random Forest and XGBoost have also demonstrated high accuracy and robustness. Kumar et al. (2020) found that combining ensemble algorithms yielded better PM2.5 forecasts in Delhi. Meanwhile, deep learning approaches, particularly LSTM models optimized with genetic algorithms, have shown promise in time-series pollutant forecasting (Drewil & Al-Bahadili, 2022) Some studies emphasized spatial generalization, as in Ghaemi et al. (2018), who used kernel-based SVR for AQI mapping in Tehran. Additionally, research leveraging satellite data and spatiotemporal kriging has further expanded the spatial resolution of pollution estimates (Zhan et al., 2018).

**Paper2:**

Air quality forecasting is vital for public health, enabling policymakers to mitigate pollution episodes. Méndez et al. (2023) surveyed 155 studies (2011–2021) across major scientific databases to analyze geographical focus, prediction targets, predictors, metrics, and models.Geographically, research aligns with pollution severity: China leads (>50% of studies on Beijing, Shanghai, Hangzhou), followed by India, Europe, North America, and Australia. East and South Asia’s high mortality rates drive academic output.Prediction targets center on the Air Quality Index (≈50% of papers) and PM₂.₅ concentrations (54 studies), reflecting their health impacts. Common predictors include pollutant levels (50%), meteorological variables—temperature, humidity, wind speed/direction, pressure, rainfall (35–50%)—and temporal/geospatial factors—hour, month, location (5–25%). Evaluation employs range-dependent metrics (RMSE/MSE in 68%, MAE in 46%) and percentage metrics (R², MAPE, accuracy each ≈20%), balancing comparability and interpretability.Deep learning (DL) dominates (77% of studies) over classical regression (37%). LSTM (36% of DL) and MLP (35%) prevail for temporal modeling; CNN, RNN, GRU, and encoder–decoder architectures appear less frequently but contribute spatial/sequential insights. Regression methods most used are Support Vector Regression (15%) and Random Forest (14%), valued for kernel flexibility and robustness. ARIMA and tree-based regressors serve niche roles.

# Expected Outcome:

By leveraging advanced machine learning and deep learning algorithms—such as random forests, adaptive boosting, and LSTM networks—the system can deliver precise and actionable forecasts, even with complex and high-dimensional data. These models not only improve the reliability and speed of air quality monitoring but also help optimize the selection of essential features, reducing costs and complexity in sensor deployments. Ultimately, this AI-assisted framework supports more effective public health decisions, smart city planning, and environmental policy by providing real-time, data-driven air quality insights

# Timeline:

|  |  |
| --- | --- |
| **Phase** | **Duration** |
| Literature Review | Week 1–2 |
| Dataset Acquisition & Preprocessing | Week 3–4 |
| Model Design & Training | Week 5–7 |
| Model Evaluation & Tuning | Week 8–9 |
| GUI Development (optional) | Week 10 |
| Final Report & Submission | Week 11–12 |

# PO/PSO Mapping:

# Methodology (if identified):

1. Data Collection: Gather diverse air quality data from sensors, satellites, or public datasets, including various pollutants, meteorological inputs, and location/time information.

2. Preprocessing: Clean the data by handling missing values, removing outliers, and normalizing features to ensure consistency and analytical accuracy.

3.Feature Selection & Engineering: Use statistical methods or AI techniques to identify the most relevant features, reducing noise and simplifying sensor requirements.

4.Model Development: Build predictive models with advanced machine learning and AI approaches (like CNN and LSTM).

5.Model Evaluation: Test the developed models using validation techniques to ensure high accuracy and reliable air quality forecasts.

6.Deployment: Integrate the optimized models into a system that can provide up-to-date air quality indices and pollutant level.

7.Continuous Improvement: Regularly update and refine models as more data becomes available, maintaining high accuracy and relevance over time.

# Unique Contribution / Novelty: (if identified)

# References: