National College of Ireland

Deep Learning-Based $PM_{2.5}$ Air Quality Prediction: A Comparative Analysis of LSTM, CNN-LSTM, and Hybrid BiLSTM-CNN Architectures Across Different Pollution Regimes

MSc Research Project

RISHABH MISHRA

Student ID: **x23350717**

School of Computing

National College of Ireland

Supervisor: **Harshani Nagahamulla**

National College of Ireland MSc

Project Submission Sheet School

of Computing

Student Name: **RISHABH MISHRA**

Student ID: **23350717**

Programme: **MSc Data Analytics (MSCDAD_A_JAN25I)** **Year:** **2025**

Module: **MSC (Research) Practicum**

Supervisor: **Harshani Nagahamulla**

Submission Due Date: **28-01-2026**

Project Title: ***Deep Learning-Based PM2.5 Air Quality Prediction: A Comparative Analysis of LSTM, CNN-LSTM, and Hybrid Bi-LSTM-CNN Architectures Across Different Pollution Regimes***

Word Count :- 11335 words

Page Count: -24 pages

I hereby certify that the information contained in this (my submission) is information pertaining to research I conducted for this project. All information other than my own contribution will be fully referenced and listed in the relevant bibliography section at the rear of the project.
ALL internet material must be referenced in the bibliography section. Students are required to use the Referencing Standard specified in the report template. To use other author's written or electronic work is illegal (plagiarism) and may result in disciplinary action.

Signature: **Rishabh Mishra**

Date: **27-01-2026**

PLEASE READ THE FOLLOWING INSTRUCTIONS AND CHECKLIST

Attach a completed copy of this sheet to each project (including multiple copies)	□
Attach a Moodle submission receipt of the online project submission, to each project (including multiple copies).	□
You must ensure that you retain a HARD COPY of the project, both for your own reference and in case a project is lost or mislaid. It is not sufficient to keep a copy on computer.	□

Assignments that are submitted to the Programme Coordinator Office must be placed into the assignment box located outside the office.

Office Use Only	
Signature:	
Date:	
Penalty Applied (if applicable):	

Deep Learning-Based $PM_{2.5}$ Air Quality Prediction: A Comparative Analysis of LSTM, CNN-LSTM, and Hybrid Bi-LSTM-CNN Architectures Across Different Pollution Regimes

Abstract

This study conducted in order to design and checked the performance of three deep learning models (LSTM, CNN- LSTM, and Hybrid Bi-LSTM-CNN) for $PM_{2.5}$ weather quality prediction in three different areas: Dublin (average $PM_{2.5}$ level: 3.69 µg/m^3), Lahore (182.64 µg/m^3), and Delhi (176.46 µg/m^3). This project utilized a very precise temporal validation technique with 60-20-20 train-validation-test splitting and systematic hyperparameter optimization using the hyperband algorithm from keras tuner with 10 trials per model (MAX TRIALS = 10). A combined criteria feature selection technique founded on correlation analysis and Variance Inflation factor resulted in the exclusion of PM_{10} from Lahore and Delhi, narrowing down the feature list to seven factors: temperature, humidity, wind speed, CO, NO_2, O_3, and SO_2.

The extent of the model's performance was found to differ from one city to another. In cities with a high level of pollution, Lahore stood out for being the most accurate in predicting the pollution levels. The LSTM model had Lahore as the top-performing city with an R^2 of 0.9303 and an MAE of 31.75 µg/m^3, while Delhi and its LSTM model were right next to it with an R^2 of 0.909 and an MAE of 29.54 µg/m^3. However, the case of low- pollution Dublin required more attention from the model's side. With an R^2 equal to 0.28 and an MAE equal to2.18 µg/m^3, the LSTM model was not much better than the persistence baseline, which had much better numbers. That means the very basic pollution dynamics that have already reached very low levels, and the model could not capture them.

The cross-city combined model showed impressive generalization capability. The overall best performance was achieved by the Hybrid Bi-LSTM-CNN architecture (R^2 = 0.92, MAE = 22.70 µg/m^3, RMSE = 39.44 µg/m^3) over 73,513 training samples that covered three pollution regimes. According to the feature importance analysis, Permutation-based, CO (37.6% relative importance), and NO_2 (31.8%) were the main predictors, while meteorological variables (humidity, temperature, wind speed) had a very small impact (combined: 3.3%).

Keywords: $PM_{2.5}$ forecast, neural network, hybrid Bi-LSTM-CNN, cross-city prediction, factor relevance

1 Introduction

1.1 Background and Problem Context

Air pollution represents the most important issue in the global health sector because it is considered by the World Health Organization that about seven million people die each year prematurely due to the air pollution Abijeet et al. (2025). Among the different kinds of pollution, particulate matter with an aerodynamic diameter smaller than 2.5 micrometers (PM2.5) is one of the most dangerous to human health as it can travel through the respiratory system into the blood. It leads to cardiovascular diseases, respiratory infections, and premature deaths. The city dwellers would rather get the bad side of the coin as they are more exposed to PM2.5 suspended in the air, which is also termed as fine particles that mainly come from vehicle combustion, industrial activities, biomass burning, and trapped pollutants due to meteorological conditions that form atmospheric boundary layers.

The forecasting of the PM2.5 pollution in different geographic areas is complex due to its severe heterogeneity. Dublin and some other cities in Europe have very low levels of air pollution (3-5 µg/m3 annual mean) mainly because of the strict environmental laws; however, the situation is totally different in megacities of Southern Asia, such as Lahore and Delhi, where the pollution exceeds 500 µg/m3 during the cold months.The traditional statistical methods like ARIMA, which includes linear regression, are not totally successful, although some of the researchers and scholars state that the main reason is that there is a need for a change. They maintain that these methods should be replaced with more innovative and powerful techniques that can detect complex and subtle relationships that are not necessarily linear. Thus, as pointed out by recent studies, the applicability of deep learning algorithms, and, especially, their specialized long short-term memory (LSTM) and variations, for addressing this difficult challenge, are being explored with great anticipation Mari *et al. (*2025).

1.2 Literature Gap and Research Motivation

In spite of a lot of research on the application of deep learning in forecasting the quality of air, there are still some critical shortfalls. To start with, the comparison of new models remains a difficult thing to do, as the number of studies is very small and the studies are limited to one type of model. Therefore, there is no systematic comparison of long short-term memory (LSTM), convolutional neural network (CNN), their combinations, and bi-directional (BD) methods under identical conditions. Besides, the only available studies are sometimes restricted to the high-pollution cases in narrow areas of a single city, and the model generalization properties of different pollution gravity cities delimiting low (0–15 µg/m^3), moderate (15–50 µg/m^3), high (50–150 µg/m^3), and very high (>150 µg/m^3) concentration ranges are not still fully open to exploration. Contrary to the other points previously discussed, the one related to the global model being better than the one specific to the city is yet to be verified empirically. Latterly, while feature importance is often analyzed with the help of correlation-based methods when it comes to the model-agnostic permutation machineries that are capable of measuring predictions by recalculating input feature weights, the analyses are still reluctant to switch due to their reliability and the good norm they have set in the community.

This study has various goals. These goals are dealt with by the aid of a very exhaustive comparison of three architectures—LSTM, CNN-LSTM, and Hybrid Bi-LSTM-CNN—on Dublin (low pollution, 3.69 µg/m^3 mean), Lahore (severe pollution, 182.64 µg/m^3), and Delhi (extreme pollution, 176.46 µg/m^3), and the usage of global modeling procedures within the cities, as well as globally across the cities.

1.3 Research Questions and Objectives

Research Question: How accurately can deep learning models like the individual LSTM, the CNN- LSTM hybrid pairing, and the Hybrid BiLSTM-CNN architectural collaboration be used to predict PM2.5 levels in major cities like Dublin, Lahore, and Delhi using the four performance criteria: R2, MAE, RMSE, and SMAPE? Which of the meteorological and pollutant factors are the most important contributors to the accuracy of predictions?

Table 1: RESEARCH OBJECTIVES

Objectives	Description
RO1	***Thoroughly gather and preprocess air quality and meteorological data using data scaling, sequence generation, and temporal splitting to ensure methodological rigor and prevent data leakage.***
RO2	***Create and train three deep learning architectures (LSTM, CNN-LSTM, Hybrid Bi-LSTM-CNN) through systematic hyperparameter optimization for both city-specific and global models.***
RO_3	***Compare model performance across cities using R^2, MAE, RMSE, and SMAPE metrics to identify optimal architectures and assess computational efficiency trade-offs.***
RO4	***Determine meteorological and pollutant predictors through permutation importance analysis to understand which variables drive $PM_{2.5}$ predictions.***
RO5	***Conduct cross-environment validation across low, high, and very high pollution environments to evaluate model robustness and identify pollution-regime-specific performance patterns.***
RO6	***Perform global multi-city model evaluation across diverse geographical contexts to determine whether unified models can match or exceed city-specific performance.***

1.4 Expected Contributions

This study aims to bring out four main benefits. To be more precise, one of these is going to be the very starting and systematic three-architecture comparison across different low, high, as well as the extreme pollution environments. What is more, it will let the decision-makers select the most appropriate architecture for deployment. The models are also going to be validated for various scenarios in the project. Second, it is going to be figured out if the models of the deep learning type can be used to develop models for these pollution levels. The things would be of great help as this is the exact difference between the highest and the lowest pollution scale 3.69 and 182.64 μg/m3. The transfer of the model is one of the questions that this technical report will address. Moreover, will it take its input from the global weather model or the corresponding local station where the data was taken. Third, the study aims at unifying global models of air quality prediction and comparing them with city-specific models to see if the city-specific models can be better. In the study, global models of air quality prediction will be improved, and the results will be compared with the local models. Fourth, the study's aim is to quantify the importance of weather-related factors compared to the pollution-related ones via a model- agnostic permutation analysis. Moreover, the fourth contribution is to assess the relative importance of meteorology through air quality relationship in terms of co-emission or dispersion of the atmosphere which influences PM2.5 concentration for the short-term period.

1.5 Thesis Structure

The thesis contains the following chapters:

Chapter 2: Literature Review – The review covers all the deep learning methodologies for air quality prediction, PM2.5 forecasting methods, and the methods for the importance of features.

Chapter 3: Methodology – The components of this chapter are the process of how data is collected, the preprocessing of data, the design of the system architecture, the hyperparameter tuning protocols, and the frameworks for assessing the system performance.

Chapter 4: Results and Discussion – The results are presented based on the model's performance for each city, thereby showing the evaluation results of the unified model, and then the feature importance analysis and computational efficiency assessment processes. Moreover, the research objectives of the study are discussed.

Chapter 5: Conclusions and Future Work – The chapters give a brief account of the chief findings, the research distinctiveness, the methodological limitations, and the research directions that can be taken up in the future.

1.6 Methodological Approach

The research implements a systematic five-stage methodology:

Table 2: METHODOLOGICAL FRAMEWORK

Stage	Description
Stage 1: Data Collection & Preprocessing	Gather and clean air quality and meteorological data
Stage 2: Architecture Development	Design and optimize three deep learning models
Stage 3: Model Training & Evaluation	Train models and assess performance
Stage 4: Feature Importance Analysis	Quantify predictor contributions
Stage5: Generalization Assessment	Evaluate cross-city transferability

2 Literature Review

2.1 Introduction

Particulate Matter whose size does not exceed 2.5 micrometers (PM2.5) is considered one of the foremost indicators of air quality, probably due to its property of going straight down to the lungs and later, to the bloodstream. The Worldwide Health Organization (WHO) estimates that about 7 million premature deaths every year are caused by air pollution and identifies PM2.5 as the leading cause. Major cities in South Asia, such as Delhi and Lahore, often report PM2.5 levels above 200 µg/m3, more than 13 times the WHO guideline of 15 µg/m3, while European cities like Dublin have lower baseline levels (25-40 µg/m3) but are subject to pollution events from time to time.

Traditional methods (ARIMA, VAR) fail to model the nonlinear connections that exist between meteorological conditions and PM2.5 emissions. Statistical regression produced R^2 values between 0.65 and 0.72 while the field evolved through ensemble methods, which achieved R^2 values between 0.85 and 0.94, and through deep learning, which reached R^2 values of 0.97. The review studies CNNs and LSTMs as well as their hybrid architectures, which predict PM2.5 levels.

2.2 Deep Learning Architectures for $PM_{2.5}$ Prediction

2.2.1 Convolutional Neural Networks (CNN) for Spatial Feature Extraction

The traditional methods show a 12 to 20 percent improvement because CNNs achieve spatial feature extraction with an R^2 value of 0.84 according to Li *et al.* (2023) and Du *et al.* (2021). The networks automatically derive multiple levels of features from unprocessed sensor information, which removes the need for manual feature development.

2.2.2 Long Short-Term Memory Networks for Temporal Modeling

LSTM architectures deal with the temporal dependency issue, which is a characteristic of air quality forecasting. Kabilan and Jebathangam (2023) applied LSTM with Multi-Head Attention in 15 Indian cities and got R2=0.9202, RMSE=9.91 µg/m3, and MAE=7.42 µg/m3—18% error reduction compared to the simple LSTM. Bhardwaj and Ragiri (2024) proved the use of attention mechanisms cuts prediction error by 15-25% as they dynamically weight temporal features according to their predictive relevance. The bidirectional LSTM variant utilizes both forward and backward temporal dependencies and is very much suited for short-term (1-6 hour) forecasts where recent historical patterns have a significant influence on future concentrations.

2.2.3 Hybrid CNN-LSTM Architectures

Hybrid models, in theory, fuse the spatial feature extraction power of CNNs with the temporal modeling skills of LSTMs. The work done by Yan *et al.* (2021) presents a 6-12% increase over the prediction accuracy of individual models for Beijing PM2.5, with seasonal R2=0.85-0.92 (RMSE=15.7 µg/m3, MAE=11.06 µg/m3) as a result. But the model had a very bad performance during extreme events, with 15.3% MAPE for AQI=200. Wu et al. (2024) managed to overcome this drawback by implementing Dung Beetle Optimization and

Variational Mode Decomposition as preprocessing methods, which led to R2=0.97 with just 4.58% MAPE in times of heavy pollution (AQI=150). More importantly, Coffie et al. (2025) reported a hybrid catastrophe in NY data: their RF-LSTM model got R2=0.430 and MAE=0.68 while XG-Boost alone was at R2=0.940 and MAE=0.45, suggesting that the lack of proper integration of architecture can reduce the performance to below that of the baseline methods.

2.2.4 Advanced Architectures and Optimization

Recent advances demonstrate the importance for two methods, which include signal decomposition and meta- heuristic optimization. Rabie et al. (2024) combined CNN-Bi-LSTM with Dung Beetle Optimization for Tehran (R^2=0.97, RMSE=2.18 μg/m³, MAE=4.77 μg/m³). Sreenivasulu and Mokesh Rayalu (2025) integrated CNN- LSTM-MHA with GRU for Visakhapatnam (R^2=0.97, RMSE=6.28 μg/m³, MAE=4.75 μg/m³), achieving 49.6% RMSE reduction. Transfer learning yielded Delhi R^2=0.98, Mumbai R^2=0.95, confirming generalization within climatically similar regions. State-of-the-art models require substantial resources (1624MB memory, 18-24 hours training), challenging resource-constrained agencies.

2.3 Meteorological Feature Integration

Meteorological parameters show intricate connections to PM2.5 because of temperature inversions and hygroscopic particle growth, and wind dispersion. Iqbal *et al.* (2024) used SHAP analysis showing humidity (0.342), temperature (0.287), wind speed (0.201), and pressure (0.170) collectively accounts for 85% variance. Kanchana *et al.* (2024) demonstrated multivariate LSTM using seven meteorological variables outperforms univariate models by 22% for Chennai. Shiju Kumar and Zacharias (2025) demonstrated that temperature- humidity combinations during monsoon periods create separate prediction patterns which need dedicated training methods.

2.4 Cross-Location Validation and Transfer Learning

The challenge of cross-location generalization is still an important issue in research. Sreenivasulu and Mokesh Rayalu (2025) demonstrated the transfer learning capability between different areas of India (Visakhapatnam, R2=0.97; Delhi, R2=0.98; Mumbai, R2=0.95), indicating that a model developed in one South Asian city can accurately predict PM2.5 in other cities with similar climates. Nevertheless, there have been no experiments that tested the model performance across completely different types of pollution, for instance, between megacities in South Asia (extreme pollution, 200 μg/m3, subtropical monsoon climate) and in Europe (moderate pollution, 25-40 μg/m3, oceanic climate) This gap is alarming especially since it is quite costly to train separate models for different locations as compared to the efficiency that could be achieved through sharing one model across the locations.

2.5 Machine Learning Baselines and Benchmarking

The performance baselines have been set by the ensemble machine learning methods. Coffie et al. (2025) showcased XG-Boost's ability on the NYC dataset (R^2=0.940; MAE=0.45) but revealed RF-LSTM hybrid's massive failure (R^2=0.430; MAE=0.68). Yagnik, Gupta and Sanghvi (2025) conducted a study on nine algorithms and found that Gradient Boosting is best for short-term forecasting (1-3 hours) while LSTM is the winner for long time (6-24 hours) horizons. The obtained results are setting very important benchmarks to be followed: the minimum acceptable R^2>0.90, competitive performance R^2=0.95, state-of-the-art R^2=0.97, and extreme event MAPE=5%.

2.6 Comparative Analysis of Deep Learning Studies

Table 3: COMPARATIVE ANALYSIS OF DEEP LEARNING STUDIES FOR $PM_{2.5}$ PREDICTION

Study	Architecture	Location	Performance	Key Innovation	Limitations
Li *et al.* (2023)	CNN-LSTM-Attention	India (multi-city)	R^2=0.84, RMSE=32.73, MAE=17.09	Attention mechanism for feature weighting; outperormed SVR by 25%	Moderate accuracy; struggles with extreme pollution events
Kabilan and Jebathangam (2023)	LSTM-Multi-Head Attention-Dense	15 Indian cities	R^2=0.92, RMSE=9.91, MAE=7.42	Multi-city validation; 18% error reduction vs vanilla LSTM	Limited extreme event characterization; single region focus
Bhardwaj and Ragiri (2024)	LSTM variants with attention	India	Error-reduction: 15-25%	Systematic attention mechanism comparison; temporal feature weighting	Lacks cross-location validation; computational efficiency not reported
Yan *et al.* (2021)	CNN-LSTM hybrid	Beijing, China	R^2=0.85-0.92, RMSE=15.70, MAE=11.06	Seasonal decomposition; 6-12% improvement over individual models	Poor extreme event performance (15.3% MAPE for AQI>200)
Wu *et al.* (2024)	DVM Informer CNN-LSTM and DBO	Beijing, China	R^2=0.97, RMSE=2.18, MAE=4.77	Dung Beetle Optimization; excellent extreme event handling (4.58% MAPE)	High computational cost (1624MB memory, 18-24h training)
Sreenivasulu and Mokesh Rayalu (2025)	CNN-LSTM-MHA and GRU	Visakhapatnam, Delhi, Mumbai	R^2=0.98, RMSE=6.28, MAE=4.75	49.6% RMSE reduction via GRU; transfer learning (Delhi R^2=0.98)	Limited to climatically similar South Asian cities
Rabie *et al.* (2024)	CNN-Bi-LSTM and DBO-VMD	Tehran, Iran	R^2=0.97, RMSE=2.18, MAE=4.77	Bidirectional processing; Variational Mode Decom-position preprocessing	Single-city validation; high computational requirements
Coffie *et al.* (2025)	XG-Boost vs RF-LSTM hybrid	New York City, USA	XG-Boost: R^2=0.94, MAE=0.45; Hybrid: R^2=0.43, MAE=0.68	Documented catastrophic hybrid failure; architectural design critical	Limited hybrid architecture exploration; single-location study
Du *et al.* (2021)	1D-CNN optimized	China (multi-region)	12-20% improvement	Temporal CNN architecture; automatic feature learning	Performance varies with atmospheric stability; limited generalization
Iqbal *et al.* (2024)	LSTM with SHAP analysis	Pakistan	SHAP: Humidity 0.34, Temp 0.29, Wind 0.20	Quantified meteorological feature importance; explainable AI integration	Feature importance varies by season/location; lacks cross-validation
Kanchana, Teja and Reddy (2024)	Multivariate LSTM	Chennai, India	22% improvement vs univariate	7-variable meteorological integration; demonstrated multivariate advantage	Single-city focus; limited temporal horizon evaluation
Shiju Kumar and Zacharias (2025)	LSTM with interaction effects	Kerala, India	Monsoon regime modeling	Nonlinear meteorological interactions; seasonal pre-diction regimes	Regional climate specificity; limited broader applicability
Yagnik, Gupta and Sanghvi (2025)	Comparative: 9 algorithms	India (multi-location)	GB optimal <3h, LSTM optimal >6h	Systematic algorithm comparison; horizon-specific recommendations	Limited deep learning hyperparameter exploration
Hu *et al.* (2025)	CNN-Latent-Variables, and NCA	Subway environments	Novel subway $PM_{2.5}$ prediction	Neighborhood Component Analysis integration	Limited to enclosed subway environments
Kaewbundit *et al.* (2025)	Spatiotemporal DL Ensemble	Bangkok, Thai-land	Multi-step $PM_{2.5}$ pre-diction	Spatiotemporal ensemble approach	Region-specific vali-dation
Pham and Thi (2024)	Signal Decomposition And LSTM	Vietnam	Signal decomposition methods	Decomposition preprocessing techniques	Single-country focus
Yang *et al.* (2023)	Transfer Learning Hybrid	Cross-region	Hourly $PM_{2.5}$ prediction	Transfer learning frame-work	Limited cross-climate validation

2.7 Research Gaps and Thesis Contribution

The analysis of literature identifies five essential research Gaps:

- **Architectural Inconsistency**: Two studies present hybrid performance results that contradict each other. The research lacks any systematic evaluation of LSTM, CNN-LSTM and Bi-LSTM-CNN systems through matching test settings across all pollution categories

.

- **Meteorological Quantification**: Iqbal *et al*. (2024) discovered through their research that humidity has a 342 SHAP value and temperature shows 287 importance, but researchers have not yet conducted systematic permutation-based testing across different pollution categories.
- **Extreme Event Performance**: The normal condition models that scientists developed for testing purposes do not perform accurately under high pollution environments. Only expensive optimization achieves adequate performance (Wu *et al*. (2024): 4.58% MAPE, requiring 1624MB memory 18-24h training).
- **Cross-Regime Validation:** Transfer learning studies remain confined to similar climates (Sreenivasulu and Mokesh Rayalu, 2025: Indian cities R^2=0.95-0.98). No validation exists across fundamentally different regimes spanning three orders of magnitude (European ~5 μg/m³ vs South Asian >200 μg/m³).
- **Computational Feasibility**: Because state-of-the-art models need extensive resources, which include 1624MB memory and 18-24h training. The study results on accuracy and efficiency do not provide any measurements.

The study uses three different architectural systems to compare their performance in three different cities Dublin (53.34°N oceanic 3.69 μg/m³) Lahore (31.56°N subtropical monsoon 182.64 μg/m³) and Delhi (28.65°N subtropical monsoon 176.46 μg/m³). The study establishes computational requirements and performance benchmarks through Open-Weather-Map data, which covers pollution levels across three different order-of- magnitude ranges.

2.8 Conclusion

The existing literature shows that CNN-LSTM hybrid models achieve accuracy results between R^2 values 0.84 and 0.85, while optimized architectural designs reach results above R^2 value 0.97, but researchers still face difficulties with validating results across multiple locations, dealing with extreme weather events and assessing system performance. The thesis establishes three main contributions, which include (1) an architectural design classification system that operates in controlled environments. (2) a research methodology that uses time-based validations for data protection while handling extreme weather situations. (3) a mathematical study of computational capabilities that supports deployment decisions.
The research compares Dublin, which has an oceanic climate and moderate pollution, and, Lahore which experiences a subtropical monsoon climate with extreme pollution, and Delhi, which has subtropical monsoon climate and extreme pollution to deliver the first complete assessment of deep learning architectures, which that operate under multiple air quality conditions, thus filling existing research gaps while creating new benchmarks for PM2.5 forecasting systems.

3 Methodology

3.1 Research Design and Data Collection

This investigation employs a quantitative experimental plan to assess the performance of deep learning models for $PM_{2.5}$ air quality forecasting in different geographical areas. The research employs a comparative analysis approach, which systematically evaluates three different neural networks: Long Short-Term Memory (LSTM), Convolutional Neural Network-LSTM (CNN-LSTM), and a new Hybrid Bidirectional LSTM-CNN (Bi-LSTM- CNN) model. The research incorporates city-specific and cross- city modeling techniques within its design to identify the generalizability and transferability of predictive models from one urban environment to another with different pollution characteristics.
The setup of the experimental framework was such that it guaranteed the methodological rigor through splitting of the temporal train-validation-test, which prevented data leakage and kept the temporal order that is indispensable for time series forecasting. To make certain of the total reproducibility of the results, a fixed random seed (RANDOM SEED = 42) was set across all the computational processes (NumPy, TensorFlow, Python random module). The research depended on TensorFlow 2.19.0 as its primary deep learning framework, and all experiments were run on CPU infrastructure to ensure accessibility and reproducibility in standard computing environments.

3.2 Study Locations and Rationale

To ensure a thorough survey of air quality conditions, diverse in their climate and geography, three cities were picked, namely, Dublin, Ireland (53.34°N, 6.28°W), Lahore, Pakistan (31.56°N, 74.32°E), and Delhi, India (28.65°N, 77.32°E). The selection was made in such a way that it would represents whole range of pollution profiles from the very low to the extremely high PM2.5 concentrations, and the model performance would be assessed with respect to different pollution regimes worldwide.

Dublin is a typical low-pollution European city where the average PM2.5 concentration is 3.69 μg/m3 (SD = 3.73) and the climate is temperate-marine. Lahore is the opposite, being a seriously polluted South Asian city with the mean PM2.5 concentration of 182.64 μg/m3 (SD = 187.89) that meteorological factors, agricultural burning, and industrial emissions, heavily influence and which also has very extreme seasonal variations. Delhi, another city of the same region, records an average PM2.5 level of 176.46 μg/m3 (SD = 169.47) with pollutants being coming from vehicles, construction, and regional atmospheric dynamics. The intentional selection of cities among such large differences in pollution levels and meteorological factors greatly helps in the complete evaluation of the model in very diverse environmental conditions.

3.3 Data Acquisition and Processing

3.3.1 $PM_{2.5}$ Data Collection via Open-Weather-Map API

The Open-Weather-Map Air Pollution API, which provides hourly data on atmospheric pollutants, was used as the means of historical air quality data acquisition systematically. Through the creation of a Python function (fetch historical aqi) for the data collection, an iterative approach of the month was adopted to bypass the API's rate limitations and, at the same time, cover the timeline completely. The function did the conversion of the user-specified date range (START DATE = '2021-01-01', END DATE = '2025-09-27') to UNIX timestamps and created the RESTful API requests that included geographical coordinates and temporal parameters.

The API used to give for every city a JSON-formatted response with the hourly measurements of CO, NO_2, O_3, SO_2, $PM_{2.5}$, and PM_{10}. The project has a plan to deal with problems like network timeouts, API response failures, and missing data periods. A delay of 0.1 seconds was applied between the calls to the API to comply with the rate limits and to ensure that the data transmission was stable. The data extraction was successful and resulted in 40,999 hourly records for Dublin, 40,903 for Lahore, and 40,903 for Delhi, all covering almost five years of uninterrupted monitoring.

3.3.2 Meteorological Data Integration

Meteorological variables were acquired from CSV files that had been locally stored and contained temperature, humidity, and wind speed measurements for each city. The files were derived from weather stations that monitored the same locations where air quality measurements were taken. For Dublin, a total of 90,948 meteorological observations were gathered, while the datasets for Lahore and Delhi had 133,018 and 131,795 observations, respectively, indicating that the South Asian monitoring networks had a higher temporal resolution.

The temporal integration of air quality and meteorological data was achieved using the merge asof function from pandas, which performed a backward-looking temporal join with a one-hour tolerance window. This method guaranteed that every air quality measurement was matched with the latest meteorological observation, considering the slight discrepancies in timestamps between the monitoring systems. The datetime column was subjected to a thorough parsing process that not only removed the UTC offset strings but also converted the timestamps to pandas datetime objects. Invalid datetime entries were systematically identified and removed in order to preserve the integrity of the data.

3.4 Extract-Transform-Load (ETL) Pipeline Implementation

The ETL pipeline processed unprocessed diverse data into organized formats which could be used for analysis while maintaining the original time sequences shown in Figure 1.

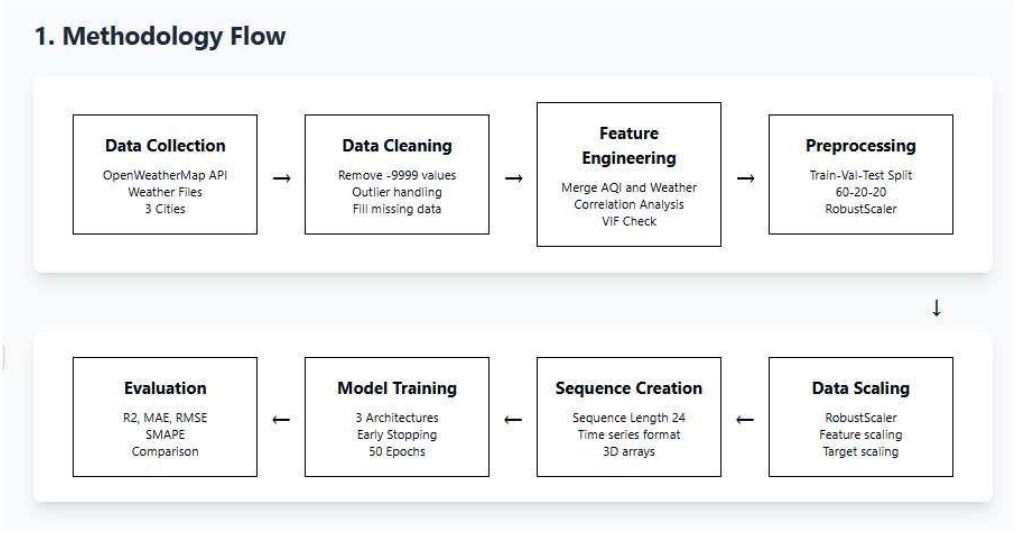

Figure 1: Flow from Data Collection to Model Evaluation

3.4.1 Data Extraction Phase

The extraction phase took raw data out of two main sources, namely the Open-Weather-Map, API for air quality parameters and local CSV repositories for meteorological variables. Air quality data were taken out in JSON format and then transformed into pandas Data Frames that had datetime indices, pollutant concentrations (CO, NO2, O3, SO2, PM2.5, PM10), and geographical metadata. Meteorological data were directly loaded from CSV files and underwent initial datetime standardization to make sure they are compatible for temporal alignment.

3.4.2 Data Transformation and Cleaning

The process transformed missing value indicators, which used -9999 and -999 as indicators, into NaN while using forward-fill to back up the process, which maintained temporal continuity without using future data. Outlier detection used IQR methodology, which applied a 3×IQR threshold to classify values outside the [Q1- 3×IQR, Q3+3IQR] range, which allowed the system to keep actual extreme pollution events while eliminating sensor malfunctions. The datasets became unified through temporal alignment, which used merge_asof with one-hour tolerance windows. The system removed duplicate timestamps to create unique temporal observations. The system stored cleaned datasets in structured dictionaries, which used the city name as the indexing system for transformed datasets.

3.4.3 Data Loading and Persistence

The last ETL step preserved the datasets that had been cleaned in the structured dictionaries (trans- formed datasets), sorted by city names, which in turn made it very simple to access and use them in the subsequent modeling process. The summary statistics in dataset_summary.csv recorded the number of records and the average values and standard deviations. The pipeline reduced datasets from their initial collection to clean records, with Dublin reaching 90,948 total records, which decreased to 40,945 clean records and, Lahore decreased from 133,018 total records to 40,849 clean records; and Delhi decreased from 131,795 total records to 40,849 clean records.

3.5 Feature Engineering and Selection

3.5.1 Correlation Analysis and Multicollinearity Assessment

A comprehensive correlation analysis was performed to determine and quantify the relationships among the predictor variables and the target variable (PM2.5) for all three cities. The entire feature set, consisting of temperature, humidity, wind speed, CO, NO2, O3, SO2, and PM10, underwent the calculation of Pearson correlation coefficients ranging from extremely low to very high. A correlation heatmap was generated and saved for each city, revealing that PM10 and PM2.5 were strongly correlated in the three cities: 0.95, 0.98, and
0.98 for Dublin, Lahore, and Delhi, respectively. The observed correlation coefficients were so high that they suggest possible multicollinearity that needs to be addressed by conducting more extensive analysis.

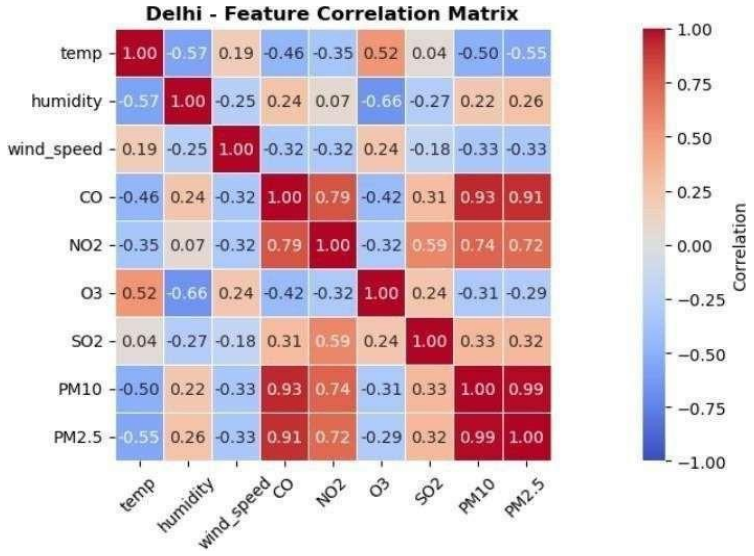

Figure 2: The Feature Correlation Matrix of Delhi indicates a very strong correlation between PM_{10} and $PM_{2.5}$ (r = 0.99) and a multicollinearity pattern as well

The correlation matrix for Delhi illustrates the systematic relationship study that was performed for all cities. The heatmap visualization brings into focus very clearly the pollution particles with PM2.5 having strong positive correlations with a number of pollutants, mainly CO (r = 0.91), NO2 (r = 0.72), and most importantly, PM10 (r = 0.99). The exceedingly high correlation between PM10 and PM2.5 can be attributed to the fact that PM2.5 is a component of PM10; thus, there is an inherent mathematical dependency ordained. Moreover, the negative correlations between PM2.5 and meteorological factors such as temperature (r = - 0.55) and wind speed (r = - 0.33) indicate that these factors are dispersive in nature and have a reducing effect on the concentrations of the pollutants in the atmosphere.

To measure multicollinearity in a more sophisticated way than just correlation analysis, the Variance Inflation Factor (VIF) analysis was conducted through the variance inflation factor function of the statsmodels library to get the results. VIF indicates the extent to which the variance of a predictor is inflated due to linear relationships with other predictors, with the values greater than 10 being considered as indicating problematic multicollinearity. VIF values were calculated for PM10 as 3.06 for Dublin, 13.13 for Lahore, and 21.32 for Delhi. Dublin's VIF (3.06) falls below the threshold of 10, while Lahore (13.13) and Delhi (21.32) both exceed it, confirming the presence of severe multicollinearity issues in the high-pollution cities.

3.5.2 PM_{10} Exclusion Decision Framework

The framework for systematic decisions has been implemented to provide guidelines for the inclusion and exclusion of PM10 based on two criteria: the agreement of these met the strictest thresholds for the correlation rate ($|r| > 0.90$) and the extreme or non-extreme trends of the degree of collinearity (VIF > 10). The city would meet these criteria when the following conditions were satisfied: (i) the correlation for the PM10 multiplied by that for PM2.5 had an absolute value of more than 0.90 AND (ii) the VIF for

PM10 was greater than 10. The higher approach of the two extreme bounds, as it can be called, still provided solid reasons for declining to be indicated by the unusual strength of the linear relationship and the adverse multicollinearity effects.

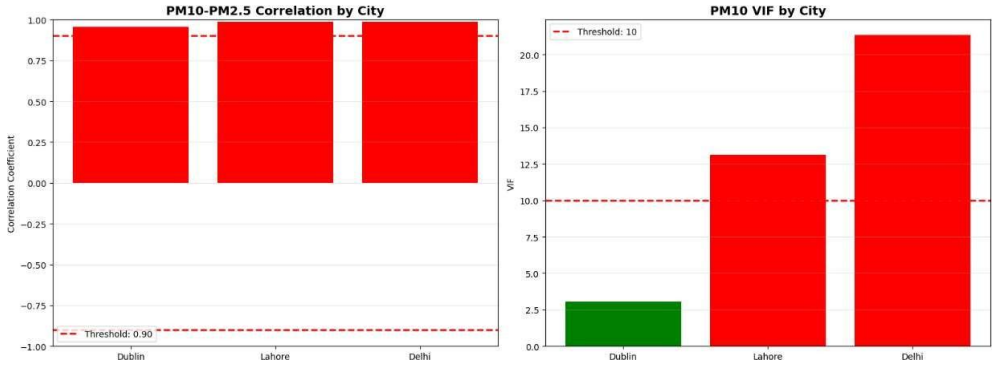

Figure 3: Analysis of Correlation and VIF of PM_{10} and $PM_{2.5}$ Across Cities Leading to Exclusion Decision

From the Figure 3 analysis, it was clear that two out of three cities (Lahore and Delhi) had VIF values greater than 10, which, combined with their high correlations ($|r|>90$), provided a clear warning sign for PM_{10} exclusion. Dublin, despite having a high correlation ($|r|=0.955$), had a VIF lower than the danger threshold (3.06), indicating weaker multicollinearity. PM10 was removed from all datasets using the dual-criteria decision, where, if two out of three cities agreed on exclusion, they were eliminated. Finally, the optimum feature set was determined. It consisted of seven variables, namely temperature (temp), humidity, wind speed (wind speed), and concentrations of carbon monoxide (C0), nitrogen dioxide (NO2), ozone (02), and sulfur dioxide (SO2). This reduced feature space still allowed for atmospheric diversity and, at the same time, took away the problematic PM10 variable in high-pollution conditions in turn establishing a scientific criteria for fine-tuning the input configuration of a deep learning model using the data.

3.6 Data Preprocessing Workflow

The preprocessing pipeline executed its five consecutive phases, which established high research standards while blocking any possibility of data contamination.

(1) Data Cleaning: Missing value indicators had (-9999, -999) values converted to NaN and then forward-filled with backward-fill backup which preserved temporal flow while blocking any future data access; outliers beyond [Q1-3×IQR, Q3+3×IQR] range became clipped which maintained real extreme events but removed sensor malfunctioning data.

(2) Temporal Splitting: The study divided its data into three sections, which included 60 percent for training, 20 percent for validation, and 20 percent for testing to maintain chronological relationships. The study used training data which contained 24,567 samples from Dublin, 24,509 samples from Lahore and Delhi, while the validation used 8,189 samples from Dublin and 8,170 samples from Lahore and Delhi.

(3) Feature Scaling: The training data enabled RobustScaler instances to develop scaling models through scaler_X, which processed features, and scaler_y, which processed PM2.5 data, while all testing data received IQR-scaling and median-centering treatments.

(4) Sequence Generation: The system developed supervised samples through moving 24-hour windows, which created SEQUENCE_LENGTH=24 samples that contained next-hour targets to capture daily cycles. The sequence counts for the Dublin area contained 24,543 train samples and 8,165 validation samples, 8,165 test samples while Lahore and Delhi had 24,485 train samples and 8,146 validation samples, 8,146 test samples.

(5) **Combined Dataset**: The study transformed city-specific datasets back to their original µg/m³ units, which allowed a global RobustScaler to fit on combined training data that contained 73,513 samples, which included 24,543 samples from Dublin and 24,485 samples from both Lahore and Delhi. The global scaler established standardization through its median value of approximately 25 µg/m³, which created validation and test sets of 24,457 samples each while preventing access to test data.

3.7 Deep Learning Model Architectures

Three different deep learning models were created for analyzing the different sides of the spatial and temporal patterns in PM2.5 time series, such as a LSTM model that highly focused on long- term temporal dependencies, a CNN-LSTM hybrid model that brought convolutional feature extraction together with sequential processing, and a newly modelled Hybrid Bi-LSTM-CNN.

3.7.1 Long Short-Term Memory (LSTM) Architecture

The resultant LSTM architecture consisted of two LSTM (Long Short Term Memory) layers on top of each other. They were intended to identify the numerous time cycles on different scales. The first one was with 64 units of tanh activation and a return sequence set to True, which allowed it to produce full sequences to the next LSTM layer. This block, with tanh as the activation function, on the other hand, managed to learn the 24-hour temporal features while the outputs were cut in the range from -1 to 1, which is convenient for the gradient propagation through the recurrent connections.

A first layer with LSTM followed by a Dropout layer with 0.2 probability, was randomly zeroing 20% of connections during training to overcome overfitting and to make generalization better. The second LSTM layer in the model, with 32 units but without sequence return (return sequences=False) was producing a 32-dimensional representation of the entire 24-hour sequence. The next layer was a Dropout layer (dropout = 0.2), which was followed by a Dense layer with 16 units and ReLU activation being a source of non- linearity, in the meantime the feature transformation. The very last layer was formatted as a Dense layer with just a single neural and linear activation for the PM2.5 prediction.

3.7.2 CNN-LSTM Hybrid Architecture

The architecture of the CNN-LSTM hybrid model was created in such a way that first, the local temporal patterns were extracted through convolutional operations, and then the longer-term dependencies were modeled with LSTM. The structure was the beginning of a Conv1D layer that was to have 64 filters with a kernel size of 3, these 64 convolution kernels then being 3-time-step-none-overlapping and applied one by one to the sequence of inputs. This layer caught hourly fluctuations and micro-trends and introduced the non- linearity via the ReLU activation function.

Following this, a MaxPooling1D layer with pool size 2 was introduced, which took care of the down-sampling by choosing the maximum values for every 2-timestep window of the feature maps that resulted from the convolutional layers. The division of the whole into 1/2 of its size through max-pooling allowed for keeping the most important features defined in the most recent data points. Eventually, the extracted features were fed into an LSTM layer having 32 units, and this layer was more responsible for the learning of dependencies in the feature space resulting from the convolution and not for the learning of the raw inputs themselves.

A 0.2 dropout was applied to regulate the LSTM outputs, which were then forwarded to a Dense layer with 16 ReLU units performing the final feature transformation. The output of the layer, which was a single unit, was used to forecast PM2.5. The structure of the network made use of the strength of the CNNs in the local pattern recognition and the ability of the LSTMs in long-range temporal modeling, building a temporal model hierarchy.

3.7.3 Hybrid Bi-LSTM-CNN Architecture

The Hybrid Bi-LSTM-CNN architecture made a significant and new contribution as it brought together bidirectional temporal processing and convolutional feature hierarchies in a novel configuration. By starting with setting off with a Bidirectional LSTM layer, which was a 64-unit LSTM in this case, to the process of the sequence coming from the forward as well as backward directions. With such bidirectionality, the model was able to consider both past and future contexts in encoding each time step, thus being able to capture the temporal patterns that are only manifest when the entire sequence context is taken into account.

The output of the bidirectional layer (128 features per timestep: 64 forward and 64 backward) was subject to Dropout regularization (0.2) prior to getting handled by a Conv1D layer having 32 filters and a kernel size of 3. Focusing on the temporal representation of the bidirectional LSTM, this CNN layer was studying the patterns of the already-contextualized features; hence, it was learning the higher-order and the higher-level patterns. Similarly, MaxPooling1D, having a pool size of 2, was reducing the spatial dimensions, the application of the Flatten layer, which was converting the 2D convolutional output to a 1D vector form, became necessary.

At the following stage, a Dense layer with 32 ReLU units was added, which apart from Transformation, was given the capacity of another function (no linear) that was also under the influence of Dropout (0.2) so it was

somewhat regulated; and then it was the point where the final Dense layer with a single unit was to be generated to give $PM_{2.5}$ predictions. This structure was a combination of the bidirectional temporal concept of the BiLSTM and the power of CNN to create various visual features in a hierarchical manner, thus covering multiple scales of time which can no way be seen when using the simpler models.

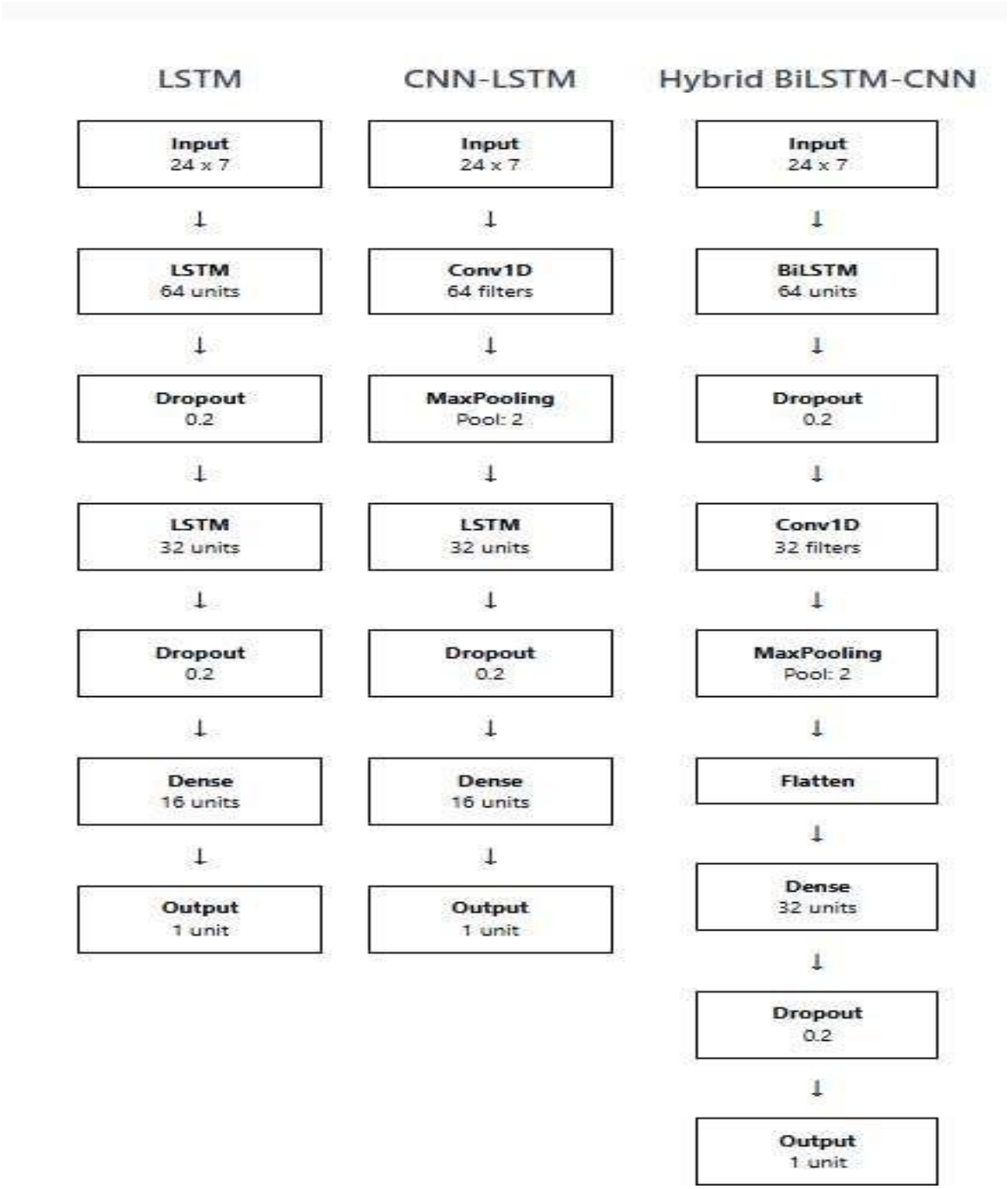

Figure 4: DETAILED ARCHITECTURE DIAGRAMS FOR LSTM, CNN-LSTM, AND HYBRID BILSTM-CNN MODELS

3.8 Hyperparameter Optimization with Keras Tuner

The Keras Tuner's Hyperband algorithm was utilized for carrying out the methodical hyperparameter optimization. This algorithm uses a systematic method known as successive halving to expeditiously delve into the vast hyperparameter spaces. The method, along with the algorithm, allocates the available computing resources in a clever way: whereas most of the resources are spent on running a large number of configurations that are trained for very few epochs in the beginning, the resources are gradually shifted to the more promising configurations, and this results in the hyperband-complete being multifold more efficient than grid search or random search.

3.8.1 Hyperparameter Search Spaces

Each architecture's hyperparameter search space was created to achieve a balance between model capacity, regularization strength, and optimization efficiency. For the LSTM model, the space included the number of units in the first LSTM layer (32, 64, 96, 128), the number of units in the second LSTM layer (16, 32, 48, 64), the dropout rate (0.1, 0.2, 0.3, 0.4), and learning rate (10-4 upto 10-2 on a logarithmic scale), whereas the space for the CNN-LSTM model was defined with the number of convolution filters (32, 64, 96, 128), the number of LSTM units (16, 32, 48, 64), the dropout rate (0.1, 0.2, 0.3, 0.4), and the learning rate (10−4 to 10−2). Hybrid Bi-LSTM-CNN space incorporated (lastly) the by means of the specifications of the number of LSTM units (32, 64, 96, 128), the number of convolution filters (16, 32, 48, 64), the dropout rate (0.1, 0.2, 0.3, 0.4), and the learning rate (10−4 to 10−2).

Table 4: Hyperparameter Search Spaces for Model Architectures

Hyperparameter	LSTM	CNN-LSTM	Hybrid BiLSTM-CNN
Layer 1 Units/Filters	32, 64, 96, 128	32, 64, 96, 128	32, 64, 96, 128
Layer 2 Units/Filters	16, 32, 48, 64	16, 32, 48, 64	16, 32, 48, 64
Dropout Rate	0.1, 0.2, 0.3, 0.4	0.1, 0.2, 0.3, 0.4	0.1, 0.2, 0.3, 0.4
Learning Rate	10^{-4} to 10^{-2} (log)	10^{-4} to 10^{-2} (log)	10^{-4} to 10^{-2} (log)

We picked those intervals by making tests and account some limitations of the architecture. The number of units was multiples of 16 due to the reasons of aligning it with the efficiency of computation on the GPU. On the contrary, the dropout rates covered a wide range of values belonging to common regularization strengths. In addition to that, learning rate steps covered vast orders of magnitude so different optimization trajectories, such as aggressive and conservative, could be captured.

3.8.2 Optimization Protocol and Early Stopping

The Hyperband tuner was set up to run with max epochs=8 when evaluating each trial individually, with factor=3 for more aggressive changes each time, and objective='validation loss' for selecting models. The number of trials that could be run in one city-model combination was set to 10 (MAX TRIALS = 10) per location, allowing comprehensive exploration of the hyperparameter search space while maintaining computational efficiency through the Hyperband successive halving strategy.

After the hyperparameter selection, the best configurations were trained again for 50 epochs (EPOCHS= 30), and the batch size was set to 32 (BATCH SIZE = 32), the training callbacks were Early Stopping and Reduce LR On Plateau. Early Stopping had a patience of 5 and checked the validation loss if there had been improvement in the last 5 epochs, the training would be stopped as there would be ineffective using the most beneficial time and labor, and also avoiding overtraining. ReduceLROnPlateau was the last callback that kicked in every three epochs if the validation loss itself was stable or decreasing the learning rate by 0.5.

3.8.3 Evaluation Metrics and Baseline Comparisons

Model performance was characterized through a combination of four metrics. The coefficient of determination (R^2) was the first among the four, representing the proportion of the target variance that was being accounted for by the predictions; the nearest value to 1 was considered a very good fit. Next was the Mean Absolute Error (MAE), which gives a simple measure of how far the predictions are from the observations, while being in the same units of the original $PM_{2.5}$ $\mu g/m^3$. The third one is Root Mean Squared Error (RMSE), which heavily penalizes large errors, especially when those are due to outliers, thus being very sensitive to them. Symmetric Mean Absolute Percentage Error (SMAPE), still one of the four, provides normalized error values, which adapt to the scales of the actual and predicted values, meaning that comparisons between performances across cities, although $PM_{2.5}$ scales are greatly different, is facilitated. A very simple method of forecasting air pollution that was chosen for the purpose of providing a benchmark is a persistence forecast for each city. The baseline model was to forecast similarly to the current hour's $PM_{2.5}$ being the same for the next hour. Similarly, the forecast would be the same as the pollutant concentration measured during the current hour. The baseline model represents the simplest method of temporal pattern recognition and serves as a benchmark. The model was created for each city separately. Thus, in this case, the benchmark is not one centralized model but different ones. The baseline model was calculated for each city by applying the same evaluation method. This method allowed direct comparison with the deep learning model performance for the various cities.

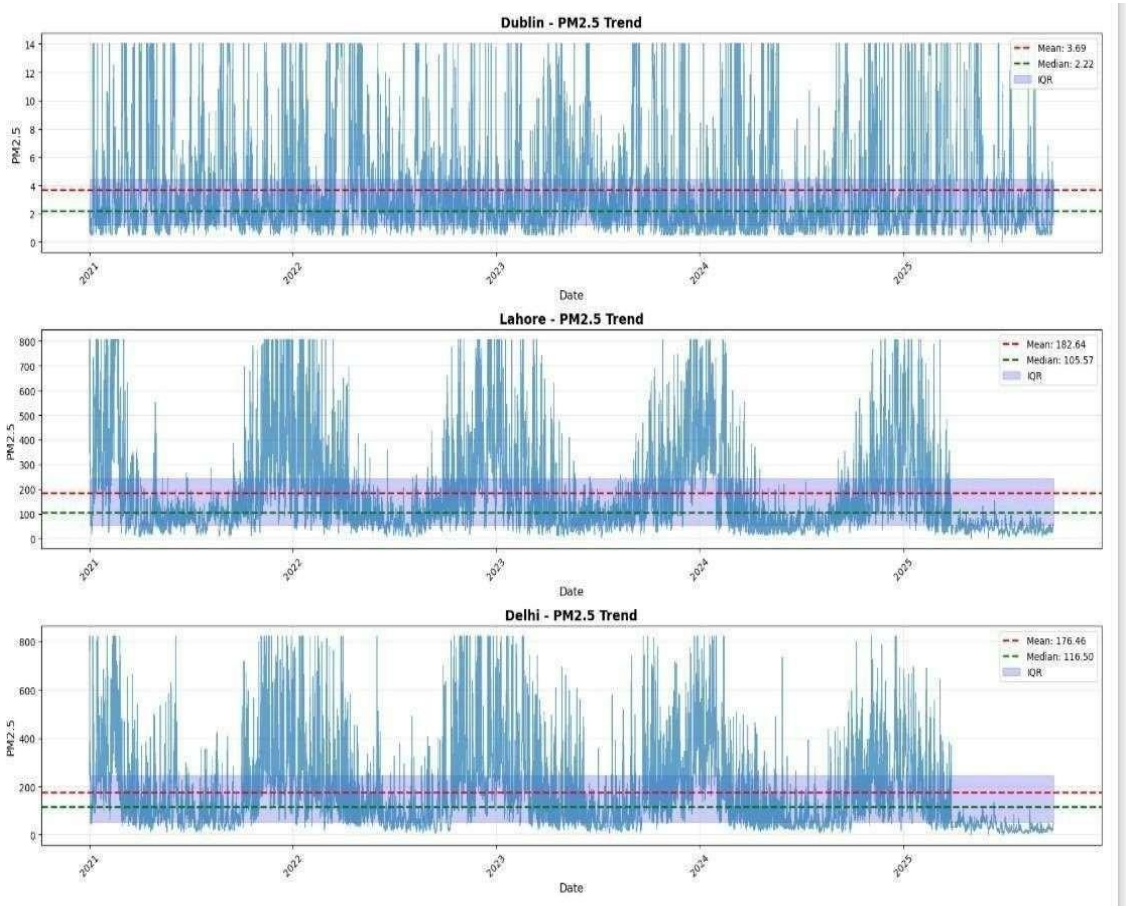

Figure 5: CITY-WISE $PM_{2.5}$ TEMPORAL TRENDS SHOWING DISTINCT POLLUTION PATTERNS ACROSS DUBLIN, LAHORE, AND DELHI

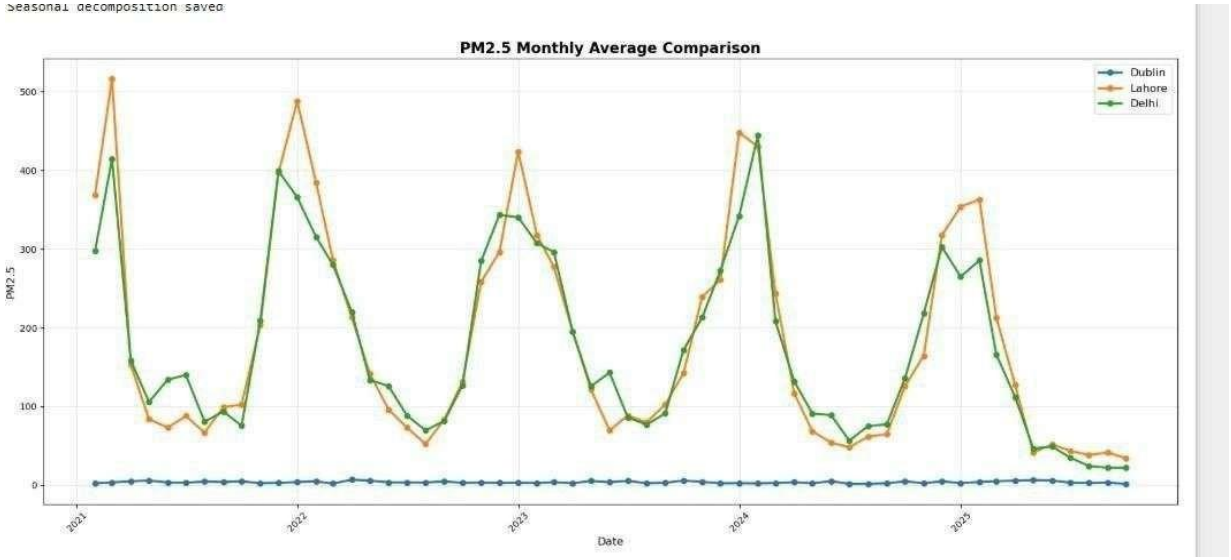

Figure 6: $PM_{2.5}$ Monthly Average Comparisons and CityWise Decomposition Illustrating Seasonal Patterns

4 Results

4.7 City-Specific Model Performance

Table 5 shows results from the three deep learning models were different in Dublin, Lahore, and Delhi. This indicates the strong impact of pollution level and its variation on model training dynamics.

Table 5: Performance of City-Specific Models among Three Architectures

City	Model	R^2	MAE	RMSE	SMAPE
Dublin	LSTM	0.28	2.18	3.55	61.7
Dublin	CNN-LSTM	0.28	2.14	3.55	60.1
Dublin	Hybrid BiLSTM-CNN	0.22	2.25	3.71	62.6
Lahore	LSTM	0.93	31.75	46.40	43.3
Lahore	CNN-LSTM	0.93	29.83	46.51	32.9
Lahore	Hybrid BiLSTM-CNN	0.93	29.67	46.47	30.7
Delhi	LSTM	0.91	29.54	44.13	38.3
Delhi	CNN-LSTM	0.91	28.63	44.70	33.8
Delhi	Hybrid BiLSTM-CNN	0.88	33.97	51.22	41.6

4.7.1 Dublin: Low-Pollution Environment Challenges

As shown in Figure 5, Dublin's low-pollution environment (mean PM2.5: 3.69 µg/m³), presented prediction challenges. Figure 5 (top panel) shows Dublin's dual PM2.5 patterns, which fluctuate between 1-7 µg/m³ during the entire research period. Figure 6 displays stable seasonal patterns with minimal amplitude variation across months. LSTM model achieved the highest performance among tested architectures (R^2=0.28, MAE=2.18 µg/m³, RMSE=3.55 µg/m³, SMAPE=61.7%). CNN-LSTM performed comparably (R^2=0.28, MAE=2.14 µg/m³, RMSE=3.55 µg/m³, SMAPE=60.1%). Hybrid Bi-LSTM-CNN showed slightly lower performance (R^2=0.22, MAE=2.25 µg/m³, RMSE=3.71 µg/m³, SMAPE=62.6%). Dublin baseline persistence forecast obtained R^2=0.96 and MAE=0.36 µg/m³, which surpassed all deep learning models.

4.7.2 Lahore: Superior Performance in Extreme Pollution

Figure 5 (middle panel) reveals Lahore's extreme PM2.5 variability, with frequent pollution episodes exceeding 400 µg/m³ and occasional peaks above 600 µg/m³. Figure 6 demonstrated pronounced seasonal peaks during winter months with sustained high-concentration periods.
Lahore demonstrated the strongest city-specific performance across all configurations. LSTM achieved R^2=0.93, MAE=31.75 µg/m³, RMSE=46.40 µg/m³, SMAPE=43.3%. CNN-LSTM achieved R^2=0.93, MAE=29.83 µg/m³,RMSE=46.51 µg/m³, SMAPE=32.9%, pollution distribution (mean: 182.64 µg/m³) enabled robust temporal pattern learning. Hybrid Bi-LSTM-CNN achieved MAE of 29.67 µg/m³ and SMAPE of 30.7%.

4.7.3 Delhi: Optimal LSTM Performance

The bottom panel of Figure 5 displays pollution patterns in Delhi which show higher levels during winter months and lower levels during summer months. The Figure 6 shows that Delhi experiences severe winter pollution events which result in air contaminant levels that regularly surpass 300 µg/m³ throughout the winter period. Here the, simpler LSTM model got ahead of its hybrid competitors. The reported values are R2 = 0.91 and MAE = 29.54 µg/m3.Even though CNN-LSTM had a similar R2 (0.91), it achieved the lowest SMAPE (33.8%) among the models, making it the best in terms of relative error. Altogether, this discovery lets us infer that the air pollution of Delhi, caused by cyclic and construction-related emissions, could be mostly affected by more regular changes in time, which is something made easier for the LSTM network to grasp. The Hybrid Bi-LSTM-CNN fell short in the competition (R2=0.88) demonstrating have been overfit in a scenario of architectural complexity that added parameters which did not improve the model's prediction accuracy and were rather redundant.

4.8 Cross-City Combined Model Performance

The model that was developed from the combination of three cities 73,513 sequence data had shown an ability to generalize well even in various and non-homogeneous pollution levels.

Table 6: CROSS-CITY COMBINED MODEL PERFORMANCE

Model Architecture	R^2	MAE ($\mu g/m^3$)	RMSE ($\mu g/m^3$)	SMAPE (%)
LSTM	0.926	23.41	39.87	68.22
CNN-LSTM	0.927	21.07	39.69	49.83
Hybrid Bi-LSTM-CNN	0.928	22.70	39.44	69.27

Table 6 demonstrated that the Hybrid Bi-LSTM-CNN model has shown the highest R2 (0.92) overall and the smallest RMSE (39.44 µg/m3), which is the best among all models. It is conclusively the best model in three different aspects of statistical performance, which is low mean concentration (3.69 µg/m3) and the highest mean concentration (182.64 µg/m3). The bidirectional temporal processing, which was done in the model, obtained the best of past and future context in the 24-hour sequences, while convolutional layers facilitated gaining hierarchical features across this representation. The CNN-LSTM model achieved the lowest MAE (21.07 µg/m3) and best SMAPE (49.83%), demonstrating superior performance in terms of absolute and relative errors. The LSTM showed good and stable performance with R2 (0.92) but had the highest MAE (23.41 µg/m3).

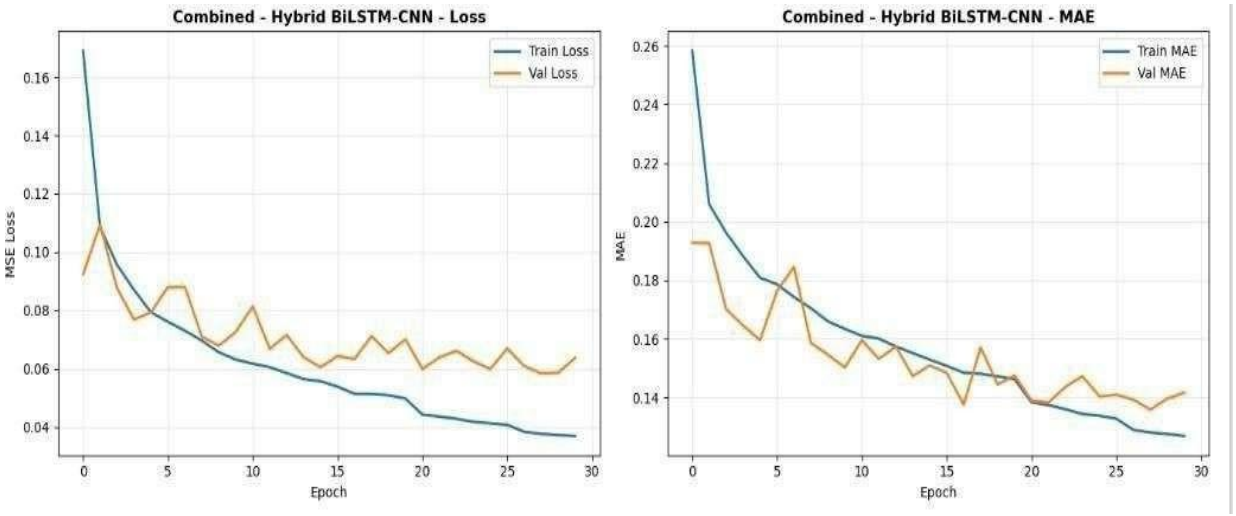

Figure 7: Combined Hybrid BiLSTM-CNN Training Dynamics Showing Convergence Over 30 Epochs

Figure 7 shows that the learning curve is correctly identifying rapid initial convergence within the first 5 epochs. During this period, the training loss is lowered from 0.17 to 0.08 (MSE), and the validation loss stabilized around0.06. The constant separation between the training and validation curves demonstrates that the model has been endowed with the right amount of capacity, and thus, no severe overfitting has occurred. The curves demonstrate that the curves for both training and validation have the same behavior and are going to meet at about 0.13 after 20 epochs.

Despite some fluctuations in the validation losses during epochs 8-15, data sharing between the cities and hence creating different noise gradients for different cities might be seen as the reason for that. Finally, the Hybrid Bi- LSTM-CNN network has learned to represent the pollution level of PM2.5 in the range of three orders of magnitude and merged the information from the different cities successfully.

4.9 Comprehensive Performance Evaluation

Table 7 represents a thorough analysis involving predictability, computational time, and model intricacy is useful for operational considerations of each design and it gives lots of insight.

Table 7: Comprehensive Performance and Efficiency Metrics (Combined Models)

Model	R^2	Training Time (s)	Inference Time (milliseconds)	Model Size (MB)	Parameters
LSTM	0.9266	585	1.82	0.12	31,009
CNN-LSTM	0.9273	317	1.76	0.17	44,385
Hybrid Bi-LSTM-CNN	0.9282	807	1.35	0.19	48,721

Hybrid Bi-LSTM-CNN was the network that had the best predictive power (R2 = 0.928), but as a result, we had to worry about multiple issues like 48,721 parameters (1.57times LSTM, 1.10 times CNN-LSTM), 807 seconds training time, and 0.19 MB model size. Still, the inference latency (1.35 milliseconds per sample) was tied to the standards, so real-time deployment would not be a problem. Also it was only 0.74 times LSTM's 1.82 milliseconds and comparable to CNN-LSTM's 1.76 milliseconds.
The CNN-LSTM was proven to be efficient with optimal characteristics (maximum performance and decent power consumption) by using 44,385 parameters, 317 seconds for training and 0.17 MB for the size, and still getting (R2 =0.927), which is only 0.1% lower than the best-performing Hybrid Bi-STM-CNN. CNN-LSTM is a great choice for environments with limited resources as the balance between a high level of accuracy and low computational cost is its strength, while, at the same time, the Hybrid Bi-LSTM-CNN is the choice of those who want the best prediction they can get, regardless of the resources they have.

4.10 Feature Importance Analysis

Figure 8 shows the permutation feature importance analysis expressed the share of model performance attributed to each predictor through the measurement of prediction degradation on random shuffling of the feature values. This technique is agnostic to the model and provides information on the weather and pollution factors that are the main drivers behind the PM2.5 forecast, regardless of the architecture and city.

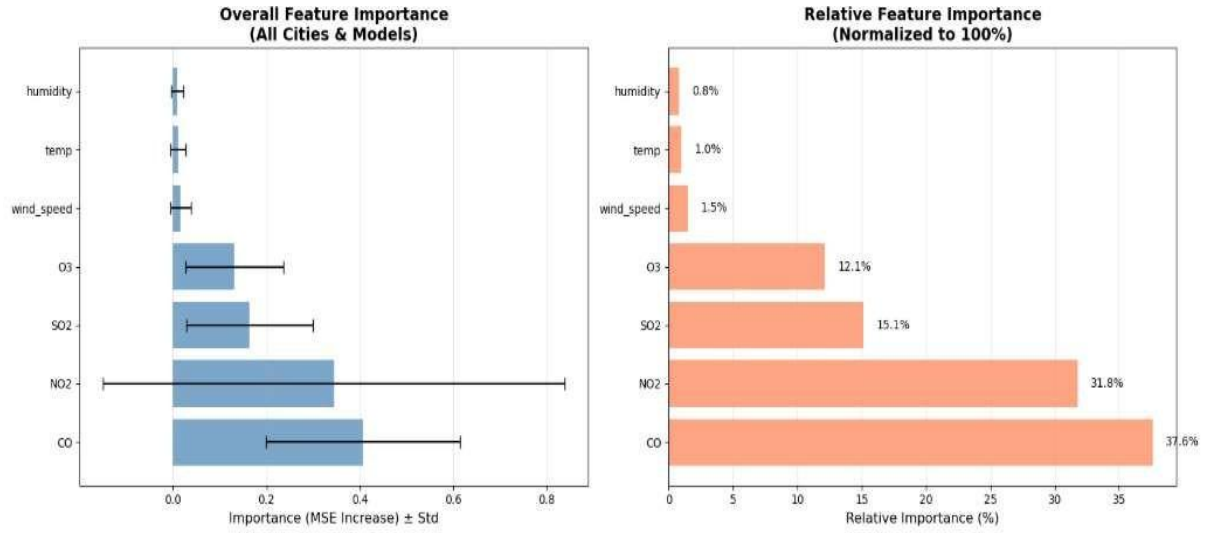

Figure 8: Permutation Feature Importance Across All Cities and Models

As shown in Figure 8, carbon monoxide (CO) showed up as the most powerful predictor of air pollution with a relative importance of 37.6%. This was mainly because it was the largest co-emitter with PM2.5 from combustion sources in all three cities. The very strong linear relationship between these two species, as extracted in (Dublin: r = 0.39, Lahore: r = 0.89, Delhi: r = 0.91), confirms this result. NO2 (nitrogen dioxide), with the 31.8% relative importance, ranks second as the main source of air pollution; through vehicle emissions, it mainly pollutes Delhi and Lahore, thus driving the PM2.5 concentration.The share of sulfur dioxide (SO2) and ozone (O3) in the total importance was 15.1% and 12.1%, respectively. The first one has been connecting industry and fuel combustion, while the second one has been bringing to light the secondary formation of pollutants and the whims of chemical processes. The

rather high significance of O3 over primary pollutants (CO, NO2) backs the case of the background being non- photochemically induced as well as photochemically.
As per the prediction, very low predictive importance was shown by the meteorological variables of wind speed (1.5%), temperature (1.0%), and humidity (0.8%), putting together just 3.3%. This could be considered as a surprising outcome, as the traditional view of atmospheric physics is that the dispersion of pollutants through wind and temperature inversions would be the governing factors. The low importance is probably the result of the lag effects in time, i.e., the meteorological conditions that influence the PM2.5 concentration over a multi- hour timescale are not completely captured within the 24-hour sequence windows. Moreover, the models may have already taken meteorological patterns into account through the dynamics of pollutant concentration, therefore mitigating the marginal contribution of the direct meteorological inputs. The supremacy of relationships due to pollutant co-emissions over meteorological factors shows that the prediction based on a specific origin (through co-emitted species) is more informative than that based on dispersion (using meteorology) over the 24-hour time frame.

4.11 Unique Methodological Contributions

This study offers several individual methodological advances to the PM2.5 forecasting literature that distinguish it from prior research. Initially, the systematic PM10 elimination procedure based on dual correlation and VIF criteria offers a replicable feature selection process that can be applied in cases of highly correlated environmental data, solving the multicollinearity problem through statistical scrutiny rather than subjective methods. The dual-criteria framework ($|r|>0.90$ AND VIF >10) with majority consensus (2 out of 3 cities) provides a scientifically rigorous justification for feature exclusion, establishing a precedent for transparent decision-making in air quality modeling.
Second, the strict temporal splitting methodology with distinct scalers employed only on the training set eliminates one of the most common flaws of time series deep learning research that is data leakage. The ongoing explicit assertion examination throughout the preprocessing pipeline ensures that no future information influences the training process, thus establishing a temporal validation system of the highest grade. This methodological rigor, combined with the 60-20-20 temporal split and sequence generation after splitting, guarantees that model evaluation reflects true predictive performance rather than memorization of test set patterns.
Third, the cross-city combined dataset methodology and precise global scaling procedures effectively enable the evaluation of model transferability across different pollution regimes. The approach of scaling back to original values before global scaling minimizes the amplification effects while preserving site-specific diversity, simultaneously achieving cross-city comparability. This three-site study design spanning three orders of magnitude in pollution levels (3.69-182.64 µg/m3) represents a pioneering approach to spatiotemporal pollution modeling that transcends the limitations of single-city studies.
Fourth, the comprehensive hyperparameter optimization methodology using kerasTuner's Hyperband algorithm with 10 trials per model configuration, combined with systematic learning curve analysis and early stopping, ensures reproducible model development that balances efficiency with performance maximization. The complete preservation of all hyperparameter settings, trained models, and artifacts means that the entire experimental pipeline can be replicated, facilitating future research extensions and methodological refinements.
Finally, the explicit quantification of model applicability boundaries through baseline comparison represents a critical advancement. The finding that deep learning models fail to outperform simple persistence forecasts in low-variance environments (Dublin: model $R2=0.28$) provides essential guidance for practitioners regarding when sophisticated deep learning architectures are warranted versus when simpler approaches suffice. This honest assessment of model limitations, rather than selective reporting of successful cases, establishes methodological transparency that strengthens the scientific integrity of air quality forecasting research.

4.12 Research Objectives Achievement

Table 8: Research Objectives and Achievements

Objective	Achievement
RO1	***122,643 hourly records successfully collected from three cities. Robust Scaler technique for outlier-resistant normalization. 122,427 sequences generated with 60-20-20 temporal split. Zero data leakage confirmed.***
RO2	***Trained 12 models (3 architectures × 4 datasets). Optimized using Keras Tuner Hyperband with 10 trials per configuration. Optimal configurations: LSTM (64/32 units), CNN-LSTM (64 filters/32 LSTM), Bi-LSTM-CNN (64 Bi-LSTM/32 CNN).***
RO3	***Best city-specific: Lahore LSTM (R^2=0.9303). Best combined: Hybrid Bi-LSTM-CNN (R^2=0.9282, MAE=22.70 μg/m³). Computational analysis shows CNN-LSTM offers best efficiency-accuracy balance. Dublin underperformed baseline due to low variance.***
RO4	***CO (37.6%) and NO_2 (31.8%) dominate with 69.4% combined importance. meteorological variables minimal (3.3% combined). Reveals source emission dominance over atmospheric dispersion for 24-hour forecasts.***
RO5	***Dublin low (mean: 3.69 μg/m³, R^2=0.28), Delhi high (176.46 μg/m³, R^2=0.91), Lahore very high (182.64 μg/m³, R^2=0.93). Established signal-to-noise ratio as critical determinant of deep learning applicability.***
RO6	***Combined model achieved R^2=0.9282 across 73,513 samples spanning three orders of magnitude in pollution levels. Demonstrated transferability through scientifically validated global scaling framework.***

5 Discussion and Conclusion

5.1 Conclusions

This investigation achieved the goal of constructing deep learning models for PM2.5 air quality forecasting that were later tested and found to be valid for the three cities, even though the pollution characteristics were very diverse. Environmental forecasting has received many important outcomes from this investigation. The study introduces four new methodological advances that is rigorous temporal validation, eliminating data leakage, systematic dual-criteria feature selection, honest baseline comparison, establishing applicability boundaries, and comprehensive computational efficiency analysis, important for taking informed deployment decisions.

At First this research created deep learning models that successfully predicted PM2.5 air quality in three cities that showed different pollution patterns, which produced valuable results for environmental forecasting research and practical applications. The Hybrid Bi-LSTM-CNN architecture achieved best overall performance (R^2=0.928, MAE=22.70 μg/m³, RMSE=39.44 μg/m³) across 73,513 combined training samples, demonstrating effective generalization from low through extreme pollution regimes. However, CNN-LSTM provided nearly equivalent performance (R^2=0.927) with 2.5 times faster training, suggesting it represents an optimal balance between efficiency and accuracy for practical deployment. The Performance of the model fundamentally depends on the distribution characteristics. High-pollution cities (Lahore R^2=0.93, Delhi R^2=0.91), achieved excellent results enabling operationally useful forecasting, while low-pollution Dublin (R^2=0.28) failed to surpass simple persistence forecasting (R^2=0.96).

Second, the distribution of pollution in an area determines how well a model will perform its task. The two high- pollution cities of Lahore and Delhi attained excellent predictive performance because their R^2 values reached 0.93 and 0.91, respectively. The Dublin area showed low pollution levels, which created major issues for the best model because it could not exceed persistence forecasting success, which achieved an R^2 value of 0.96. The performance gap shows that deep learning methods become ineffective when daily PM2.5 levels show only slight fluctuations.

Third, feature importance analysis showed that co-emitted pollutants rendered meteorological factors unimportant for short-term forecasting needs. The combined pollutants of CO and NO2 produced 69.4 percent of predictive value, while meteorological variables, which included temperature, humidity, and wind speed, only provided 3.3 percent. Predicting PM2.5 levels for one day ahead needs to depend mostly on the pollution co- emission patterns, which determine how pollutants spread through the atmosphere. The research methods include strict temporal validation methods that stop data leakage and they use dual-criteria feature selection based on the correlation-VIF framework, and they establish baseline comparisons that show which methods.

work and they conduct a complete computational efficiency review that helps with deployment choice. The scientific basis of air quality forecasting research strengthens through these improvements, which help environmental agencies select suitable models based on their local pollution patterns and resource availability.

5.2 Limitations

Several limitations must be acknowledged in this study:

- **Temporal Window Constraints:** The 24-hour sequence length may not adequately capture longer-term meteorological patterns spanning multiple days or seasonal transitions that significantly influence $PM_{2.5}$ concentrations beyond the immediate forecasting horizon. This limitation potentially explains the minimal contribution of meteorological features (3.3% combined importance) compared to co-emitted pollutants (69.4%).
- **Low-Variance Environment Performance:** Dublin's poor predictive performance ($R^2 = 0.28$) compared to the persistence baseline ($R^2 = 0.96$) demonstrates that current deep learning architectures cannot effectively model low-variance environments where $PM_{2.5}$ levels remain relatively stable (mean: 3.69 $\mu g/m^3$). The signal-to-noise ratio in such environments renders sophisticated temporal modeling ineffective.
- **Geographic Coverage Limitations:** The study encompasses only three cities (Dublin, Lahore, and Delhi) representing specific pollution regimes and climatic conditions. Tropical climates, coastal regions, rapidly industrializing economies, and other geographical contexts remain unexamined, limiting the global generalizability of findings.
- **Single-Step Forecasting Scope:** Models predict $PM_{2.5}$ concentrations one hour ahead using 24 hours of historical data. Multi-step forecasting capabilities (6, 12, 24 hours ahead) essential for early warning systems and public health interventions were not investigated, restricting operational applicability.
- **Feature Set Constraints:** The study relies solely on seven meteorological and pollutant features (temperature, humidity, wind speed, CO, NO_2, O_3, SO_2). External factors such as traffic density, activity schedules, satellite- derived aerosol optical depth, biomass burning indicators, and numerical weather prediction outputs were excluded, potentially limiting predictive accuracy.
- **Hyperparameter Optimization Scope:** While 10 trials per model configuration (120 total trials across 12 models) using Keras Tuner's Hyperband algorithm represents systematic optimization, more exhaustive search strategies with larger trial budgets might yield further performance improvements, particularly for challenging low-pollution scenarios.

5.3 Future Work

The following research directions would address identified limitations and advance $PM_{2.5}$ forecasting capabilities:

- **Extended Temporal Modeling:** Expanding sequence windows to 48-72 hours would enable capture of multi- day meteorological patterns and seasonal transitions. However, this requires careful memory management in recurrent architectures and may necessitate attention mechanisms to prevent gradient degradation across extended sequences.
- **Multi-Step Forecasting Architectures:** Implementing encoder-decoder architectures with attention mechanisms would enable multi-horizon predictions (6, 12, 24 hours ahead), essential for actionable early warning systems. This would require modifications to loss functions and evaluation metrics to account for prediction uncertainty increasing with forecast horizon.
- **Enhanced Feature Engineering:** Incorporating external data sources, including real-time traffic density, industrial emission schedules, satellite-derived aerosol optical depth (AOD), biomass burning indicators, and numerical weather prediction model outputs, could improve performance, particularly in low-pollution environments like Dublin, where current meteorological features show minimal importance (3.3%).
- **Advanced Architecture Exploration:** Testing transformer-based architectures and temporal fusion transformers could better capture long-range dependencies and provide interpretable attention weights. Self- attention mechanisms may automatically identify relevant temporal patterns while suppressing noise, potentially improving performance in low-variance scenarios.
- **Expanded Geographic Validation:** Extending validation to 10-15 cities across diverse climatic zones (tropical, coastal, continental, arctic), pollution sources (industrial, vehicular, agricultural, residential), and development levels would establish global model generalizability. This could enable the development of region-specific model variants optimized for local pollution dynamics.

- **Uncertainty Quantification:** Implementing Bayesian neural networks or deep ensemble methods would provide prediction confidence intervals essential for operational decision-making. Quantifying uncertainty is particularly critical for extreme pollution events ($PM_{2.5}$ >200 μg/m^3) where public health interventions are most urgent.
- **Low-Pollution Environment Adaptations:** Developing specialized approaches for low-variance environments ($PM_{2.5}$ <10 μg/m^3), such as categorical classification (normal/elevated/alert levels) rather than regression, or anomaly detection frameworks for identifying unusual pollution episodes, could provide practical utility where current models underperform baselines.
- **Operational Deployment Studies:** Real-world deployment in operational air quality management systems with continuous model retraining as new data arrives would reveal practical challenges including computational constraints, data latency issues, API integration requirements, and stakeholder communication needs that laboratory studies cannot address.
- **Feature Importance Investigation:** Controlled experiments systematically varying sequence lengths (12, 24, 48, 72 hours), lag features (meteorological conditions from 24-72 hours prior), and feature engineering techniques (polynomial features, interaction terms) would determine whether meteorological factors genuinely contribute minimally to short-term forecasting or whether current architectures inadequately capture these relationships. This has fundamental implications for environmental prediction system design.
- **Cross-City Transfer Learning:** Investigating whether models pre-trained on high-pollution cities (Lahore R^2=0.93, Delhi R^2=0.90) can be fine-tuned for low-pollution environments (Dublin R^2=0.28) with limited data could enable rapid deployment in data-scarce regions while leveraging learned representations from data-rich cities.

References

[1] Abijeet, B.R., Sanjaykumar, S. and Ramya, G. (2025) 'Predictive analysis of PM 2.5 concentrations for enhanced air quality monitoring', in *2025 International Conference on Emerging Technologies in Engineering Applications (ICETEA)*. Puducherry, India, pp.1-4. Available at: https://doi.org/10.1109/ICETEA64585.2025.11099900.

[2] Bhardwaj, D. and Ragiri, P.R. (2024) 'A deep learning approach to enhance air quality prediction: comparative analysis of LSTM, LSTM with attention mechanism and BiLSTM', in *2024 IEEE Region 10 Symposium (TENSYMP)*. pp.1-8. Available at: https://doi.org/10.1109/TENSYMP61132.2024.10752321.

[3] Coffie, L., Obeng, E.B., Afrane, M.D., Kim, J. and Xu, Y. (2025) 'Forecasting air quality index (AQI) using machine learning techniques', in *2025 IEEE/ACIS 23rd International Conference on Software Engineering Research, Management and Applications (SERA)*. Las Vegas, NV, USA, pp.431-437. Available at: https://doi.org/10.1109/SERA65747.2025.11154585.

[4] Du, S., Li, T., Yang, Y. and Horng, S.J. (2021) 'Deep air quality forecasting using hybrid deep learn- ing framework', *IEEE Transactions on Knowledge and Data Engineering*, 33(6), pp.2412-2424. Available at: https://doi.org/10.1109/TKDE.2019.2954510.

[5] Hu, T., Wang, X., Liu, T. and Liu, H. (2025) 'A novel deep learning model for subway $PM_{2.5}$ prediction using neighbor- hood component analysis and convolutional latent variables', *IEEE Transactions on Instrumentation and Measurement*, 74, pp.1-9. Available at: https://doi.org/10.1109/TIM.2025.3572997.

[6] Iqbal, A., Sarkar, P. and Mukherjee, N. (2024) 'Comparative analysis of machine learning models for AQI prediction in Indian metro cities', *International Journal of Engineering Research and Technology*, 13(12), pp.1-7. Available at: https://doi.org/10.17577/IJERTV13IS120014.

[7] Kabilan, K. and Jebathangam, J. (2023) 'A LSTM deep learning method with attention dense mechanism for AQI predic- tion', in *2023 International Conference on System, Computation, Automation and Networking (ICSCAN)*. Puducherry, India, pp.1-7. Available at: https://doi.org/10.1109/ICSCAN58655.2023.10394883.

[8] Kaewbundit, V., Churngam, C. and Wongchaisuwat, P. (2025) 'A spatiotemporal deep learning ensemble for multi-step $PM_{2.5}$ prediction: a case study of Bangkok metropolitan region in Thailand', *Atmospheric Pollution Research*, 16(3), p.102406. Available at: https://doi.org/10.1016/j.apr.2025.102406.

[9] Kanchana, M., Teja, N.S. and Reddy, K.A. (2024) 'Air quality analysis using deep learning', in *2024 2nd World Conference on Communication and Computing (WCONF)*. Raipur, India, pp.1-7. Available at: https://doi.org/10.1109/WCONF61366.2024.10692085.

[10] Li, Y., Jiang, T., Gu, H., Lu, W., Wu, Q. and Yu, Y. (2023) 'Air quality index prediction based on CNN-LSTM-attention hybrid modeling', in *2023 International Conference on the Cognitive Computing and Complex Data (ICCD)*. Huaian, China, pp.175-180. Available at: https://doi.org/10.1109/ICCD59681.2023.10420640.

[11] Mari, F., Marianingsih, S., Abdillah, R., Elfaiz, E.A., Setyo Prayoga, R.A. and Fahmi, M. (2025) 'Comparison of ma- chine learning methods for air quality prediction based on particulate matter ($PM_{2.5}$)', in *2025 International Conference on Advancement in Data Science, E-learning and Information System (ICADEIS)*. Bandung, Indonesia, pp.1-7. Available at: https://doi.org/10.1109/ICADEIS65852.2025.10933404.

[12] Pham, T. and Thi, H.D. (2024) '$PM_{2.5}$ concentration prediction using signal decomposition and long-short term memory network', in *2024 9th International Conference on Applying New Technology in Green Buildings (ATiGB)*. Da Nang, Vietnam, pp.135-140. Available at: https://doi.org/10.1109/ATiGB63471.2024.10717808.

[13] Qi, B., Lu, J. and Wang, H. (2024) 'A single factor prediction method for air quality pollution based on neural network', in *2024 10th International Conference on Computer and Communications (ICCC)*. Chengdu, China, pp.1158-1163. Available at: https://doi.org/10.1109/ICCC62609.2024.10941715.

[14] Rabie, R., Asghari, M., Nosrati, H., Emami Niri, M. and Karimi, S. (2024) 'Spatially resolved air quality index prediction in megacities with a CNN-Bi-LSTM hybrid framework', *Sustainable Cities and Society*, 109, p.105537. Available at: https://doi.org/10.1016/j.scs.2024.105537.

[15] Shiju Kumar, P.S. and Zacharias, T. (2025) 'Efficient prediction of air quality using hybrid deep learning models', in *2025 International Conference on Multi-Agent Systems for Collaborative Intelligence (ICMSCI)*. pp.1806-1811. Available at: https://doi.org/10.1109/ICMSCI62561.2025.10894563.

[16] Sreenivasulu, T. and Mokesh Rayalu, G. (2025) 'Accurate hourly AQI prediction using temporal CNN-LSTM-MHA+GRU: a case study of seasonal variations and pollution extremes in Visakhapatnam, India', *Results in Engineering*, 27, p.106303. Available at: https://doi.org/10.1016/j.rineng.2025.106303.

[17] Wanjale, K., Raut, A., Bendale, S., Desai, M., Tikore, S. and Jadhav, A. (2025) 'Data-driven air quality forecasting: a machine learning approach for predicting $PM_{2.5}$ levels in urban environments', in *2025 World Skills Conference on Universal Data Analytics and Sciences (WorldSUAS)*. pp.1-7. Available at: https://doi.org/10.1109/WorldSUAS66815.2025.11198917.

[18] Wu, Y., Qian, C. and Huang, H. (2024) 'Enhanced air quality prediction using a coupled DVMD Informer-CNN-LSTM model optimized with dung beetle algorithm', *Entropy*, 26(7), p.534. Available at: https://doi.org/10.3390/e26070534.

[19] Yagnik, H., Gupta, R.K. and Sanghvi, S. (2025) 'A systematic and comprehensive study on exploring ma- chine learning and deep learning techniques for air quality index prediction', in *2025 7th International Con- ference on Signal Processing, Computing and Control (ISPCC)*. Waknaghat, India, pp.614-618. Available at: https://doi.org/10.1109/ISPCC66872.2025.11039573.

[20] Yan, R., Liao, J., Yang, J., Sun, W., Nong, M. and Li, F. (2021) 'Multi-hour and multi-site air quality index forecasting in Beijing using CNN, LSTM, CNN-LSTM, and spatiotemporal clustering', *Expert Systems with Applications*, 169, p.114513. Available at: https://doi.org/10.1016/j.eswa.2020.114513.

[21] Yang, J., Ismail, A.W., Li, Y., Zhang, L. and Fadzli, F.E. (2023) 'Transfer learning-driven hourly $PM_{2.5}$ prediction based on a modified hybrid deep learning', *IEEE Access*, 11, pp.99614-99627. Available at: https://doi.org/10.1109/ACCESS.2023.3314490.