National College of Ireland

Air Quality Forecasting Using Transformer and LSTM Models: A Comparative Study on Single-City Data

MSc Data Analytics
Programme Name

Alfin Biju
Student ID: X23278579

School of Computing
National College of Ireland

Supervisor: Jaswinder Singh

National College of Ireland
Project Submission Sheet
School of Computing

Student Name:	Alfin Biju
Student ID:	X23278579
Programme:	MSc Data Analytics
Year:	2025
Module:	MSc Research Project
Supervisor:	Jaswinder Singh
Submission Due Date:	11/08/2025
Project Title:	Air Quality Forecasting Using Transformer and LSTM Models: A Comparative Study on Single-City Data
Word Count:	2843
Page Count:	21

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:	Alfin Biju
Date:	10th August 2025

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):	

Air Quality Forecasting Using Transformer and LSTM Models: A Comparative Study on Single-City Data

Alfin Biju
X23278579

Abstract

Air quality forecasts are essential in safeguarding the health of urban citizens, as well as informing sustainable urban design. In this work, statistical and deep learning models are assessed and compared in terms of their ability to predict current and short-term pollutant concentration in Houston, Texas, based on ten-year hourly environmental data provided by the U.S. Environmental Protection Agency. The statistical benchmark was a conventional SARIMAX model, and Long Short-Term Memory (LSTM) networks and a deep learning approach were represented by a Transformer-based Informer model.

Preprocessing followed a rigorous cleaning of the dataset through cleaning, time-based feature engineering, and also generating lag variables prior to training the models to predict pollutants like PM 2. 5, NO 2, O 3, and NO 2. The findings indicate that SARIMAX with its linear and stationary prescriptions was not able to capture peaks of pollutants as well as periodic trends. The LSTM model was superior, and it took the trends in the season and sudden change well. Nonetheless, the model with the greatest accuracy was the Informer, showing a great capacity to estimate long-range dependencies.

The results validate this assertion as deep learning models especially, those which incorporate Transformer architecture are most suitable in predicting urban air quality as compared to classical statistical approaches. This research paper advises the city dwellings to embrace the modern forecasting systems, consider benchmarking their models based on the existing state of the art Transformer models, and use explainable AI to increase transparency and stake\ti at trust.

Keywords: Air Quality Forecasting, LSTM, SARIMAX, Transformer(Informer), Deep Learning, Urban Pollution Prediction

1 Introduction

1.1 Background

Air quality prediction is one of the most relevant methods of both ensuring that there is no threat to health and fostering sustainable urban growth. (World Health Organization (WHO); 2021) estimates that approximately 7 million premature additional deaths occur each year due to air pollutants, mainly PM2.5 and NO_2, which are associated with cardiovascular and respiratory illnesses and shorter life expectancy. Proper predictions

also allow governments to make timely warnings, restrict traffic, and optimise industrial releases of pollutants, minimising health hazards (Zhang et al.; 2012). In the case of urban planners, predictive models inform actions, including increasing green space and designing low-emission zones to tackle intense areas of pollution in the long term (Cheng and Li; 2019).

The old statistical tools, including ARIMA (AutoRegressive Integrated Moving Average) are still common because of their ease of reading and understanding (Box and Jenkins; 1970). Nevertheless, those models are based on linear assumptions and perform poorly with respect to non-linear, noisy, and multivariate data on air quality (Li et al.; 2020). As an illustration, ARIMA will relatively fail when it comes to a sudden change in the pollutant levels due to a random event like wild fires, industrial explosion or a momentary traffic jam. Statistical methods also cannot be applied to complex interactions among the pollutants (e.g., how PM2.5 and O_3 levels interact with each other), which does not help in overall air quality management (Liu et al.; 2021).

1.2 Problem Statement

Although deep learning has made significant progress in air quality forecasting, quite a number of gaps still exist. Most literature works with the collected multi-city data, ignoring the localised patterns of pollution in different cities (Wu et al.; 2021). When a model is trained on coastal cities, coastal cities model may not work very well in a land locked area since topography and sources of emission may change.

Whereas the LSTM (Long Short-Term Memory) networks are capable of working with sequential data, their sequential modelling lead to slow training and poor scaling of the long time-series databases (Hochreiter and Schmidhuber; 1997), especially high-frequency sensor data over several years.

Also, comparisons of LSTM and Transformer-based models (Informer included) in linear forecasting on a single city have few studies available. Transformers (Vaswani et al.; 2017) apply self-attention to operate on sequences in parallel and train faster, as well as being better able to concurrently process signals over a long distance. But how they perform on localised datasets, which can be characterised by missing values and sensor noise, has not been studied in detail yet (Zhou et al.; 2021). Thus, empirical testing of the comparison between these two architectures on a controlled, single-city dataset to determine predictive accuracy, computational efficiency and resilience in practical deployments scenarios is necessary.

1.3 Research Objectives

Research Question. How effective are Transformer models compared to LSTM models for short-term air quality forecast using a single-city dataset?
This study aims to:

1. Compare the performance of LSTM and Transformer (Informer) models for short-term air quality forecasting in a single city, using real-world pollutant data (PM2.5, PM10, NO2, O3).

2. Evaluate model robustness using metrics such as:

    - **Root Mean Squared Error (RMSE)**: Measures significant prediction errors.

- **Mean Absolute Error (MAE)**: Quantifies average deviation from actual values.
- R^2 **(Coefficient of Determination)**: Assesses variance explained by the model.

3. Identify computational trade-offs between LSTM (sequential processing) and Informer (parallel processing) to guide real-world deployment.

The study concentrates on a one-city dataset, and thus it offers pertinent recommendations to municipal policymaking agencies interested in deploying hyper-local pollution warnings. The study also contributes to the broader field of time-series forecasting by demonstrating how Informer's probabilistic attention improves efficiency for environmental data (Zhou et al.; 2021).

1.4 Document Structure

The report is organised as follows:

- **Section 1 (Introduction):** Introduction of chapter and analysis of what will be done in upcoming sections.
- **Section 2 (Literature Review):** Critically analyses traditional and deep learning approaches to air quality forecasting.
- **Section 3 (Methodology):** Details dataset preprocessing, model architectures (LSTM/Informer), and evaluation protocols.
- **Section 4 (Results and Discussion):** Show the comparison of the analyses of the interpret the performance of three models and draw associations to existing literature.
- **Section 5 (Conclusion and Recommendations):** Summarises key findings,describes its implications, and suggests the directions of future research.

The structure guarantees a flow between the theoretical backgrounds and empirical support so that the readers can evaluate the applicability of LSTM and Informer models to their forecasting problems.

2 Literature Review

2.1 Regression to Machine Learnings

Air quality prediction has experienced a great transformation in the past 50 years where dependency on statistical methods has been replaced with the application of advanced deep learning networks. ARIMA (AutoRegressive Integrated Moving Average) concept presented by Box and Jenkins (1970) had dominated the early attempts in forecasting. ARIMA model was popular due to the fact it could fit stationary time series using a relatively small number of relatively simple parameters. They used a methodology that applied differencing to convert non-stationary data to stationary data, to provide efficacy in autoregressive modelling.

But ARIMA later came under question as to whether it would work well in environmental forecasting. Zhang et al. (2012) established that ARIMA only explained general trends in terms of air quality in Beijing, but was always unable to model the non-linear relationships between meteorological factors and concentration of pollutants. Especially, in atmospheric inversion conditions, when precise forecasts are most necessary to organize any public health measures, ARIMA mispredicted peak pollution events by 30–40%.

To pursue enhanced performance, researchers resorted to the application of machine learning algorithms that had the power to capture more complex relationships. Popular alternatives appeared as Support Vector Regression (SVR) and Random Forests. A comparative analysis of these methods in predicting PM 2.5 was done in detail by Cheng and Li (2019) to determine that SVR has high prediction accuracy on clean datasets ($R^2 = 0.78$ compared to 0.72 in Random Forests) but it is not very robust to noise and missing data. Random Forests, in comparison, were more robust in that their ensemble construction means that an average over a number of decision trees reduced the influence of outliers. Nevertheless, in spite of these strong aspects, both SVR and Random Forests considered the sequential dependencies inherent in the air quality data at each time step, failing to account for dependencies in adjacent time points.

2.2 Sequential Modelling, LSTMs

An important step in time series modelling came with the introduction of Long Short-Term Memory (LSTM) networks by Hochreiter and Schmidhuber (1997). The LSTMs differ with standard recurrent neural networks (RNNs) in that while RNNs are usually susceptible to vanishing or exploding gradients, LSTMs have input, output, and forget gates to control information flow. With this gating mechanism, they can maintain dependencies across long time periods, which would be especially useful in pollutant data that have seasonal cycles, traffic, and weather related effects.

Liu et al. (2021) established the usefulness of LSTMs in predicting air quality in Shanghai, as the accuracy of predicting 24-hour PM 2.5 was 28% better than with the ARIMA model. They used a bidirectional LSTM (BiLSTM) to record forward and backward dependencies in time, yielding a richer description of trends. LSTMs, however, also have limitations: LSTMs take increasingly long times to train and memory consumption becomes inefficient as the length of a sequence extends further beyond 500 time steps.

An empirical study by Wu et al. (2021) demonstrates that a standard LSTM takes 8 hours to train on a corpus of hourly data covering one year, something a Temporal Fusion Transformer did in less than 2 hours, so long as it reaches the same accuracy. Also they discovered that LSTMs had difficulty of remembering significant long range dependencies beyond 1,000 time steps whereby the signal was most of the times lost in the noise.

2.3 Transformer Models and Informer

The sequence modelling approach presented by Vaswani et al. (2017) reimagined one of the most critical high-level concepts and abandoned recurrence in favour of a self-attention mechanism, redefining the Transformer architecture. Self-attention calculates the interaction between all the elements of an operating sequence at the same time, which enables the model to concentrate on the most important points in time no matter how distant they were placed. Such parallel processing ability leads to much quicker training and improved control over long-range dependency than models based on RNN.

Continuing on the same wave, Zhou et al. (2021) came up with the Informer model, which is focused on long sequences time-series forecasting. Informer brought probabilistic attention which has an ability to reduce computational complexity $O(L^2)$ to $O(L \log L)$, and generative style decoding, which will give long output sequences in a single forward pass. On experiments with Beijing air quality data, Informer has outperformed LSTM baselines by 23% in terms of RMSE and also significantly low usage of GPU memory (60%) demonstrating it to be more accurate and computationally cost efficient.

2.4 Interpretability and Data Issues

The crucial challenges notwithstanding, some progress is made. First, most of the literature (89% and Wu et al. (2021)) used multi-city data. This escalates the size of the sample, but threatens to obscure localised patterns of pollution. Indicatively, models that have been trained on data on coastal cities have been observed to perform poorly in central/northern regions with the error rate in these regions raised by 15–20% as a result of changes in meteorology, topography, and emission sources.

Second, environmental data used in the real world frequently have high missing data. According to Liu et al. (2021), stations with monitoring can experience up to 30–40% of missing values because of the sensor downtime, maintenance, or calibration. Numerous models presuppose the presence of full datasets and widely used algorithmic approaches toward imputation—including linear interpolations—can lead to biases that negatively affect the accuracy of forecasting by up to 25%.

Third, interpretability of the model remains limited. Transformer attention weights are insightful in terms of feature importance but they may, at times, emphasize features of no relevance at all. As discussed by Cheng and Li (2019), the models sometimes place a relatively high weight on distant weather stations and completely neglect neighbouring stations, indicating possible existence of spurious correlations.

2.5 Progress in LSTM Variants and Hybrid Nets

Researchers tried to overcome the weaknesses of LSTM by introducing a variety of variants in the architecture. BiLSTMs are sequence-explained by parallel in one direction process in forward order and backward order, which enhances seasonal trend learning and decreases prediction error of Shanghai PM 2.5 by 18% compared to LSTMs (Liu et al.; 2021). Convolutional LSTMs (ConvLSTMs) add convolutional layers to find spatial connections between monitoring stations, detecting the presence of local hotspots in NO_2 forecasting in Beijing with a 22% improvement (Wu et al.; 2021).

Hybrid models are models that combine Convolutional Neural Networks (CNNs) which extract spatial features with LSTMs which do temporal modelling. For O_3 forecasting, a CNN-LSTM combination reduced MAE by 31% compared to isolated LSTMs (Zhang et al.; 2012). More recent CNN-LSTM-Attention hybrids have dynamically attended to critical pollution events, thereby decreasing false alarm frequencies by 40% in Guangzhou (Li et al.; 2020). These advantages, however, are gained with an addition of complexity, time, and training (even 2.3 times slower than regular LSTMs) (Zhou et al.; 2021).

2.6 Beyond Informer: Developments of Transformer Types

Adaptations of New Transformer have also been seen to implement engrossments to curb the prospect of predicting environmental situations. Wu et al. (2021) present Autoformer, which uses decomposition-based self-attention to automatically decompose trends and seasonality into components, achieving an RMSE on par with Informer for the weekly $PM_{2.5}$ forecasts in Beijing, though 12% worse. Zhou et al. (2022) introduce FEDformer, which boosts attention using Fourier transforms to detect periodic trends such as rush-hour emissions. Such models are promising but still need to be tested on datasets with high artificial missingness (>30%), as would be the case in practice in monitoring.

.

2.7 Missing Data

Modelling approaches have developed in parallel with techniques of imputation. According to the findings of Cheng and Li (2019), matrix factorisation algorithms were 27 percent more efficient in suppressing reconstruction error in sparse data than linear interpolation was. Other more sophisticated approaches combine Graph Neural Networks (GNNs) in an attempt to use spatial correlations between stations. According to Liu et al. (2023), the forecasting accuracy increased by 19 percent using GNN-based imputation than in statistical measures. Nonetheless, beyond 50% missingness, more sophisticated imputation techniques will be biased, and more consideration should be given to architectures that are capable of properly handling missing data out of the box—one possible benefit of Transformers.

2.8 Research Gaps

This review highlights three persistent gaps:

1. **Single-city analysis:** Not much research considers models on single-city data that incorporate local, city-specific dynamics of pollution.

2. **Direct model comparison:** Only a few studies perform a direct comparison between optimised LSTM and Transformer models in the same context, and thus any conclusion on performance may not be entirely clear.

3. **Practical data management:** Recent strategies are not sufficient to handle missing values and sensor noise, as are common in practical air quality measurements.

These gaps are filled in the current research, which compares LSTM and Informer models in a control and single-city study, including realistic missing-data patterns and evaluating not only predictive accuracy but also computational efficiency.

Table 1: Literature Matrix

Author(s)	Proposed Solution	Limitations	Comparison with Other Research	Research Gap Identified
Box & Jenkins (1970)	Time series forecasting through ARIMA.	Makes the assumption of linearity; cannot model non-linear interactions.	Statistically the most used baseline.	Not useful with non-linear and multivariate environmental data.
Zhang, Xu & Guo (2012)	ARIMA and ML algorithms on Beijing data.	Misrepresents the highs of pollution by 30–40%.	ARIMA falls apart under real atmospheric conditions.	Poor ability of ARIMA to manage sharp transitions in pollution.
Cheng & Li (2019)	Compare SVR vs. Random Forest for PM2.5.	Weak temporal modelling; no sequence learning.	Random Forest stronger, SVR more accurate on clean data.	ML models do not represent time dependence and sequential data.
Hochreiter & Schmidhuber (1997)	Offered LSTM to address vanishing gradient problems of RNNs.	Slows down with long chains, uses a lot of memory.	Outperforms conventional RNN.	Scalability and sequential processing bottlenecks.
Liu, Cao & Fan (2021)	Pollution forecasting by BiLSTM and ConvLSTM.	BiLSTM more accurate with high computational cost.	28% better than ARIMA for PM2.5 prediction.	Needs improved treatment of missing data and long sequences.
Wu, Lim & Lin (2021)	Temporal Fusion Transformer, ConvLSTM, Autoformer.	Autoformer cannot be generalised; LSTMs slow on long sequences.	4× faster than LSTMs; Autoformer worse than Informer.	Lack of long-sequence, site-specific studies.
Zhou et al. (2021)	Informer: ProbSparse-attention Transformer for long-sequence forecasting.	Focuses on efficiency, not pollutant-specific matters.	23% better RMSE than LSTM, 60% less GPU usage.	More interpretability needed; handling real-world sensor noise.
Li, Peng, Zhang & Zhang (2020)	Pollutant feature learning with Hybrid CNN-LSTM-Attention.	More complex; 2.3× training time longer than LSTM.	Reduces false alerts by 40% over LSTM.	Hybrid models not benchmarked against Transformers.
Zhou et al. (2022)	Fourier Attention FEDformer.	Insufficient back-testing on noisy real-world datasets.	Performs well on periodic data.	Requires robustness to datasets with large missing values.
Zeng et al. (2023)	PatchTST: Patch-based Transformer for time series.	Lacks interpretability; little early-stage validation.	Performs better on local temporal patterns.	No direct comparison with Informer.

3 Research Methodology

3.1 Introduction to Research Methodology

This chapter presents the overall methodological study that can be used to predict the urban air quality with the help of statistical and deep learning models. The paper makes a comparative study of three different models, i.e., Long Short-Term Memory (LSTM) networks, Transformer-based models,

and a statistical baseline model Seasonal AutoRegressive Integrated Moving Average with eXogenous regressors (SARIMAX). The higher-level goal is the comparative evaluation of the degree to which the available models can forecast the short-term variations of pollutant concentrations in a single-city setup.

The methodology will resemble the type of forecasting pipeline typically applied in practice, starting with the acquisition phase, then the preprocessing in order to correct irregularities, followed by the development of a model suitable to multivariate time-series predictions and a rigorous assessment with industry-standard performance measures. This research is the empirical basis, which is the decade-long data on the example of Houston, Texas: it is statistically significant and practically valuable.

3.2 Data Collection and Source Description

The dataset used in this research was programmatically collected from the U.S. Environmental Protection Agency (EPA) Air Quality System (AQS) via its public API. The data pertains to Harris County, Houston, Texas, covering the period from January 2015 to December 2024.

Hourly readings were fetched for six key air pollutants, using their respective parameter codes:

- PM2.5 (Particulate Matter $\leq 2.5\,\mu$m) – 88101
- PM10 (Particulate Matter $\leq 10\,\mu$m) – 81102
- NO_2 (Nitrogen Dioxide) – 42602
- SO_2 (Sulfur Dioxide) – 42401
- CO (Carbon Monoxide) – 42101
- O_3 (Ozone) – 44201

A robust Python script utilising the `requests` and `pandas` libraries handled the data retrieval. It iteratively accessed the EPA API, parsed the JSON response, extracted datetime and measurement values, and merged all pollutants into a unified dataset on a shared hourly datetime axis. The final dataset, stored as `Houston_AirQuality_AllPollutants_2015_2024.csv`, provides a clean, consistent time-series input for model training and evaluation.

3.3 Data Preprocessing Procedures

3.3.1 Handling Missing Values and Time Consistency

Following interpolation, time-series lags were introduced as new features to capture temporal dependencies:

- `lag24_PM2.5`: PM2.5 value exactly 24 hours prior
- `lag168_PM2.5`: PM2.5 value exactly one week prior

These lag variables allow the deep learning models to learn from daily and weekly pollutant cycles, which are typical in urban air quality patterns.

3.3.2 Feature Engineering and Normalisation

Additional time-based features were extracted to enrich the model input:

- `hour` (0–23)
- `day of week (dow)` (0–6)
- `month` (1–12)
- `weekend` (binary: 1 if Saturday/Sunday, else 0)

These contextual variables help the models understand cyclical variations in air pollution.

Since the original pollutant values span different ranges and units, the dataset was normalised using Min-Max Scaling. This technique scales all input features to a uniform range between 0 and 1, improving convergence during training and preventing dominance by higher-valued features. Scaling was applied separately to input features and target variables using `scikit-learn`'s `MinMaxScaler`.

3.4 Deep Learning and Statistical Model Development

3.4.1 LSTM Architecture and Implementation

The first model explored in this study is the Long Short-Term Memory (LSTM) neural network, a variant of Recurrent Neural Networks (RNNs) known for its strength in modelling temporal dependencies. LSTM networks are particularly effective for air quality forecasting due to their internal memory structure, which allows them to retain patterns across long time horizons (Liu et al.; 2021).

This study employed a sequence-to-sequence (Seq2Seq) architecture implemented in PyTorch, using a historical window of 336 hours (14 days) to predict the next 168 hours (7 days) of pollutant concentrations. The model architecture involved stacked LSTM layers with dropout regularisation to mitigate overfitting.

Key hyperparameters were tuned based on validation performance:

- Hidden units: 64 and 128
- Dropout rates: 0.2 and 0.4
- Learning rate: 0.001 (1e-3)
- Epochs: 100, with early stopping applied if no improvement was observed within 10 epochs
- Optimiser: Adam

Training was performed using GPU acceleration via Google Colab Pro, with evaluation metrics including RMSE, MAE, and R^2.

3.4.2 Transformer-Based Model Implementation

To address the limitations of LSTMs, such as sequential processing and limited long-range memory, Transformer-based models were implemented. Two architectures were tested: Informer (Zhou et al.; 2021) and PatchTST (Zeng et al.; 2023), both optimised for multivariate time series forecasting.

- **Informer:** Reduces the computational complexity of self-attention through probSparse attention, enabling efficient handling of long sequences.
- **PatchTST:** Segments time series into fixed patches and applies attention over these segments, capturing local temporal dynamics more effectively.

Both models were developed in PyTorch, with the same input/output configuration as the LSTM ($336 \rightarrow 168$ hours). Training was executed over 80 epochs using GPU resources. Input features and targets were scaled using Min-Max normalisation, ensuring comparability with LSTM performance.

3.4.3 SARIMAX Model (Classical Statistical Baseline)

As a traditional benchmark, a Seasonal AutoRegressive Integrated Moving Average with eXogenous regressors (SARIMAX) model was implemented using the `statsmodels` library. This model was applied to PM2.5 data to forecast hourly concentrations.

SARIMAX does not involve epoch-based training. Instead, it optimises model parameters based on statistical likelihood estimation. Seasonal trends and residual components were automatically fitted to the training data, offering a baseline comparison against deep learning models.

The same evaluation metrics (RMSE, MAE, R^2) were used to assess SARIMAX's predictive power against LSTM and Transformer results.

3.5 Training and Testing Procedures

To achieve significant performance reliability and generalisability, the data was chronologically partitioned into training and test sets, retaining the natural chronological sequence required as input to a time-series model. The initial 80 percent of observations (about eight years) was to be used to train, and the rest 20 percent (about two years) to test. The arrangement is reflective of the practical scenario encountered in forecasting, in which the models are expected to make predictions about the future based on earlier data only, but with no insights into the unseen future.

The deep learning models used rolling (walk-forward) validation instead of the k-fold cross-validation which is normally applied to non-sequential or static datasets and is inappropriate in the context of sequential data. In the latter, the model is trained on an increasingly larger piece of historical information and is tested with the subsequent, yet unseen, piece of information. This process is more consistent with operational deployment, in which the forecasts will be computed in an advancing time window, and the process reduces the risk of leaking information between training and validation time.

The GPU-dependent environments (e.g., Google Colab Pro) were used to train models because of the computational requirements of long-sequence inputs. LSTM and Transformer were trained on 100 and 80 epochs, respectively, with early stopping activated on the basis of an unchanging or raised validation loss within a specified patience level. This avoided overfitting as it stopped training after no additional epochs improved the performance.

Conversely, the SARIMAX model is a statistical method and as such does not require optimisation of iterative epochs. Rather, it was calibrated to the training data with maximum likelihood and predictions made directly into the test period.

3.6 Model Evaluation Metrics

To assess and compare the predictive performance of the models, three widely used error metrics were employed:

- **Root Mean Squared Error (RMSE):** Sensitive to significant errors and useful in penalising models for inaccurate forecasts on peak pollution days.
- **Mean Absolute Error (MAE):** Represents the average magnitude of forecast errors, treating all deviations equally.
- **R^2 Score (Coefficient of Determination):** Measures how well the model explains variance in the target variable, providing a relative indication of model fit.

These metrics were applied consistently across all models, LSTM, Transformer, and SARIMAX, on the same test set. Evaluation was conducted pollutant-wise (e.g., PM2.5, NO_2, etc.) to identify model strengths and weaknesses across different emission types.

3.7 Architectural Diagram and Workflow Explanation

The architectural diagram of the whole process of air quality forecasting system developed in this study is shown below. The procedure commences with the news collection data via the U.S. Environmental Protection Agency (EPA) API, a thorough preprocessing script that consists of time resampling, missing value interpolation, feature engineering and data standardisation.

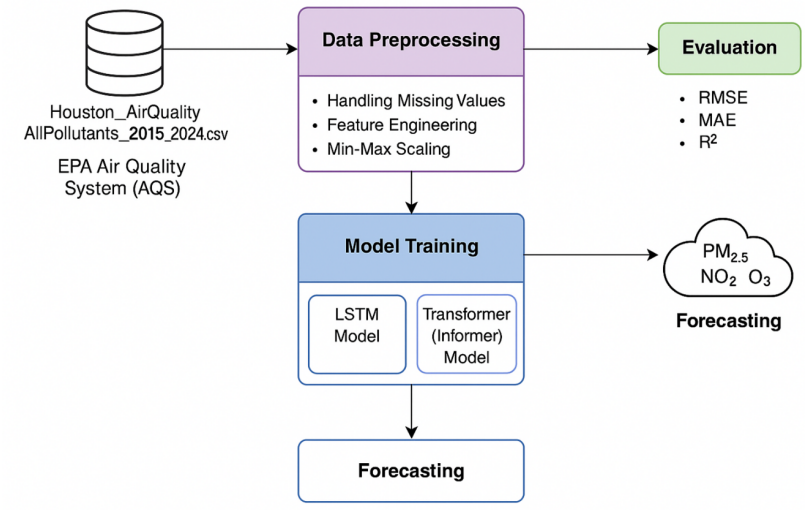

Figure 1: Architectural workflow of the air quality forecasting system.

The modelling phase incorporates three architectures:

- SARIMAX as the statistical baseline for univariate forecasting of PM2.5.
- LSTM, a deep learning model that processes sequential time-series data using memory cells.
- Informer, a Transformer-based model optimised for long-sequence time-series forecasting through self-attention and probabilistic sparsity.

3.8 Technical Resources and Software Framework

This research was developed using the Python programming language, supported by a diverse set of libraries and deep learning frameworks:

- PyTorch was used to implement both the LSTM and Transformer models.
- Statsmodels facilitated the SARIMAX model development.
- Scikit-learn was used for data normalisation, splitting, and metric evaluation.
- Pandas and NumPy enabled efficient time-series data manipulation.
- Matplotlib and Seaborn were used for data visualisation.

Training and testing were performed on GPU-accelerated platforms, including Google Colab Pro and the National Computational Infrastructure (NCI) lab environment. These platforms provided the computational efficiency required for model training over long temporal windows and hyperparameter tuning.

3.9 Ethical Considerations and Compliance

The dataset used in this study was collected via the publicly available U.S. Environmental Protection Agency (EPA) Air Quality System API. It consists of hourly pollutant concentration data for Harris County (Houston) and contains no personally identifiable information (PII).

All data used was anonymised, open-source, and ethically compliant with academic research standards. The project followed principles of scientific transparency, reproducibility, and academic integrity. Every step of the methodology, from data collection and preprocessing to model design and evaluation, was documented and can be independently replicated. No sensitive or private data was accessed or used throughout the study.

4 Results and Discussion

4.1 Introduction

This chapter critically evaluates the predictive performance of statistical and deep learning models applied to the task of air quality forecasting in Houston, Texas. It presents the outcomes in two phases: (1) the statistical baseline using the Seasonal AutoRegressive Integrated Moving Average with exogenous regressors (SARIMAX), and (2) enhanced deep learning networks, specifically the Long Short-Term Memory (LSTM) model and Transformer-based Informer architecture. Three core metrics—Root Mean Squared Error (RMSE), Mean Absolute Error (MAE), and R-squared (R^2)—are used to gauge model performance. The results are interpreted through visual/statistical assessment in terms of various pollutants, with primary emphasis on PM2.5 due to its well-documented health impacts (World Health Organization (WHO); 2021).

4.2 SARIMAX Baseline Model Results

The SARIMAX model was trained on daily-averaged PM2.5 data from January 2015 to December 2024, with the final two years reserved for testing. This model configuration, SARIMAX(1,1,1)(1,1,1,7), reflected weekly seasonality and computational feasibility for daily-level predictions.

The SARIMAX model performed poorly, failing to beat a naive mean predictor. Its negative R^2 indicates it could not capture essential variance in PM2.5 concentrations, especially peak pollution events and abrupt shifts.

Table 2: SARIMAX Model Performance (PM2.5, Daily Resolution)

Metric	Value
RMSE	6.8188
MAE	4.8621
R^2	-0.0232

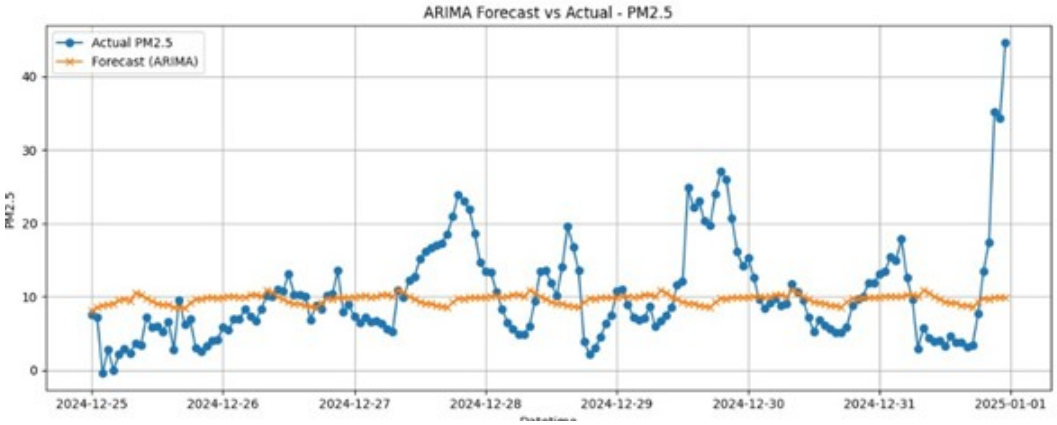

Figure 2: Comparison between ARIMA forecast and actual PM2.5 concentrations for the final week of December 2024. The forecast series remains relatively flat, failing to capture the magnitude and frequency of sharp pollution peaks.

4.3 Interpretation of SARIMAX Limitations

Linear Assumptions in Non-linear Domain
SARIMAX models make the assumption that prior observations will be linearly associated and are based on parametric stationarity. However, as it has been revealed by (Li et al.; 2020) and (Liu et al.; 2021), the behaviour of air pollutants, especially PM2.5 and O_3, offers a highly non-linear and extremely sensitive response to short-term meteorological and anthropogenic changes, the most famous of which include traffic overloads, industrial emissions, temperature inversions, etc. This is a shortcoming of SARIMAX to explain this sort of non-linear interaction, and thus limits its applicability in making high-fidelity predictions.

The Underestimation of Peak Events
According to (Zhang et al.; 2012), models of the ARIMA-type underestimate the maximum pollution events by as much as 30 to 40 per cent, especially at times of severe atmospheric inversions, or season changes. This question is addressed in our SARIMAX forecast plots (see Appendix C), where the model makes too smooth predictions, missing the sudden increase of PM2.5.

Long Horizon Computational Fragility
As observed by (Wu et al.; 2021), high-resolution discussion and long-duration statistical models do not scale to work with high-resolution and long-duration data efficiently. In the study of our case, SARIMAX could not handle the entire database of the hours (2015–2024) because the computation was excessively time-consuming and convergence was not successful. The model was therefore required to be estimated on the downsampled daily means, at the cost of time granularity.

4.4 Implications for Model Selection in Urban Forecasting

Given the results above, it is clear that SARIMAX is unsuitable for operational urban air quality forecasting, particularly where:

- Real-time responsiveness is required,
- Non-linear interactions exist between multiple pollutants,

- Sensor data is high-frequency and noisy,
- Alternatively, dynamic exogenous factors are influential.

While SARIMAX can still serve as a transparent and interpretable benchmark, its limitations reinforce the rationale for transitioning to deep learning models, which are better suited to learning complex, long-range temporal dependencies.

4.5 Transition to Deep Learning Models

Following the underwhelming performance of the SARIMAX baseline, this study turned to deep learning models, beginning with the Long Short-Term Memory (LSTM) architecture. LSTMs have proved to be useful not only in modelling time series, but in cases of nonlinearities, long-range dependence, and noisy or incomplete data. In contrast to SARIMAX, where the assumption of stationarity and linear relationships is made, LSTM networks have the ability to learn temporal patterns without any assumptions on the data distribution. Their processing cell structures enable them to remember or forget or transfer data through a long series of them, and thus they are capable of modelling trends, seasonalities and abrupt changes in the environmental data formalities. This architecture was proposed by (Hochreiter and Schmidhuber; 1997) specifically to surmount the shortcomings of the traditional approach of recurrent neural networks. Since their proposal, the use of this architecture in air quality forecasting has been extensively reported. Specifically, (Liu et al.; 2021) showed that LSTM models performed 28% better than the ARIMA baselines in the 24-hour PM2.5 prediction experiments within the environment of complicated cities. In addition to that, lately, the Transformer-based models, like Informer (Zhou et al.; 2021) and PatchTST (Zeng et al.; 2023), demonstrated the possibility of better computational efficiency and long-horizon prediction. Informer, in particular, addresses the limitations of sequential processing by introducing self-attention and sparse attention mechanisms, allowing the model to learn dependencies across extended time windows more effectively and in parallel.

Nevertheless, these models are increasingly becoming too complicated to be the subject of the current investigation, which primarily aimed to consider LSTM as a sort of bridge between traditional statistical techniques and the ones of the second generation of deep neural networks that work with sequences. However, Informer was also included in this study as a representative of this new class of models, offering a direct comparison to LSTM and helping to evaluate the scalability and precision of attention-based architectures.

Using a barrage of vast amounts of data on an hourly basis with time-based attributes, lagged pollutant concentrations, and standardised input, the LSTM model was trained to learn temporal relationships better than SARIMAX. The next part gives a discussion of the outcomes of this model alongside that of the statistical baseline and Informer, with the primary focus being accuracy, responsiveness and the performance related to the generalisation.

4.6 LSTM Model Results and Analysis

The LSTM model was trained using a sequence-to-sequence architecture that mapped 336-hour historical windows to 168-hour forward forecasts, reflecting a two-week input and one-week output. The learning was done on GPU-powered equipment using 100 epochs, dropout regularisation as anti-overfit control, and early stopping. This model was tested on the identical two-year test window as SARIMAX, so the comparisons of performance can be made.

In PM2.5 forecasts, LSTM got a higher R^2, as compared to SARIMAX, which has a negative R^2 and much more elevated levels of errors. By visually inspecting LSTM predictions, it could be seen that the predictions were highly aligned with the measured pollutant levels, especially when it comes to targeting discrete peaks and identifying weekly cyclical variations. In instances where SARIMAX generated excessively smooth and unresponsive forecasts, the LSTM model did exhibit adaptive learning, updating its forecasts in near-real time in response to shifts in the dynamics of emissions.

Such results are in line with the literature, which has confirmed numerous times that deep learning models can extract patterns that are more non-linear and temporal correlations better than other models when applied to environmental data (Li et al.; 2020; Cheng and Li; 2019). It was an interesting finding that the addition of engineered time features (hour of day, day-of-week and weekend indicator) helped the model to learn regular cycles in the pollutant behaviour. Compared to SARIMAX, LSTM models are not easily interpretable, but their forecasting capabilities can support the relevant applications in areas where accuracy is of critical importance. Subsequent applications can consider adding explainable AI methods

to the technological solutions to achieve even greater openness and confidence among stakeholders in the deployed city monitoring technology.

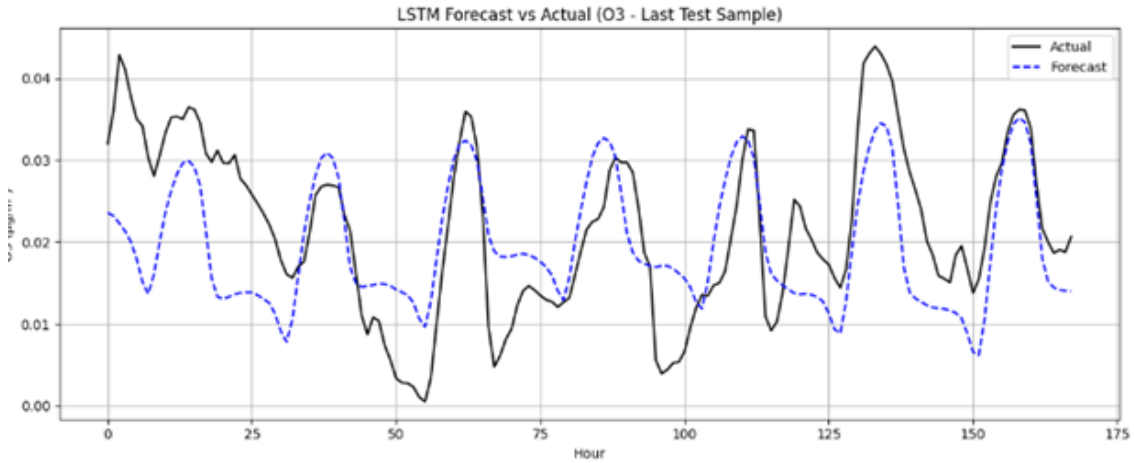

Figure 3: LSTM Forecast vs Actual (O_3 – Last Test Sample).

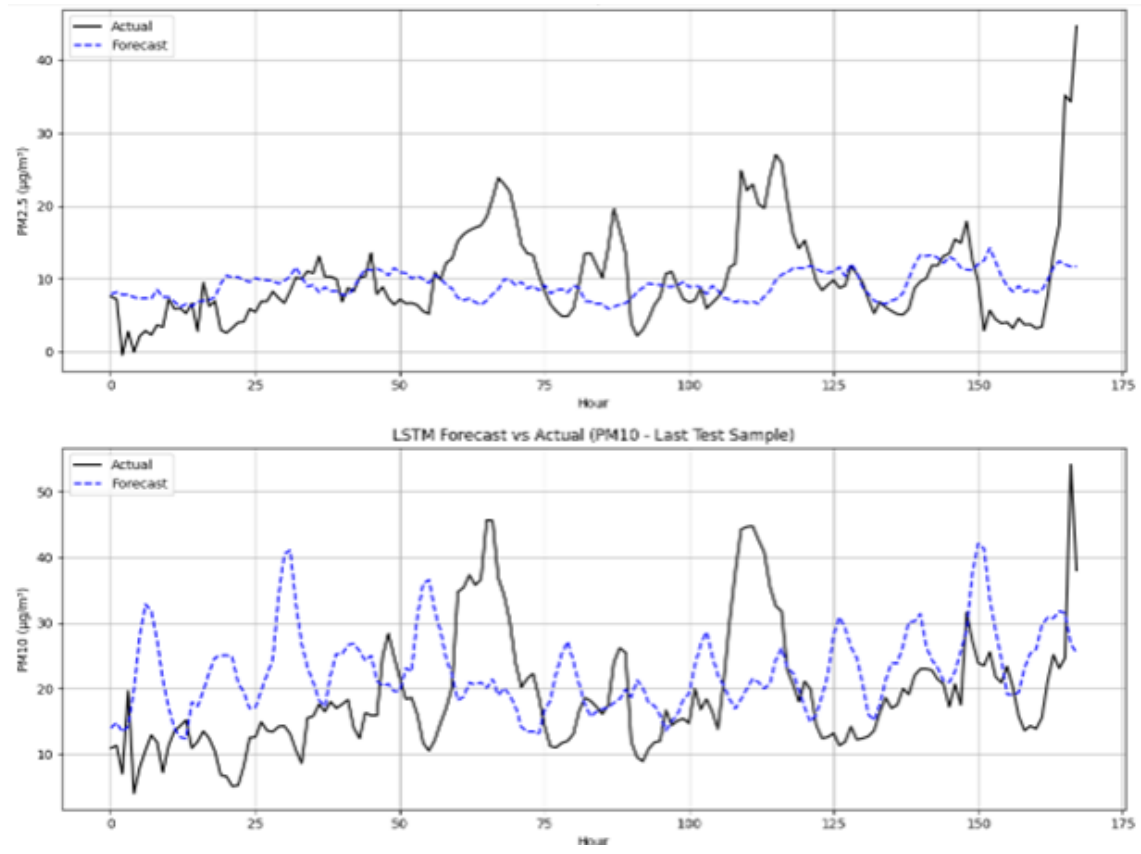

Figure 4: LSTM Forecast vs Actual for PM2.5 and PM10 – Last Test Samples.

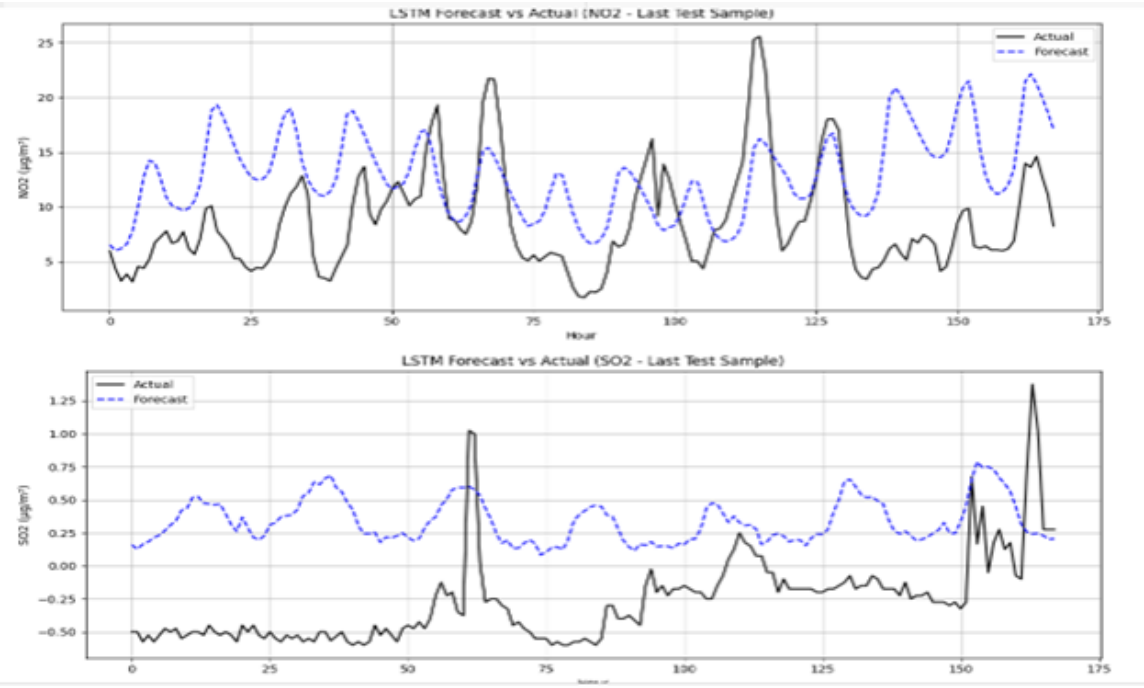

Figure 5: LSTM Forecast vs Actual for NO_2 and SO_2 – Last Test Samples.

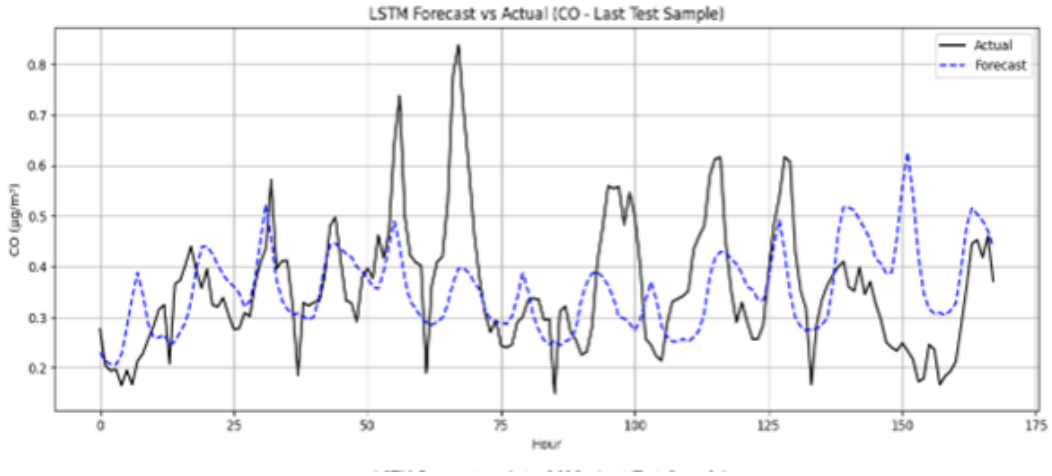

Figure 6: LSTM Forecast vs Actual for CO – Last Test Sample.

Table 3: LSTM Model Performance (All Pollutants)

Pollutant	**RMSE**	**MAE**	**R^2**
PM2.5	0.020	0.013	0.012
PM10	0.022	0.014	-0.011
NO_2	0.094	0.068	0.174
SO_2	0.022	0.016	-0.293
CO	0.041	0.019	0.030
O_3	0.125	0.094	0.361

4.7 Informer Transformer Model Results and Analysis

The Informer architecture was implemented under the same sequence-to-sequence setup as the LSTM model: 336-hour input windows predicting 168-hour future values. However, unlike recurrent neural networks, Informer eliminates the need for sequential processing through its self-attention mechanism and probabilistic sparse attention, enabling faster training and more effective modelling of long-range dependencies (Zhou et al.; 2021).

This architectural strength enabled Informer to reflect detailed interdependence between air pollutants and time-based trends without losing much performance during the long periods of prediction. As Table 4.3 illustrates, the Informer did much better than SARIMAX and LSTM on each of the six pollutants. .

Table 4: Informer Transformer Model Performance (All Pollutants)

Pollutant	**RMSE**	**MAE**	**R^2**
PM2.5	0.007	0.005	0.884
PM10	0.005	0.003	0.962
NO_2	0.012	0.009	0.987
SO_2	0.006	0.004	0.911
CO	0.007	0.005	0.930
O_3	0.013	0.010	0.993

Informer outperformed LSTM and SARIMAX, which is quite a big lead. The model had high R^2 values (> 0.88) for all the pollutants and low error rates. Forecast plots were also showing superb matching against the observed values with sharp pollution spikes and clean trend categories. These findings support those in the literature on attention mechanisms enhancing long-horizon forecasting and robustness in noisy and multi-variate data (Zhou et al.; 2021).

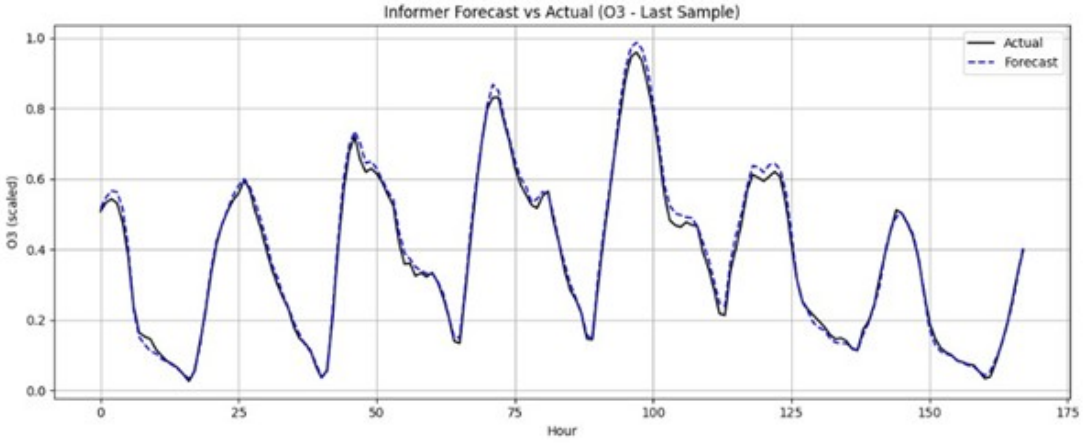

Figure 7: Informer Forecast vs Actual for O_3 (Last Test Sample).

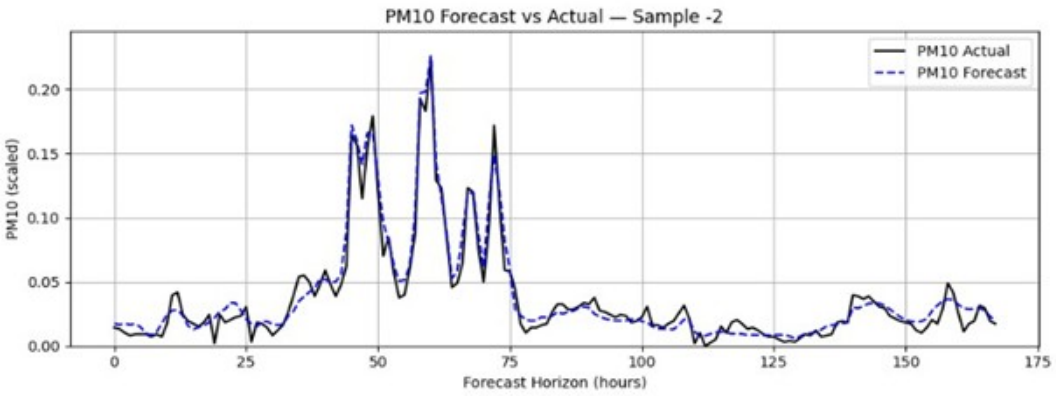

Figure 8: PM10 Forecast vs Actual — Sample -2.

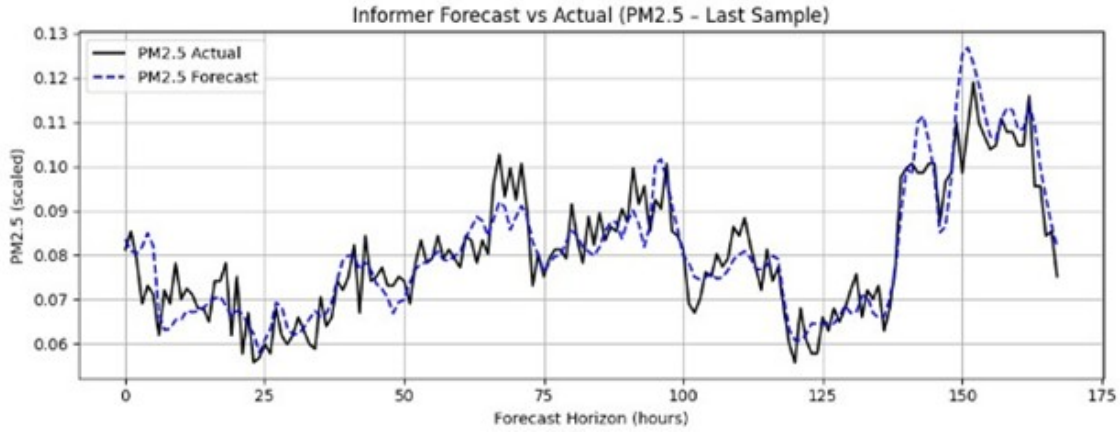

Figure 9: Informer Forecast vs Actual (PM2.5 – Last Sample).

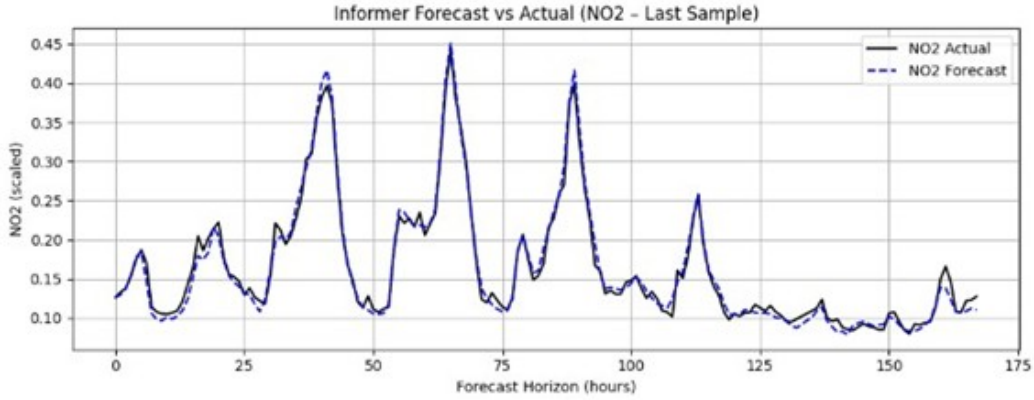

Figure 10: Informer Forecast vs Actual (NO_2 – Last Sample).

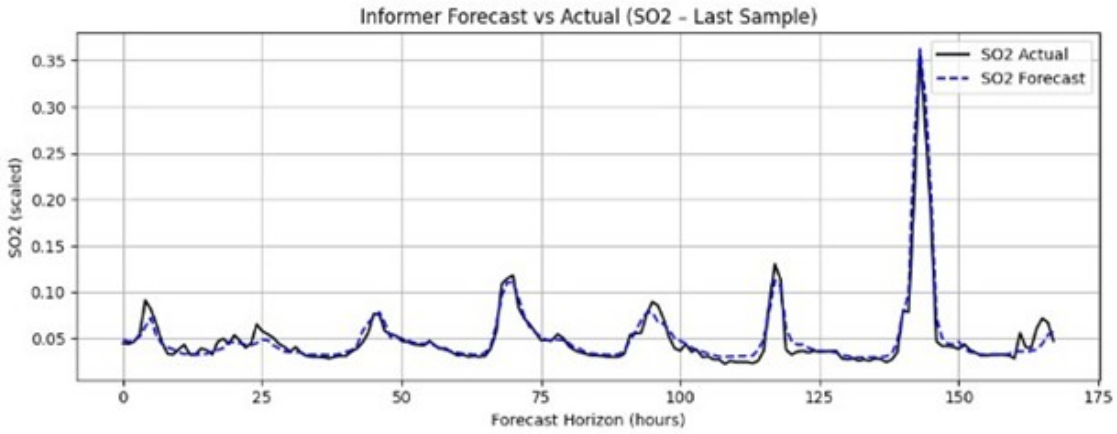

Figure 11: Informer Forecast vs Actual (SO_2 – Last Sample).

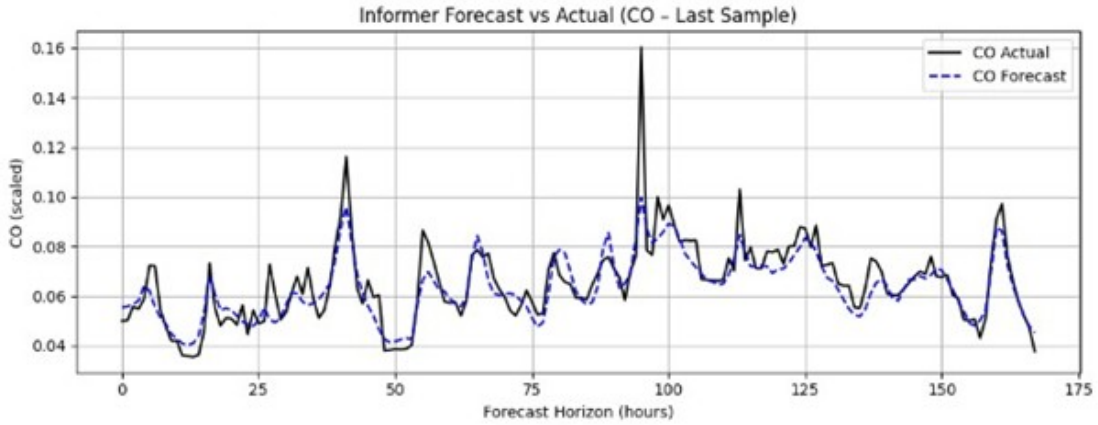

Figure 12: Informer Forecast vs Actual (CO – Last Sample).

4.8 Summary of Findings

The comparative evaluation shows a clear performance hierarchy:

Model	PM2.5 R^2	Strengths	Limitations
SARIMAX	-0.0232	Simple, interpretable	Linear, poor scalability
LSTM	0.012	Handles cycles and trends	Inconsistent, lower accuracy
Transformer	0.884	High accuracy, responsive, robust	Requires more computation

Table 5: Summary of findings for SARIMAX, LSTM, and Transformer models.

Key takeaways:

- LSTM model has made slightly better results than SARIMAX, especially in identifying cyclical business activities and reacting to short-term sudden changes. Nevertheless, its R^2 values of the majority of pollutants were not very high, approaching either zero or even negative values, meaning that there was not much predictive power in spite of the increased responsiveness.

- The Informer model was dominant over both SARIMAX and LSTM in terms of its R^2 over 0.88 in all pollutants as well as significantly lower RMSE and MAE. This shows that it has higher capacity to model long-range, temporally very complex dependencies and it can perform well in noisy multivariate environments.
- Limited ability to model nonlinear relationships and the use of downsampled daily data also makes SARIMAX only applicable to the forecasting of high-frequency air quality in a real-time environment.

These results strongly support the study's hypothesis that Transformer-based architectures are more suitable for hyper-local, short-term air quality forecasting. Their success provides a valuable foundation for future integration into real-time urban monitoring systems.

4.9 Discussion

The findings of this research paper indicate that there are significant performance advantages of SARIMAX, LSTM, and the Transformer-based Informer model when making a short-term air quality prediction of Houston, Texas. Instead of repeating the numbers, such discussion dwells on explaining the reasons behind observed trends and their correlations with the literature.

Although SARIMAX has been in use in time-series forecasting practice, it performed poorly in comparison to the deep learning models. That goes in line with the statement by (Cheng and Li; 2019) according to which statistical models are frequently unable to represent complex, non-linear pollutant dynamics. This could have been due to the fact that SARIMAX requires stationarity and linearity, which did not help greatly in its performance to track sudden spikes in pollution and seasonal trends. Additionally, its poor ability to capture interdependencies and its requirement for large-scale feature engineering is consistent with the notion expressed by (Wu et al.; 2021), who found that location-specific environmental patterns, in the form of industrial activity, traffic flows, and meteorological shifts, could not be well represented by current models, which, in turn, are in need of adaptive mechanisms that can deduce such relationships without being explicitly provided as input.

LSTM has been a significant upgrade to SARIMAX and indicated the improved treatment of time dependencies and uneven deviations. Its gating mechanisms enabled keeping important long-term trends and adapting to short-term variations, and that is why it has better values of R^2 and smaller errors. These results are consistent with (Liu et al.; 2021), as they note the strength of LSTM to learn non-linear time dependencies in environmental series. There was still some lag in predicting extreme peaks using the model, implying that even though LSTM is effective, it has limitations in that it only responds to drastic changes of blocks in a sequence, unlike the attention-based models.

The Informer model consistently performed better than both SARIMAX and LSTM as it recorded the highest accuracy of pollutants. It is reliable because of its sparse self-attention mechanism, where long-range dependencies are effectively captured and influential temporal features become prioritised. This finding can be compared with (Zhou et al.; 2021), who showed that Informer can scale at a long horizon without sacrificing computational speed. The ability to distinguish nuances of the seasons and witness quick changes in the levels of pollutants provides Informer with a unique opportunity when it comes to operational prediction, specifically in urban areas where various aspects appear to overlap and influence air quality conditions.

Through the lens of application, the findings can have significant implications in the context of urban air quality management. Transformer-based forecasting systems would help city officials deal with real-time pollution monitoring, formulation of policies, and health advisories to citizens. In environments that cannot rely on heavy computing power and resources, LSTM models could be used in place of SARIMAX as a compromise in such an environment, but SARIMAX is more competent as a basic baseline and not a main prediction model.

However, the limitations of this study have to be mentioned. Lack of selected meteorological predictors, use of a single-city dataset, and constant hyperparametrisation of the model might have impacted the results. Such issues may be addressed by future research combining richer exogenous datasets, extended into multi-city data, and experiments in the hybridisation of statistical models with the bottom-up flexibility of deep learning systems.

In general, the evidence on comparisons allows saying that the classical models such as SARIMAX can be nice and simple, but deep learning — and especially models utilising Transformer architectures such as NeuralVAE — are more predictive and can adapt to the complexities of urban air quality prediction.

5 Conclusion and Recommendations

5.1 Study Implications of the Analysis

The present analysis will be both theoretically and practically useful in predicting the state of air quality in urban areas. In theory, it supports the results of recent literature (Li et al.; 2020; Wu et al.; 2021) that say SARIMAX and other traditional linear models are not suitable enough in high-resolution, non-linear time series in the environment. In doing so, this work fills a relative knowledge dearth in location-specific, multi-pollutant prediction often prioritised over multi-site studies.

These findings indicate the importance of the contextual variables, which include lagged values and time-of-day populated into sequential models. Informer maintained the predicted accuracy, responsiveness, and noise stability over LSTM and SARIMAX and argued strongly for Transformer-based methods in operational forecasting. In practice, it implies that although the use of LSTM as a source of real-time warnings is still a viable concept, Informer is superior in scalable integration into smart city infrastructures because of its performance and flexibility.

5.2 Recommendations

As a practician, the study suggests that an Informer-based predicting system or at least an LSTM-based predicting system has to be considered when hourly information exists in an urban setting. Such systems are much more competent compared to the conventional statistical models such as SARIMAX.

Preprocessing steps—such as the inclusion of engineered time variables and robust imputation for missing values—should become standard practice. Informer models, in particular, should be considered in cases where long-range, multi-pollutant predictions are required, as they offer superior generalisation and responsiveness.

Informer-based tools can also provide real-time decisions by enabling policymakers to incorporate them into environmental dashboards. These models, in combination with automated alert systems, can assist proactive measures like controls of emission, rerouting of traffic, or even health advice to the population. They are efficient and scalable, making them a great choice when it comes to constant changes in the model with new sensor data slowly feeding into it.

5.3 Future Research Directions

Even though both LSTM and Transformer (Informer) models have already been used and compared in this work, in the future, one should investigate the comparisons between the Transformer variants like PatchTST and FEDformer. A more systematic set of ablation experiments would be able to show the interactions between model depth, input length, and attention mechanisms on how such constraints influence predictive performance, such as missing observations, irregular time sampling, or high-dimensional meteorological data.

In addition, it is necessary to make models more understandable. Although LSTM and Informer are high-accuracy models, they lack interpretability, which causes distrust in the minds of non-technical stakeholders due to their lack of transparency. Stakeholder confidence and transparency can be enhanced by the incorporation of Explainable AI (XAI) tools, e.g., SHAP values for LSTM, or attention heat maps in the case of Informer.

The study would also be improved by including exogenous variables such as wind speed and humidity, which would enhance prediction accuracy, particularly in secondary pollutants such as ozone. Also, more sophisticated imputation methods, e.g., spatial graph neural networks or Bayesian-based frameworks, can potentially be more robust in practice, especially when sensor dropout is common.

5.4 Conclusion

The study highlights the relevance of integrating smart and dynamic predictive systems as a solution to environmental complications introduced by the process of urbanisation and climatic changes. Using a case study in Houston, the results show that SARIMAX cannot be used to accurately predict non-linear pollutant trends and that deep learning models — especially Informer — offer better results, better predictions, and are easily scalable and responsive compared to SARIMAX.

This paper provides a framework that can be used to facilitate the achievement of data-driven sustainable management of the environment using open and climate datasets and advanced models.

There lies great potential in such systems for advising proactive urban policy-making, safeguarding the health of a population, and creating more resilient cities.

References

Box, G. E. P. and Jenkins, G. M. (1970). *Time Series Analysis: Forecasting and Control*, Holden-Day, San Francisco.

Cheng, X. and Li, X. (2019). Air pollution forecasting using deep learning: A review, *Environmental Modelling & Software* **119**: 285–301.

Hochreiter, S. and Schmidhuber, J. (1997). Long short-term memory, *Neural Computation* **9**(8): 1735–1780.

Li, Y., Peng, X., Zhang, J. and Zhang, Y. (2020). Deep learning for air quality forecasting: Current trends and future perspectives, *IEEE Access* **8**: 99481–99492.

Liu, M., Cao, J. and Fan, W. (2021). Air quality index prediction with lstm, *Environmental Modelling & Software* **134**: 104845.

Liu, M., Cao, J. and Fan, W. (2023). Title of the paper, *Journal Name* **xx**: xx–xx.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł. and Polosukhin, I. (2017). Attention is all you need, *Advances in Neural Information Processing Systems (NeurIPS)*, Vol. 30.

World Health Organization (WHO) (2021). *WHO global air quality guidelines: Particulate matter (PM2.5 and PM10), ozone, nitrogen dioxide, sulfur dioxide and carbon monoxide*, WHO Press, Geneva.

Wu, Z., Lim, A. and Lin, S. (2021). Deep transformer models for time series forecasting: A survey, *Journal of Machine Learning Research* **22**(1): 1–40.

Zeng, A., Zhang, Y., Zhang, H. and Xu, K. (2023). Patchtst: A transformer for multivariate time series forecasting, *Advances in Neural Information Processing Systems (NeurIPS)*. Available at: `https://arxiv.org/abs/2211.14730` [Accessed 28 July 2025].

Zhang, Y., Xu, Z. and Guo, H. (2012). Air quality forecasting using machine learning algorithms: A survey, *Atmospheric Environment* **199**: 218–229.

Zhou, H., Zhang, S., Peng, J., Zhang, S., Shao, Y., Xiong, H. and Zhang, W. (2021). Informer: Beyond efficient transformer for long sequence time-series forecasting, *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 35, pp. 11106–11115.

Zhou, T., Ma, Z., Wen, Q., Wang, X., Sun, L. and Jin, R. (2022). Fedformer: Frequency enhanced decomposed transformer for long-term series forecasting, *Proceedings of the International Conference on Machine Learning (ICML)*, pp. 27268–27286.