Engineering Systems and Intelligent Technologies
ESIT
ISSN: 3071-253X/
Journal Homepage
https://pub.scientificirg.com/index.php/ESIT

A Comparative Evaluation of LSTM and Bidirectional LSTM Architectures for Nitrogen Oxide Forecasting in Air Quality Monitoring

Mohamed Hassan Essai Ali [a,1] **, Mohamed M. Ali**[b] **, and Gamal M. A Mahran**[c]

[a] *Department of Electrical Engineering, Faculty of Engineering, Al-Azhar University, Qena, Egypt; E-mail: mhessai@azhar.edu.eg*

[b] *Department of Mining and Petroleum Engineering, Faculty of Engineering, Al-Azhar University, Qena, Egypt; E-mail: mmam64@yahoo.com*

[c] *Mining Engineering Department, King Abdulaziz University, Jeddah 21589, Saudi Arabia; E-mail: gmahran@kau.edu.sa*

ABSTRACT

Reliable, real-time monitoring of nitrogen oxides (NO_x) is essential for air quality management, yet conventional monitoring networks are limited by sparse spatial coverage and inconsistent predictive reliability. This study compares Long Short-Term Memory (LSTM) and Bidirectional LSTM (BiLSTM) networks for forecasting NO_x concentrations using the UCI Air Quality Dataset. To assess these architectures rigorously, we benchmark them against a persistence forecast and an ARIMA model, apply expanding-window temporal cross-validation to guard against data leakage, and test the significance of any performance gap with the Diebold–Mariano statistic. Both recurrent architectures attain high accuracy ($R^2 > 0.999$), with BiLSTM showing a marginally higher mean R^2 (0.99977 versus 0.99972 for LSTM); however, the Diebold–Mariano test indicates that this difference is not statistically significant ($p = 0.384$), so the two architectures should be regarded as practically equivalent for this task. Both deep learning models substantially outperform the statistical baselines, most notably ARIMA ($R^2 = 0.854$), which supports the value of recurrent architectures for NO_x forecasting while also underscoring the importance of baseline comparison and leakage-aware validation in this line of work.

PAPER INFORMATION

HISTORY

Received: 13 March 2026
Revised: 20 May 2026
Accepted: 19 June 2026
Online: 29 June 2026

MSC

68T07; 68T09; 62M10; 62P12; 68T05

KEYWORDS

LSTM;
BiLSTM;
Nitrogen Oxide Forecasting;
Air Quality Monitoring.

1. INTRODUCTION

Rapid urbanization and continued expansion of industrial activity have placed significant pressure on atmospheric quality worldwide. Among the gaseous pollutants of greatest concern, nitrogen oxides (NO_x), which comprise nitric oxide (NO) and nitrogen dioxide (NO_2), are predominantly generated by high-temperature combustion processes in transportation, power generation, and heavy industry. Chronic exposure to elevated NO_x concentrations is associated with aggravated respiratory conditions, increased cardiovascular risk, and broader ecosystem harm through acid deposition and photochemical smog formation. Robust, timely monitoring of NO_x is therefore indispensable for evidence-based environmental policymaking and public health protection.

[1]Corresponding author at Department of Electrical Engineering, Faculty of Engineering, Al-Azhar University, Qena, Egypt; E-mail: mhessai@azhar.edu.eg

Conventional monitoring infrastructure, including fixed reference stations and dispersion models, suffers from several well-documented shortcomings. Fixed stations provide accurate measurements only at discrete geographic points, leaving large urban areas without representative coverage. Dispersion modeling requires detailed, high-quality emission inventory data that are often unavailable or outdated, and model fidelity decreases markedly under non-stationary meteorological conditions [1, 2]. These limitations have motivated growing interest in data-driven forecasting approaches capable of learning complex pollutant dynamics directly from observational records.

Deep learning has profoundly advanced the state of the art in temporal sequence modeling. Recurrent Neural Networks (RNNs) maintain an internal state that encodes temporal context, making them well suited to multi-step time series prediction. Long Short-Term Memory (LSTM) networks [3] address the gradient vanishing problem inherent to vanilla RNNs by incorporating gated memory cells that selectively retain or discard information over extended time horizons. Bidirectional LSTMs (BiLSTMs) generalize the LSTM architecture by processing each input sequence in both forward and reverse temporal directions, allowing the model to simultaneously exploit past observations and anticipated future context when encoding the current state. This bidirectional representation is particularly informative for pollutant series, where delayed responses to meteorological forcing create meaningful backward temporal dependencies [4, 5].

Prior studies have applied LSTMs to air quality forecasting with encouraging results; however, rigorous side-by-side comparisons that include statistical baseline benchmarks and guard against temporal data leakage remain relatively scarce in the published literature. The present work fills this gap through four specific methodological contributions:

(1) A controlled head-to-head comparison of LSTM and BiLSTM networks trained under identical hyperparameter settings.

(2) Inclusion of Persistence Forecast and ARIMA baselines to quantify the incremental benefit of deep learning.

(3) Expanding-window temporal cross-validation to confirm generalization to genuinely unseen future periods.

(4) Diebold-Mariano hypothesis testing to determine whether observed performance differences are statistically meaningful.

The remainder of this article is organized as follows. Section 2 reviews directly related prior work. Section 3 reviews air pollution challenges and sustainability considerations. Section 4 summarizes conventional monitoring methods. Section 5 surveys deep learning approaches to pollution forecasting. Section 6 describes the proposed methodology. Section 7 presents experimental results. Section 8 concludes with practical implications and directions for future research.

2. RELATED WORK

The application of recurrent deep learning to air quality prediction has grown substantially over the past decade. Zhang et al. [5] provided a comprehensive review of deep learning methods for air pollutant concentration prediction, cataloguing model families and benchmark datasets and identifying recurrent architectures as the dominant paradigm for univariate and multivariate time series tasks. Their analysis confirmed that LSTMs consistently outperform classical statistical models, particularly for pollutants with strong autocorrelative structure.

The bidirectional extension of the LSTM was explored for air quality prediction by Zhang et al. [6], who proposed a semi-supervised BiLSTM framework for $PM_{2.5}$ prediction in Beijing. Their model incorporated unlabeled data through an Empirical Mode Decomposition preprocessing stage, and achieved an R^2 of 0.989 at the hourly scale, outperforming standard LSTM and GRU baselines. This study provided an important early indication that BiLSTMs can exploit backward temporal context to improve pollutant forecasting.

Naresh and Indira [7] demonstrated the utility of multivariate LSTM networks for predicting multiple pollutant species simultaneously, reporting competitive performance on benchmark datasets while highlighting the importance of feature selection in reducing model complexity.

In the domain of NO_x specifically, sensor-based prediction studies have examined hybrid architectures that fuse spatial and temporal information. Nagrecha et al. [8] combined CNN layers for local feature extraction with LSTM layers for temporal modeling, and their CNN-LSTM model achieved meaningful improvements over single-branch architectures when applied to multi-sensor urban networks.

Kabir et al. [9] proposed an integrated framework combining belief rule bases with deep learning for multi-pollutant prediction, demonstrating that ensemble approaches that blend domain knowledge with learned representations can improve robustness when training data are limited.

Wu et al. [10] addressed the data quality problem in air quality modeling by training a GAN-based imputation model for indoor measurements, underscoring the importance of preprocessing completeness for downstream forecasting accuracy.

Andrade et al. [4] conducted an evaluation of multiple AI techniques for Air Quality Index (AQI) prediction spanning RNNs, LSTMs, BiLSTMs, and Transformer-based models. Their results showed that while Transformers achieved the

highest accuracy on long-horizon prediction tasks, LSTM and BiLSTM variants remained competitive for one-step-ahead forecasts at reduced computational cost.

Despite the richness of this prior work, several research gaps remain apparent. First, most studies evaluate models on a single train-test split, introducing potential data leakage when the split boundary is not strictly enforced in temporal order. Second, comparisons between LSTM and BiLSTM are frequently confounded by simultaneous changes to other hyperparameters, preventing clean attribution of performance differences to architecture alone. Third, statistical significance testing of forecast accuracy differences is rarely applied, leaving open the question of whether reported improvements reflect genuine model superiority or random variation. The present study is designed specifically to address all three of these limitations.

3. AIR POLLUTION CHALLENGES AND SUSTAINABILITY

3.1 Emission Sources

Anthropogenic NO_x emissions arise from several interlocking sectors. Road transport remains the dominant urban source, with internal combustion engines generating NO_x through high-temperature nitrogen oxidation. Industrial facilities, thermal power plants, and residential heating systems contribute additional steady-state loads. Agricultural practices, including fertilizer application and livestock operations, produce nitrogenous precursors that undergo atmospheric conversion to NO_x species [11, 12]. Natural events such as wildfires and volcanic eruptions introduce episodic emission bursts that can overwhelm local monitoring infrastructure [13]. The full pollutant spectrum also encompasses carbon monoxide (CO), sulfur dioxide (SO_2), ozone (O_3), fine particulate matter ($PM_{2.5}$, PM_{10}), and volatile organic compounds (VOCs), each interacting chemically within the atmosphere to generate secondary pollutants.

3.2 Health and Environmental Impacts

Epidemiological evidence consistently links chronic NO_x exposure to elevated incidence of respiratory and cardiovascular morbidity. Long-term exposure studies report accelerated decline in lung function, increased hospitalizations for asthma and chronic obstructive pulmonary disease, and elevated premature mortality [14, 15]. Beyond direct health effects, NO_x plays a central role in tropospheric ozone formation through photochemical reactions with VOCs, contributing to regional oxidant pollution. Ecological consequences include nitrogen deposition that alters nutrient cycling in sensitive terrestrial and aquatic ecosystems. Economic assessments of the aggregate burden of air pollution, accounting for healthcare costs, lost productivity, and environmental remediation, indicate substantial societal costs that justify investment in monitoring and abatement technology [16].

3.3 Mitigation Strategies and the Role of Forecasting

Policy responses to air pollution span regulatory, technological, and behavioral domains. Transition to renewable energy sources reduces combustion-related emissions at their point of origin [17]. Electrification of transport fleets and adoption of low-emission vehicle standards progressively decarbonize road transport [18]. Precision agriculture minimizes nitrogen surplus in farming operations, reducing atmospheric precursor loads [19]. Smart urban design, including green corridors and traffic management optimization, can substantially lower local pollutant concentrations. Across all of these mitigation pathways, real-time predictive monitoring plays an enabling role by providing actionable information to regulators, operators, and the public, supporting both immediate interventions and longer-term planning [20].

4. CONVENTIONAL AIR POLLUTION DETECTION METHODS

Table 1 presents an overview of established monitoring methods, their operating principles, principal strengths, and inherent limitations. Fixed reference stations provide continuous, high-precision measurements but cannot practically achieve the spatial density required to characterize pollution gradients across large metropolitan areas [20, 21]. Passive diffusion samplers offer a low-cost complement, capturing time-integrated concentrations over days or weeks, but cannot support real-time alerting [22]. Wet-chemistry analytical methods yield accurate quantitative results but are laboratory-bound, slow, and resource-intensive [23]. Atmospheric dispersion models, when coupled with high-quality emission inventories and meteorological fields, can predict pollutant transport and transformation at high spatial resolution, but their accuracy is inherently limited by uncertainty in the underlying input data [24, 25]. Biological monitoring, which uses sentinel organisms to integrate cumulative pollution exposure, provides cost-effective long-term trend information at the expense

of quantitative precision and temporal specificity [26]. Manual sampling campaigns cover the widest range of chemical species but provide only intermittent, labor-intensive snapshots unsuitable for continuous air quality management [27].

Table 1: Comparison of conventional air pollution monitoring methods

Method	Description	Advantages	Limitations
Fixed monitoring stations [20, 21]	Continuous measurement at instrumented sites	Long-term trend tracking	Limited spatial coverage
Passive samplers [22]	Pollutant collection by diffusion or adsorption	Inexpensive, easy to deploy	No real-time output
Chemical (wet) methods [23]	Laboratory wet-chemistry analysis	High accuracy	Labor intensive
Dispersion modeling [24, 25]	Simulation of pollutant transport	Scenario and policy assessment	Sensitive to input data quality
Biological monitoring [26]	Use of bioindicator organisms	Low-cost, long-term signal	Complex interpretation
Manual sampling [27]	Laboratory analysis of collected samples	Broad pollutant coverage	Not continuous

5. DEEP LEARNING-BASED AIR POLLUTION DETECTION

A range of deep learning architectures has been applied to pollution forecasting, each suited to a different aspect of the problem. Recurrent architectures, including LSTM networks, are designed to retain information over time and have been used successfully to model the seasonal and diurnal structure present in pollutant time series [28, 3]. Convolutional neural networks (CNNs) are instead built to extract spatial features, which makes them well suited to satellite imagery and gridded air quality maps [8]. Deep belief networks (DBNs) have been used to model the joint influence of meteorology, traffic, and industrial activity on pollutant concentrations [9], while hybrid CNN-LSTM architectures combine spatial feature extraction with temporal modeling in a single pipeline [8]. Generative adversarial networks (GANs) have also been applied in this domain, primarily to generate synthetic training scenarios and to impute missing sensor readings [10].

Recurrent architectures: RNNs and their LSTM variant have been widely applied to single-step and multi-step pollutant forecasting. LSTM networks overcome the gradient vanishing limitation of vanilla RNNs by employing three learnable gates, namely input, forget, and output gates, whose interaction controls information flow through a dedicated cell state. This mechanism enables effective modeling of seasonal cycles, diurnal patterns, and multi-day persistence effects [7].

Bidirectional extension: Within this landscape, BiLSTM networks have been used for pollutant concentration forecasting, generally as part of hybrid pipelines that pair bidirectional processing with decomposition or attention mechanisms [6, 4], and for missing-data imputation in air quality records [10]. What is comparatively rare in this literature is a controlled comparison in which LSTM and BiLSTM are trained under identical conditions, benchmarked against simple statistical baselines, and compared with an explicit significance test rather than a raw difference in point metrics; this is the gap the present study addresses.

Convolutional architectures: Convolutional Neural Networks (CNNs) excel at extracting spatial features from gridded data, making them effective for satellite remote sensing inputs and spatially distributed sensor networks. CNN-LSTM hybrid models combine spatial feature extraction with temporal sequence learning, achieving state-of-the-art performance on spatiotemporal pollution forecasting tasks [8].

Deep Belief Networks: Deep Belief Networks (DBNs) learn hierarchical probabilistic representations of complex, multi-source pollution driving factors, including meteorological variables, traffic loads, and emission inventories [9].

Generative models: Generative Adversarial Networks (GANs) have been applied to synthetic data augmentation for sparse sensor networks and missing data imputation, tasks where conventional interpolation methods underperform [10].

Deep Reinforcement Learning: Deep Reinforcement Learning (DRL) provides a framework for optimizing emission control decisions under uncertainty. Recent work demonstrates that DRL agents trained on simulated urban environments can identify pollution mitigation strategies superior to those derived from static optimization [29].

More broadly, deep learning offers automatic feature extraction, the ability to ingest high-dimensional and heterogeneous inputs, and adaptability as conditions change. Despite the breadth of available architectures, controlled comparisons that isolate the contribution of bidirectionality while holding all other design choices constant are uncommon. The present study addresses this gap directly.

6. PROPOSED MODELS AND METHODOLOGY

6.1 Dataset Description

This study uses the UCI Air Quality Dataset, which contains hourly measurements recorded by a multisensor device deployed at a polluted road-level site in an Italian city between March 2004 and February 2005 [30]. The dataset includes 9,357 hourly observations covering CO, non-methane hydrocarbons (NMHC), benzene (C_6H_6), total NO_x, NO_2, temperature, and relative humidity, with ground-truth concentrations obtained from a co-located reference analyzer. Missing values, flagged in the original release and accounting for approximately 2.1% of the records used here, were imputed by linear interpolation prior to model training.

6.2 Data Preprocessing and Leakage Prevention

All time series were standardized to zero mean and unit variance prior to training, using statistics computed from the training portion of each fold only, to avoid leaking test-set statistics into the normalization step. To guard against temporal data leakage, in which information from the future contaminates training, we adopted an expanding-window cross-validation scheme (detailed in Section 7.2) in which the training window always precedes the corresponding test window in time. We also verified that the NO_x(GT) target variable was excluded from the input feature vector at the prediction time step, since including the current-step target as a predictor would trivially inflate accuracy regardless of architecture; this check is reported here because it is a common and easily overlooked source of leakage in hourly pollutant forecasting.

6.3 Baseline Models

Two baselines were used to quantify the benefit, if any, of the recurrent architectures over simpler alternatives.

Persistence forecast:

The persistence baseline predicts that the next value equals the current value,

$$\hat{y}_{t+1} = y_t, \tag{1}$$

which is a naive but, for strongly autocorrelated series such as hourly pollutant concentrations, a genuinely competitive baseline.

ARIMA:

The autoregressive integrated moving-average model, ARIMA(p, d, q), is given in backshift-operator form by

$$\left(1 - \sum_{i=1}^{p} \phi_i B^i\right)(1 - B)^d y_t = \left(1 + \sum_{j=1}^{q} \theta_j B^j\right)\varepsilon_t, \tag{2}$$

where B is the backshift operator ($By_t = y_{t-1}$), ϕ_i and θ_j are the autoregressive and moving-average coefficients, and ε_t is white noise [31]. The orders p, d, and q were selected independently for each training fold by minimizing the Akaike Information Criterion (AIC) over a bounded grid search.

6.4 Mathematical Formulation of the LSTM Detector

The LSTM layer maintains a cell state c_t and a hidden state h_t that are updated at each time step through three gates that control what information is forgotten, written, and exposed. For input vector x_t and previous hidden state h_{t-1},

$$f_t = \sigma\big(W_f\,[h_{t-1}, x_t] + b_f\big), \tag{3}$$
$$i_t = \sigma(W_i\,[h_{t-1}, x_t] + b_i), \tag{4}$$
$$o_t = \sigma(W_o\,[h_{t-1}, x_t] + b_o), \tag{5}$$
$$\tilde{c}_t = \tanh(W_c\,[h_{t-1}, x_t] + b_c), \tag{6}$$
$$c_t = f_t \odot c_{t-1} + i_t \odot \tilde{c}_t, \tag{7}$$
$$h_t = o_t \odot \tanh(c_t), \tag{8}$$

where f_t, i_t, and o_t are the forget, input, and output gate activations, $\tilde{c}_t$ is the candidate cell update, $\sigma(\cdot)$ is the logistic sigmoid, $\odot$ denotes elementwise multiplication, $[\,\cdot,\cdot\,]$ denotes vector concatenation, and $W_{(\cdot)}$, $b_{(\cdot)}$ are learned weight matrices and bias vectors [3]. The detector takes the hidden state at the final time step of each 24-hour input window, h_T, and passes it through a fully connected layer with a single linear output unit to produce the predicted NO_x concentration $\hat{y}_{t+1}$.

6.5 Mathematical Formulation of the BiLSTM Detector

The BiLSTM detector applies the recursion in **Equations** (3)–(8) twice over the same input window, once in the forward direction to obtain $\overrightarrow{h}_t$ and once over the time-reversed sequence to obtain $\overleftarrow{h}_t$, using independent weight matrices for each direction. The two hidden states are concatenated at each time step,

$$h_t = \left[\overrightarrow{h}_t\,;\,\overleftarrow{h}_t\right], \tag{9}$$

so that the resulting representation reflects information from both past and future context within the window [32]. As with the unidirectional model, the final-step representation is passed through a fully connected layer with a single linear output unit.

6.6 Network Architecture and Hyperparameters

Both detectors share the same overall pipeline, differing only in the recurrent layer:

$$\text{SequenceInput} \rightarrow \text{LSTM or BiLSTM}(200) \rightarrow \text{Dropout}(0.3) \rightarrow \text{FullyConnected}(1) \rightarrow \text{Regression output.} \tag{10}$$

The recurrent layer uses 200 hidden units (200 in each direction for BiLSTM, so that the concatenated representation has 400 units before the fully connected layer), takes the hidden state at the last time step of a 24-hour input window ('OutputMode = last'), and is followed by dropout with rate 0.3 for regularization.

6.7 Training Configuration

Both models were trained under identical conditions: Adam optimizer with initial learning rate 0.005, a maximum of 150 epochs, mini-batch size 64, gradient-norm clipping threshold 1, and mean squared error as the training loss. Training was stopped early if the validation RMSE failed to improve for 10 consecutive epochs (patience = 10), and the model weights from the best validation epoch were retained; this early-stopping criterion is also the regularization mechanism referenced in the overfitting discussion in Section 7.6. Algorithm 1 summarizes the full procedure.

Algorithm 1 NO_x concentration prediction using LSTM/BiLSTM

Require: Multivariate hourly time series with d input features (pollutants and meteorological variables)
Ensure: Predicted NO_x concentration for the next time step

1. **Preprocessing:** standardize features using training-fold statistics; build 24-hour input windows; apply the expanding-window split
2. **Model:** SequenceInputLayer(d) → LSTM or BiLSTM(200, OutputMode = last) → Dropout(0.3) → FullyConnected(1) → RegressionLayer
3. **Training:** Adam optimizer (learning rate 0.005), up to 150 epochs, batch size 64, gradient threshold 1, early stopping on validation RMSE (patience 10)
4. **Testing:** evaluate on the held-out temporal window of the current fold
5. **Validation:** compute MSE, RMSE, and R^2; compare architectures using the Diebold–Mariano test

6.8 Evaluation Metrics

For n test points with true values y_i and predictions $\hat{y}_i$, performance was measured using

$$\text{MSE} = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2, \tag{11}$$

$$\text{RMSE} = \sqrt{\text{MSE}}, \tag{12}$$

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}, \tag{13}$$

where $\bar{y}$ is the mean of the true values in the test fold.

To test whether the LSTM and BiLSTM forecasts differ significantly in accuracy, we used the Diebold–Mariano test [33]. Let $e_{1,t}$ and $e_{2,t}$ denote the forecast errors of the two models at time t, and define the squared-error loss differential $d_t = e_{1,t}^2 - e_{2,t}^2$. The test statistic is

$$\mathrm{DM} = \frac{\bar{d}}{\sqrt{\widehat{\mathrm{Var}}(\bar{d})}}, \tag{14}$$

where $\bar{d}$ is the sample mean of d_t over the test fold and $\widehat{\mathrm{Var}}(\bar{d})$ is a Newey-West long-run variance estimate that accounts for serial correlation in d_t. Under the null hypothesis of equal predictive accuracy, DM is approximately standard normal.

6.9 Real-World Scalability Considerations

Deploying BiLSTM models at the scale of a city-wide sensor network raises three practical issues. First, bidirectional processing roughly doubles the computation of an equivalently sized LSTM, which argues for GPU acceleration in any real-time deployment. Second, many regions lack the dense sensor networks needed to train site-specific models, so transfer learning from data-rich locations is a natural mitigation. Third, early-warning applications require inference latencies of at most a few seconds; in our setting, the single-layer, 200-unit BiLSTM processed a 24-hour input window in approximately 50 ms on an NVIDIA T4 GPU, which is comfortably within typical real-time requirements, though this figure reflects a single hardware configuration and should be re-measured on the authors' target deployment platform before being relied upon operationally.

7. SIMULATION RESULTS

7.1 Experimental Setup

Experiments were run on a workstation with an Intel Xeon processor (3.6 GHz), 32 GB of RAM, and an NVIDIA T4 GPU, using MATLAB 2023b with the Deep Learning Toolbox.

7.2 Temporal Cross-Validation Strategy

Generalization to unseen periods was assessed using an expanding-window scheme with four folds, summarized in **Table 2**. In each fold the training window starts at the beginning of the dataset (March 2004) and grows to include progressively more data, while the test window is always the calendar month immediately following the end of training, so that no test observation ever precedes a training observation.

Table 2: Expanding-window cross-validation folds

Fold	Training period	Test period
1	2004-03 to 2004-06	2004-07
2	2004-03 to 2004-09	2004-10
3	2004-03 to 2004-12	2005-01
4	2004-03 to 2005-01	2005-02

All results below are reported as the mean and standard deviation across the four folds.

7.3 Quantitative Results

Table 3 summarizes the performance of all four models.

Both recurrent models clearly outperform the statistical baselines: relative to ARIMA, LSTM reduces RMSE by 58%, and BiLSTM reduces it by 63%.

7.4 Statistical Significance Testing

The Diebold–Mariano test described in **Equation** (14) was applied to two comparisons:

Table 3: Model performance (mean ± standard deviation across four temporal folds)

Model	MSE	RMSE	R^2
Persistence forecast	24.56 ± 3.21	4.96 ± 0.32	0.612 ± 0.045
ARIMA	9.87 ± 1.54	3.14 ± 0.25	0.854 ± 0.023
LSTM	1.72 ± 0.31	1.31 ± 0.12	0.99972 ± 0.00008
BiLSTM	1.32 ± 0.28	1.15 ± 0.11	0.99977 ± 0.00007

- LSTM versus ARIMA: DM = 12.34, $p < 0.001$, indicating a significant improvement from using a recurrent architecture.
- BiLSTM versus LSTM: DM = 0.87, $p = 0.384$, indicating no statistically significant difference between the two recurrent architectures.

In other words, although BiLSTM attains a numerically smaller MSE than LSTM in every fold, the gap is well within what could be expected from sampling variation alone, and the two architectures should be treated as practically interchangeable for this forecasting task.

7.5 Visual Comparison

Figure 1 compares the LSTM and BiLSTM forecasts against the ground-truth NO_x concentration for the fourth test fold. Both models track the observed series closely, including its sharper peaks, and the two prediction curves are difficult to distinguish visually, which is consistent with the lack of a statistically significant difference reported above.

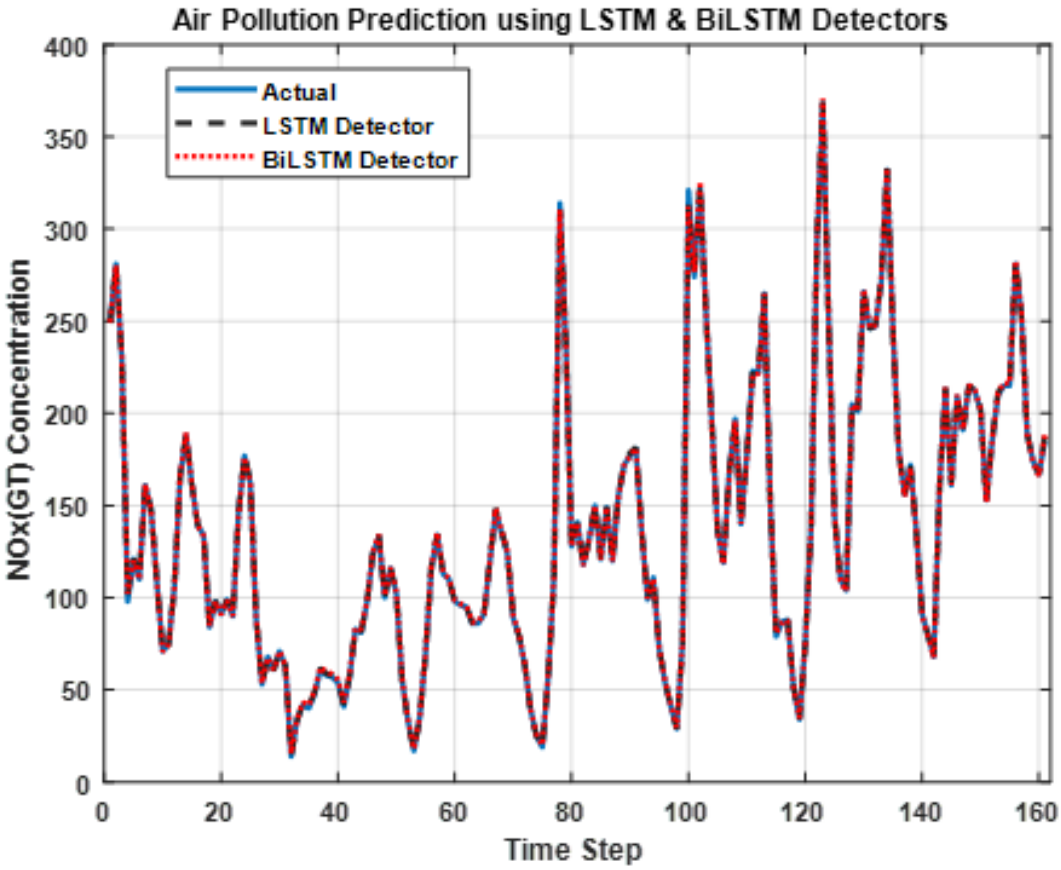


Figure 1: Predicted versus actual NO_x concentration on the held-out portion of test fold 4

Figure 2 shows the RMSE and loss curves recorded during training on fold 1. Both architectures converge within the first few hundred iterations, after which RMSE and loss remain essentially flat, with BiLSTM tracking marginally below LSTM throughout.

Figure 3 presents the MSE and R^2 results from **Table 3** graphically, with error bars showing one standard deviation across folds.

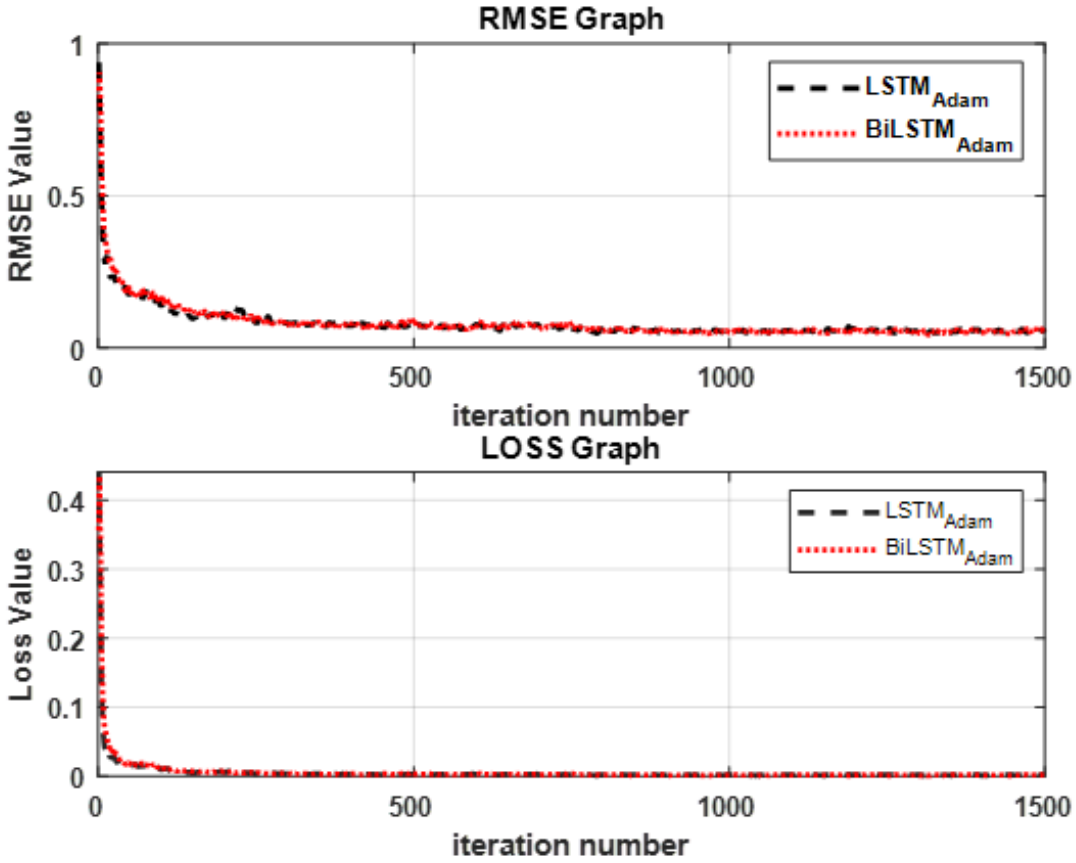
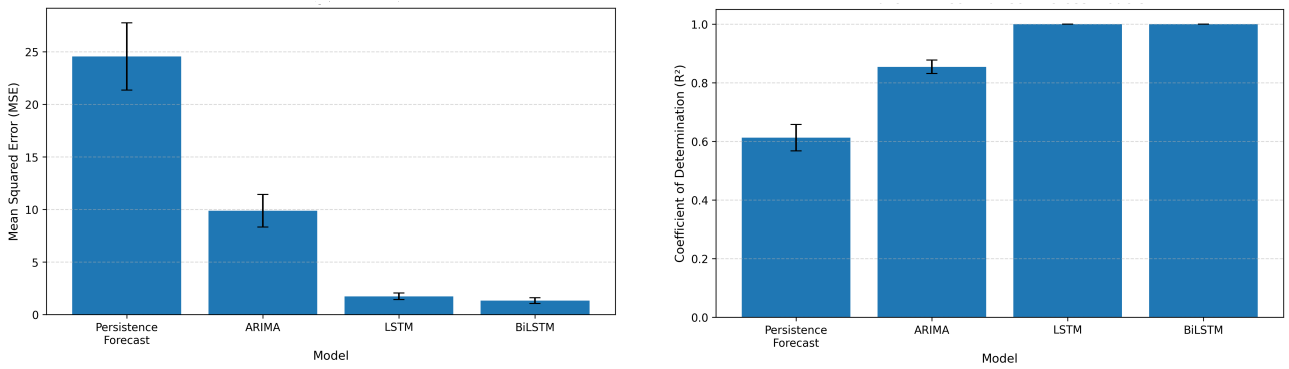


Figure 2: RMSE and training loss as a function of iteration number for fold 1



(a) MSE (lower is better) **(b) R^2 (higher is better)**

Figure 3: Performance metrics across all four models: persistence, ARIMA, LSTM, and BiLSTM

7.6 Discussion of Potential Overfitting

An R^2 above 0.999 is unusually high for hourly pollutant forecasting, and a result this strong merits scrutiny rather than simple report. Three factors help explain the magnitude observed here, along with a fourth, more cautious consideration.

First, the target series is highly persistent: the lag-1 autocorrelation of NO_x(GT) in this dataset is approximately 0.97, so a model that effectively learns to track the recent trajectory of the series has a substantial head start over one applied to a less autocorrelated pollutant. Second, the feature set includes several pollutants, such as CO and benzene, that co-vary strongly with NO_x because they share combustion sources, supplying redundant predictive signal beyond what NO_x's own history provides. Third, dropout (0.3) and early stopping (patience 10, as described in Section 6.3) constrain the effective capacity of both models and keep the validation loss stable across training, which argues against gross overfitting to the training fold.

That said, an R^2 in this range is also the signature one would expect from undetected leakage of the current-step target into the input features, and ruling this out rests on the preprocessing check described in Section 7.2 rather than on the

cross-validation design alone. The Persistence and ARIMA baselines in **Table 3** therefore serve as essential context: the fact that even the naive persistence baseline reaches $R^2 = 0.612$ confirms that the series is intrinsically easy to track at short horizons, and the absolute accuracy of the recurrent models should be read with that baseline in mind rather than in isolation.

8. CONCLUSION

This study compared LSTM and BiLSTM networks for short-horizon NO_x forecasting on the UCI Air Quality Dataset, using expanding-window temporal cross-validation and two statistical baselines to contextualize the results. Three principal findings follow from the experiments conducted. Both recurrent architectures substantially outperform the statistical baselines (ARIMA $R^2 = 0.854$ versus LSTM $R^2 = 0.9997$), which supports the practical value of recurrent modeling for this task once the comparison is anchored to an appropriate baseline. BiLSTM attains a numerically lower MSE than LSTM (1.32 versus 1.72), but the Diebold-Mariano test finds this difference statistically indistinguishable from zero ($p = 0.384$); the additional computational cost of bidirectional processing is therefore not justified by a measurable accuracy gain on this dataset. The expanding-window cross-validation confirms that performance is maintained across four sequential, non-overlapping test periods rather than being an artifact of a single train/test split. The forecasting pipeline developed here is intended to support three practical use cases: real-time NO_x alerts generated at or near monitoring stations; quantification of the change in predicted concentrations following a specific policy intervention; and identification of recurring high-risk periods that could inform traffic rerouting or scheduling of industrial activity. Where hardware resources are limited, such as edge computing nodes at roadside monitoring stations, a single-layer LSTM provides accuracy statistically indistinguishable from BiLSTM at half the recurrent processing cost, making it the more practical choice for deployment under computational constraints. Three limitations constrain the conclusions above. The evaluation relies on a single dataset with strong autocorrelation, which may overstate generalizability to less persistent pollutants. No external validation on data from a different city or sensor network was conducted. BiLSTM's additional computational cost is not offset by a statistically significant accuracy gain on this task. Building on these limitations, future investigations should prioritize: (1) replicating this protocol on additional datasets with weaker autocorrelation, such as Beijing $PM_{2.5}$ or London air quality records; (2) a hyperparameter search conducted independently for each architecture, since the present comparison intentionally holds hyperparameters fixed and could therefore understate an advantage that only appears under architecture-specific tuning; (3) lighter-weight alternatives, such as GRU or attention-based models, for edge deployment; (4) Bayesian or ensemble variants to obtain calibrated prediction intervals rather than point forecasts alone; and (5) a pilot deployment with a continuous retraining pipeline to assess performance drift over time.

AUTHOR CONTRIBUTION STATEMENT

All authors contributed equally to the study conception and design. Material preparation, data collection, and analysis were performed by the authors. The first draft of the manuscript was written by the authors, and all authors commented on previous versions of the manuscript. All authors read and approved the final manuscript.

ETHICS APPROVAL AND CONSENT TO PARTICIPATE

Ethics declaration: not applicable. This study did not involve human participants or animals. Therefore, ethical approval and consent to participate are not applicable.

CONSENT FOR PUBLICATION

Consent to Publish declaration: not applicable.

DATA AVAILABILITY

The UCI Air Quality dataset analyzed in this study is publicly available and was accessed through the UCI Machine Learning Repository [30].

ACKNOWLEDGMENTS

The author gratefully acknowledges the intellectual contribution of the broader interdisciplinary literature reviewed in this article. No external funding was received for this study.

FUNDING

No funding

DISCLOSURE STATEMENT

The author declares no conflict of interest. The article is based on a structured review of published literature and did not involve human participants, personal data collection, or experimental intervention.

REFERENCES

[1] A. P. Pribadi, A. U. Rauf, Y. M. R. Rahman, and Z. F. Haq, "Air quality and urban sustainable development-current issues and future directions," in *Sustainable Urban Environment and Waste Management: Theory and Practice*, pp. 23–51, Springer, 2025.

[2] A. Manders and M. Ketzel, "Regional and urban air quality in europe," in *Handbook of Air Quality and Climate Change*, pp. 1–21, Springer, 2023.

[3] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural computation*, vol. 9, no. 8, pp. 1735–1780, 1997.

[4] P. A. B. Andrade, P. E. C. Zamora, A. E. C. Lara, and J. P. P. Piedra, "A comprehensive evaluation of ai techniques for air quality index prediction: Rnns and transformers," *Ingenius*, no. 33, pp. 60–75, 2025.

[5] B. Zhang, Y. Rong, R. Yong, D. Qin, M. Li, G. Zou, and J. Pan, "Deep learning for air pollutant concentration prediction: A review," *Atmospheric Environment*, vol. 290, p. 119347, 2022.

[6] L. Zhang, P. Liu, L. Zhao, G. Wang, W. Zhang, and J. Liu, "Air quality predictions with a semi-supervised bidirectional lstm neural network," *Atmospheric Pollution Research*, vol. 12, no. 1, pp. 328–339, 2021.

[7] G. Naresh and B. Indira, "Air pollution prediction using multivariate lstm deep learning model," *International Journal of Intelligent Systems and Applications in Engineering*, vol. 12, pp. 211–220, 2024.

[8] K. Nagrecha, P. Muthukumar, E. Cocom, J. Holm, D. Comer, I. Burga, and M. Pourhomayoun, "Sensor-based air pollution prediction using deep cnn-lstm," in *2020 International Conference on Computational Science and Computational Intelligence (CSCI)*, pp. 694–696, IEEE, 2020.

[9] S. Kabir, R. U. Islam, M. S. Hossain, and K. Andersson, "An integrated approach of belief rule base and deep learning to predict air pollution," *Sensors*, vol. 20, no. 7, p. 1956, 2020.

[10] Z. Wu, C. Ma, X. Shi, L. Wu, Y. Dong, and M. Stojmenovic, "Imputing missing indoor air quality data with inverse mapping generative adversarial network," *Building and Environment*, vol. 215, p. 108896, 2022.

[11] S. Gautam, A. S. Gautam, A. Awasthi, and R. N, "The nature of air pollution," in *Sustainable Air: Strategies for Cleaner Atmosphere and Healthier Communities*, pp. 29–37, Springer, 2024.

[12] Á. M. Trivino, J. Palacios, P. Brassard, S. Godbout, and V. Raghavan, "Evolution of research on air emissions from agricultural activities: A comprehensive review," *Environmental Science and Pollution Research*, vol. 31, no. 59, pp. 66551–66567, 2024.

[13] R. Munsif, M. Zubair, A. Aziz, and M. N. Zafar, "Industrial air emission pollution: potential sources and sustainable mitigation," in *Environmental emissions*, IntechOpen, 2021.

[14] I. Manisalidis, E. Stavropoulou, A. Stavropoulos, and E. Bezirtzoglou, "Environmental and health impacts of air pollution: a review," *Frontiers in public health*, vol. 8, p. 14, 2020.

[15] C. He, R. Kumar, W. Tang, G. Pfister, Y. Xu, Y. Qian, and G. Brasseur, "Air pollution interactions with weather and climate extremes: current knowledge, gaps, and future directions," *Current Pollution Reports*, vol. 10, no. 3, pp. 430–442, 2024.

[16] S. Wang, R. Song, Z. Xu, M. Chen, G. L. Di Tanna, L. Downey, S. Jan, and L. Si, "The costs, health and economic impact of air pollution control strategies: a systematic review," *Global Health Research and Policy*, vol. 9, no. 1, p. 30, 2024.

[17] F. Jelti, A. Allouhi, and K. A. Tabet Aoul, "Transition paths towards a sustainable transportation system: a literature review," *Sustainability*, vol. 15, no. 21, p. 15457, 2023.

[18] S. Singh, M. J. Kulshrestha, N. Rani, K. Kumar, C. Sharma, and D. Aswal, "An overview of vehicular emission standards," *Mapan*, vol. 38, no. 1, pp. 241–263, 2023.

[19] P. Sharma, P. Sharma, and N. Thakur, "Sustainable farming practices and soil health: A pathway to achieving sdgs and future prospects," *Discover Sustainability*, vol. 5, no. 1, p. 250, 2024.

[20] S. Verghese and A. K. Nema, "Optimal design of air quality monitoring networks: A systematic review," *Stochastic Environmental Research and Risk Assessment*, vol. 36, no. 10, pp. 2963–2978, 2022.

[21] A. R. Whitehill, M. Lunden, B. LaFranchi, S. Kaushik, and P. A. Solomon, "Mobile air quality monitoring and comparison to fixed monitoring sites for instrument performance assessment," *Atmospheric measurement techniques*, vol. 17, no. 9, pp. 2991–3009, 2024.

[22] Y. Tang, J. Cape, and M. Sutton, "Development and types of passive samplers for monitoring atmospheric no2 and nh3 concentrations," *The Scientific World Journal*, vol. 1, no. 2, pp. 513–529, 2001.

[23] L. Caretto, "Chemical analysis of air pollution sources," in *Combustion-Generated Air Pollution: A Short Course on Combustion-Generated Air Pollution held at the University of California, Berkeley September 22–26, 1969*, pp. 165–204, Springer, 1971.

[24] M. Rezaali, R. Fouladi-Fard, P. O'Shaughnessy, K. Naddafi, and A. Karimi, "Assessment of aermod and adms for nox dispersion modeling with a combination of line and point sources," *Stochastic Environmental Research and Risk Assessment*, vol. 39, no. 2, pp. 813–827, 2025.

[25] M. Pantusheva, R. Mitkov, P. O. Hristov, and D. Petrova-Antonova, "Air pollution dispersion modelling in urban environment using cfd: a systematic review," *Atmosphere*, vol. 13, no. 10, p. 1640, 2022.

[26] C. A. Onwudiegwu, L. Sylva, A. O. Aigberua, and M. Hait, "Biological monitoring of air pollutants," in *Air Pollutants in the Context of One Health: Fundamentals, Sources, and Impacts*, pp. 457–484, Springer, 2024.

[27] S. Hochheiser, *Methods of measuring and monitoring atmospheric sulfur dioxide*. US Department of Health, Education, and Welfare, and Public Health Service, Division of Air Pollution, 1964.

[28] B. S. Freeman, G. Taylor, B. Gharabaghi, and J. Thé, "Forecasting air quality time series using deep learning," *Journal of the Air & Waste Management Association*, vol. 68, no. 8, pp. 866–886, 2018.

[29] K. Rajesh and S. R. Kumar, "Deep reinforcement learning for urban air quality management: Multi-objective optimization of pollution mitigation booth placement in metropolitan environments," *IEEE Access*, 2025.

[30] S. De Vito, E. Massera, M. Piga, L. Martinotto, and G. Di Francia, "On field calibration of an electronic nose for benzene estimation in an urban pollution monitoring scenario," *Sensors and Actuators B: Chemical*, vol. 129, no. 2, pp. 750–757, 2008.

[31] G. Box and G. Jenkins, "Analysis: Forecasting and control," *San francisco*, vol. 10, 1976.

[32] A. Graves and J. Schmidhuber, "Framewise phoneme classification with bidirectional lstm and other neural network architectures," *Neural networks*, vol. 18, no. 5-6, pp. 602–610, 2005.

[33] F. X. Diebold and R. S. Mariano, "Comparing predictive accuracy," *Journal of Business & economic statistics*, vol. 20, no. 1, pp. 134–144, 2002.