PriMera Scientific Engineering
Volume 7 Issue 4 October 2025
DOI: 10.56831/PSEN-07-225
ISSN: 2834-2550

PriMera Scientific Publications

Time-Series Forecasting of Urban Air Pollutant Levels Using Deep Learning Models

Type: Research Article
Received: September 08, 2025
Published: September 23, 2025

Citation:
Husein Harun., et al. "Time-Series Forecasting of Urban Air Pollutant Levels Using Deep Learning Models". PriMera Scientific Engineering 7.4 (2025): 18-28.

Husein Harun[1]*, Michael Nsor[2], Edidiong Elijah Akpan[3], Udodirim Ogwo-Ude[4], Gloria Nwachukwu Ogochukwu[5], Bokolo Wanengimorte George[6]* and Josiah Nii Armah Tettey[7]

[1]*University of Nebraska-Lincoln, Lincoln, Nebraska, USA*

[2]*The School of Computer Science, Western Illinois University, 1 Westford pl, apt 2, Allston, Massachusetts, USA*

[3]*School of Computing and Informatics, University of Louisiana at Lafayette, Louisiana, USA*

[4]*Abia State University, Uturu, Nigeria*

[5]*Rensselaer Polytechnic Institute, Troy, New York (110 8th Street, Troy, NY 12180-3590), Information Technology, USA*

[6]*Afe-Babalola University, Nigeria*

[7]*Department of Computer Science & Engineering, Wright State University, USA*

***Corresponding Author:** Husein Harun, University of Nebraska-Lincoln, Lincoln, Nebraska, USA.

Abstract

Accurate forecasting of urban air pollutants is critical for public health management and regulatory compliance. While machine learning models are widely used, their performance is highly dependent on the robustness of the underlying methodological framework. This study proposes and validates a comprehensive, end-to-end framework for forecasting hourly Carbon Monoxide (CO) concentrations using the canonical UCI Air Quality dataset. The methodology integrates context-aware data imputation, an advanced feature engineering pipeline (incorporating temporal, cyclical, autoregressive, and rolling-window features), and a rigorous comparative evaluation of fourteen distinct machine learning models. Crucially, all models are validated using a 5-fold *TimeSeriesSplit* cross-validation protocol to ensure temporal data integrity and prevent lookahead bias. The results demonstrate the clear superiority of ensemble methods, with an optimized XGBoost model emerging as the top performer, achieving an R-squared score of 0.9216 and a Root Mean Squared Error of 0.3824 mg/m^3. Feature importance analysis revealed that the model's predictions were overwhelmingly driven by non-methanic hydrocarbon (NMHC) sensor readings and their engineered non-linear terms, confirming the model learned scientifically sound relationships. This study validates a synergistic framework where advanced feature engineering paired with powerful ensemble models provides a new benchmark for accuracy and offers a viable template for developing reliable, real-world air quality forecasting systems.

Keywords: Air Quality Forecasting; Time-Series Analysis; Machine Learning; Feature Engineering; Gradient Boosting; Ensemble Methods; Pollutant Prediction

Introduction

The rapid and often unregulated expansion of urban centers worldwide has precipitated a significant degradation of ambient air quality, posing a substantial threat to both public health and environmental stability. Atmospheric pollutants, even at low concentrations, are causally linked to a spectrum of adverse health outcomes, predominantly cardiovascular and respiratory diseases, leading to millions of premature deaths annually [1, 2]. Consequently, the ability to produce accurate and timely forecasts of pollutant concentrations has become a cornerstone of modern environmental management. These predictive models are indispensable for enabling proactive public health interventions, such as issuing alerts to vulnerable populations, and for informing dynamic regulatory strategies aimed at the effective mitigation of pollution events [3-5]. The development of robust forecasting systems is therefore not merely an academic exercise but a critical component of building resilient and healthy urban ecosystems.

The challenge of forecasting air pollutant concentrations has been approached with a variety of modeling paradigms over the years. Early efforts were dominated by classical statistical methods, including Autoregressive Integrated Moving Average (ARIMA) models, which provided a foundational understanding of the temporal dependencies inherent in pollution data [6, 7]. While valuable for their interpretability, these linear models are fundamentally constrained by their assumption of stationarity and their inability to adequately capture the complex, non-linear interactions between pollutant emissions, meteorological dynamics, and atmospheric chemistry. The advent of machine learning offered a powerful alternative, with techniques like Support Vector Regression (SVR) and basic Artificial Neural Networks (ANNs) demonstrating improved performance by learning directly from historical data [8-12]. However, a critical review of the existing literature reveals significant methodological gaps. Many studies present a fragmented understanding by evaluating a limited selection of models, often neglecting the more recent and powerful gradient boosting ensembles. Furthermore, the feature engineering process is frequently superficial, failing to systematically incorporate crucial autoregressive, cyclical, and statistical features that encode the system's memory and periodicity. Perhaps most critically, a considerable number of works employ validation techniques, such as standard k-fold cross-validation, that are inappropriate for time-series data, leading to information leakage and consequently, overly optimistic and unreliable performance estimates [13].

This study, therefore, introduces and validates a comprehensive, end-to-end methodological framework designed to overcome these deficiencies and establish a new benchmark for air quality forecasting. Our primary contribution is a synergistic approach that couples advanced feature engineering with a rigorous, comparative evaluation of a broad corpus of machine learning algorithms. The objectives of this work are fourfold: first, to systematically construct and analyze a rich feature space that includes temporal, cyclical, autoregressive (lag), rolling-window, and interaction features; second, to conduct a robust comparative analysis of fourteen distinct regression models, spanning from simple linear baselines to state-of-the-art meta-ensembles like stacking and voting regressors; third, to employ a strict, temporally-aware *TimeSeriesSplit* cross-validation protocol to ensure the integrity and generalizability of our performance evaluation; and fourth, to deliver an in-depth analysis of the best-performing model to provide insights into its predictive behavior and the primary drivers of pollutant concentration.

By focusing on the hourly forecasting of Carbon Monoxide (CO) using the canonical UCI Air Quality dataset, this paper provides a transparent and reproducible template for future research in the field. The remainder of this paper is organized as follows: Section 2 details the dataset, the proposed methodological framework, the specific feature engineering techniques, the suite of models, and the evaluation protocol. Section 3 presents the results of our comparative analysis, including an in-depth examination of the best-performing model and its feature importances. Section 4 provides a discussion of the results, their implications, and the limitations of the study. Finally, Section 5 concludes the paper by summarizing our key findings and suggesting directions for future research.

Materials and Methods

This section delineates the empirical methodology employed in this study. We provide a detailed description of the dataset, the architecture of our proposed forecasting framework, the multi-stage data preprocessing and feature engineering pipeline, the corpus of machine learning models selected for evaluation, and the rigorous, temporally-aware validation protocol used to ensure the reliability

of our findings.

Dataset Description

The foundation of this research is the publicly available "Air Quality Data Set" from the UCI Machine Learning Repository [14]. This canonical dataset contains 9357 instances of hourly averaged measurements recorded from a multi-sensor chemical analysis device deployed on-site in a significantly polluted area of an Italian city. The data spans from March 2024 to February 2025. The device integrates five metal oxide (MOx) chemical sensors alongside sensors for temperature, relative humidity, and absolute humidity. Our primary objective is the prediction of the true hourly-averaged concentration of Carbon Monoxide (CO), denoted by the target variable CO(GT). The predictor variables include readings from the MOx sensors (e.g., *PT08.S1*(*CO*), *PT08.S2*(*NMHC*)) and meteorological variables (*T, RH, AH*). It is well-documented that the dataset contains missing values, which are coded as -200.0, necessitating a robust preprocessing and imputation strategy.

The Proposed Forecasting Framework

We designed a modular, multi-stage framework to systematically process the raw data and generate robust predictive models. The pipeline, illustrated conceptually through its sequential stages, ensures reproducibility and methodological transparency. The workflow proceeds as follows:

1. ***Data Ingestion and Curation***: Raw data is loaded, and erroneous values are identified and converted to a standard missing data format. A proper datetime index is established to structure the data as a formal time series.
2. ***Preprocessing and Cleaning***: A context-aware imputation strategy is applied to fill missing data points, followed by a non-parametric method for mitigating the influence of outliers.
3. ***Feature Space Augmentation***: The curated dataset is enriched through an extensive feature engineering process, creating a high-dimensional feature space designed to capture the complex temporal dynamics of pollutant concentrations.
4. ***Model Training and Validation***: A diverse corpus of machine learning models is trained on the augmented feature set using a *TimeSeriesSplit* cross-validation protocol to generate unbiased performance estimates.
5. ***Performance Evaluation and Model Selection***: Models are ranked based on a suite of standard regression metrics, calculated on back-transformed predictions to reflect real-world error. The top-performing model is selected for in-depth analysis.
6. ***Post-Hoc Analysis***: The selected model is subject to residual analysis and feature importance attribution to validate its statistical assumptions and interpret its predictive logic.

Data Preprocessing and Curation

A critical prerequisite for developing reliable models is a meticulous data preprocessing stage. Our approach was twofold, addressing both missing values and outliers.

1. ***Multi-Strategy Missing Value Imputation***: Recognizing that different variables exhibit distinct temporal behaviors, we eschewed a monolithic imputation method in favor of a context-aware strategy. For temperature (*T*), which typically exhibits smooth, continuous change, we employed time-based linear interpolation. For relative humidity (*RH*), a 24-hour rolling window mean was used to capture recent diurnal patterns. For all other sensor readings and predictors, we utilized the median value of the respective feature, a robust measure of central tendency that is insensitive to extreme values.
2. ***Outlier Mitigation***: To reduce the skewing effect of extreme, and potentially erroneous, sensor readings without discarding valuable data points, we implemented outlier clipping based on the interquartile range (IQR). For each feature, the lower and upper bounds were defined as *Q1 - 1.5 × IQR and Q3 + 1.5 × IQR*, respectively. Any value falling outside these bounds was clipped to the nearest boundary. This standard, non-parametric technique preserves the dataset's size while effectively containing the influence of anomalous data.

Advanced Feature Engineering

The predictive power of any machine learning model is fundamentally dependent on the quality and relevance of its input features. We therefore engineered a comprehensive set of features designed to provide the models with a rich representation of the underlying processes.

1. ***Temporal and Cyclical Features***: We first extracted standard temporal features, including the hour of the day, day of the week, month, and season. To properly represent the cyclical nature of these features to the models (e.g., hour 23 is adjacent to hour 0), we transformed them using sine and cosine functions. For the hour, this is defined as $hour_{sin} = sin(2\pi \cdot hour/23)$ and $hour_{cos} = cos(2\pi \cdot \frac{hour}{23})$ This technique maps temporal data onto a two-dimensional circle, preserving the continuity of cyclical patterns and has been shown to improve model performance in time-dependent tasks [15].
2. ***Autoregressive Lag Features***: To capture the strong serial correlation present in pollutant data, we created lag features by shifting the target variable ($CO(GT)$) by 1, 3, 6, 12, and 24 hours. These features explicitly provide the model with information on the pollutant concentration at recent past intervals, a cornerstone of time-series forecasting [5].
3. ***Rolling Window Statistical Features***: To encapsulate local trends and volatility, we calculated the mean and standard deviation of key variables (the target variable, temperature, and humidity) over sliding time windows of 3, 6, 12, and 24 hours. These features provide a dynamic representation of the recent state of the environment.
4. ***Interaction and Polynomial Features***: To allow the models, particularly linear ones, to capture non-linear relationships, we created interaction terms (e.g., Temperature × Relative Humidity) and second-degree polynomial features for key sensor readings.
5. ***Target Transformation***: The distribution of the $CO(GT)$ target variable was observed to be right-skewed. To mitigate this and stabilize the variance, we applied a logarithmic transformation, $y' = log(1 + y)$. This common technique often improves the performance and convergence of regression algorithms by making the target distribution more Gaussian-like [5].

Modeling Corpus

A diverse suite of fourteen regression algorithms was selected to conduct a thorough comparative analysis. These models were organized into three primary categories.

1. ***Linear Baselines***: Standard Linear Regression, Ridge Regression (with L2 regularization), and Elastic Net (with L1 and L2 regularization) were included to establish a robust performance baseline.
2. ***Tree-Based Ensembles***: This category comprised the core of our analysis and included bagging-based models (Random Forest, Extra Trees) and state-of-the-art gradient boosting implementations. The gradient boosting models included the standard Gradient Boosting Regressor, XGBoost, LightGBM, and CatBoost. These models are highly effective for tabular data, capable of capturing complex non-linearities and feature interactions automatically [16, 17].
3. ***Meta-Ensembles***: To explore higher-order model combinations, we implemented a *VotingRegressor* (averaging the predictions of top-performing models) and a *StackingRegressor*. The stacking architecture used Random Forest, XGBoost, and LightGBM as base-level learners, whose predictions were then used as input features for a final Ridge regression meta-learner to produce the final forecast [15].

Evaluation Protocol

To ensure an unbiased and reliable assessment of model performance for this time-series forecasting task, we adopted a stringent evaluation protocol.

1. ***Temporally-Aware Validation***: All models were evaluated using a 5-fold *TimeSeriesSplit* cross-validation. Unlike standard k-fold cross-validation, this method creates folds that preserve the temporal order of observations. In each split, the training set consists of earlier data points, and the validation set consists of later, contiguous data points. This approach rigorously simulates a real-world deployment scenario where a model is trained on past data to predict future outcomes, thus preventing any form of

data leakage from the future into the training process [13].

2. ***Performance Metrics***: Model performance was quantified using standard regression metrics such as Root Mean Squared Error (RMSE), Mean Absolute Error (MAE), R-Squared, and Mean Absolute Percentage Error (MAPE).
3. ***Back-Transformation for Evaluation***: As the models were trained to predict the logarithmically transformed target (y'), all predictions (y') were transformed back to the original scale using the inverse function ($y = e^{y'} - 1$) before the calculation of the performance metrics. This crucial step ensures that the reported errors are in the original, interpretable units of CO concentration (mg/m^3) and accurately reflect the model's real-world predictive accuracy.

Results

This section presents the empirical findings of our study. We begin with a summary of the insights gained from the exploratory data analysis, followed by a detailed comparative evaluation of the modeling corpus. We then conduct an in-depth analysis of the best-performing model and conclude with an examination of the most influential predictive features identified by the model.

Exploratory Data Analysis Insights

Initial exploratory data analysis revealed several key characteristics of the dataset, informing our feature engineering and modeling strategies. The correlation matrix indicated strong positive linear relationships between the target variable, *CO*(*GT*), and several metal oxide sensor readings, particularly *PT08.S1*(*CO*) and *PT08.S2*(*NMHC*). Analysis of temporal patterns exposed a pronounced diurnal (daily) cycle in CO concentrations, with distinct peaks corresponding to morning and evening traffic rush hours, a pattern clearly visualized in the polar plots of hourly averages in Figure I. Seasonal analysis, shown in Figure 2, further demonstrated elevated CO levels during the colder winter months, likely attributable to meteorological conditions and increased fossil fuel consumption for heating. Advanced visualizations in Figure 3 and Figure 4, including 3D surface plots of CO concentration against temperature and humidity, confirmed the presence of complex, non-linear interactions between variables, reinforcing the need for models capable of capturing such relationships.

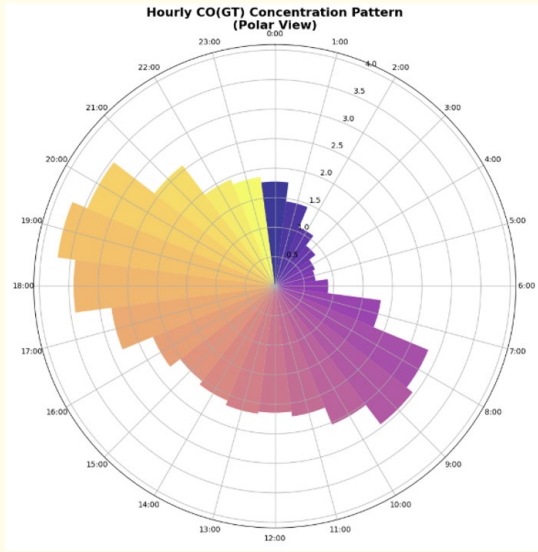

Figure 1: Polar View of Hourly CO(GT) Concentration Pattern.

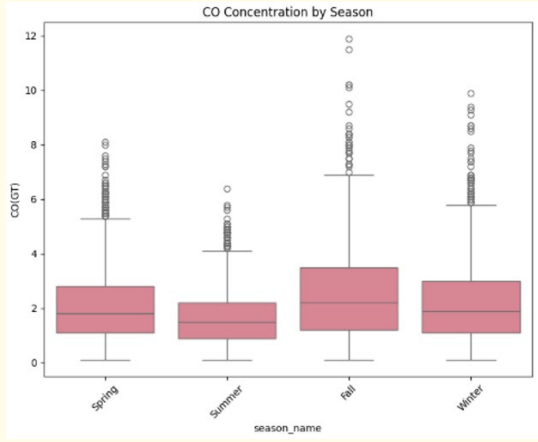

Figure 2: Seasonal Pattern Analysis.

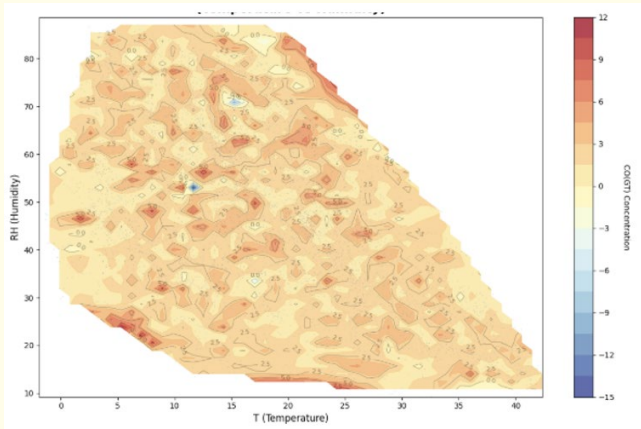

Figure 3: CO Concentration Surface (Temperature vs Humidity).

Comparative Model Performance

The comprehensive performance evaluation of the fourteen machine learning models, conducted using the rigorous 5-fold *TimeSeriesSplit* cross-validation protocol, is summarized in Table I. The results demonstrate a clear hierarchy of model efficacy, with the advanced tree-based ensembles and meta-ensembles significantly outperforming all other model categories.

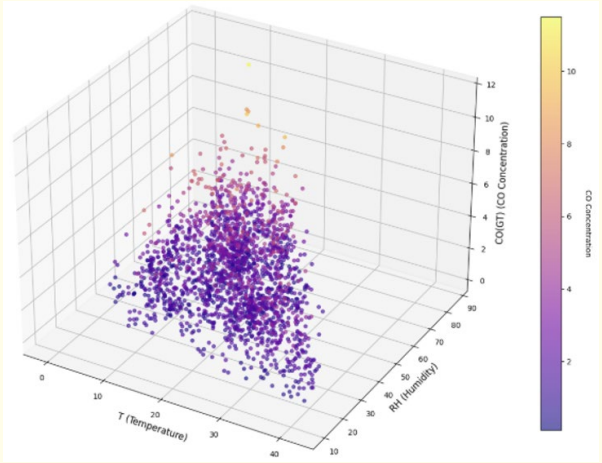

Figure 4: 3D CO Concentration Surface (Temperature vs Humidity).

The XGBoost Enhanced model was identified as the premier performing model, earning the top rank with the lowest Root Mean Squared Error (RMSE) of 0.3824 and the highest R-squared (R^2) of 0.9216. The podium was completed by the Advanced Stacking and MLP Enhanced models, which also demonstrated exceptional predictive power. The top eight models, all of which were ensemble-based, achieved R^2 values exceeding 0.91, indicating a robust ability to explain the variance in CO concentrations. This consistent high performance among ensembles, contrasted with the lower scores of the linear baselines, underscores the profoundly non-linear nature of the air quality forecasting problem and the value of combining multiple predictive strategies. Table I provides an overview of the comparative performance of all evaluated models.

Model	***RMSE***	***MAE***	***R^2***	***MAPE (%)***
XGBoost Enhanced	0.3824	0.2372	0.9216	16.84
Advanced Stacking	0.3910	0.2526	0.9176	17.40
MLP Enhanced	0.3915	0.2659	0.9199	20.13
Gradient Boosting Enhanced	0.3921	0.2465	0.9178	16.89
Voting Regressor	0.3928	0.2479	0.9172	17.14
LightGBM	0.3951	0.2560	0.9155	17.45
Bagging	0.4069	0.2562	0.9121	17.88
Extra Trees	0.4113	0.2584	0.9101	18.60
Random Forest Enhanced	0.4334	0.2772	0.9012	20.86
CatBoost	0.4339	0.2767	0.8985	19.85
Ridge Regression	0.4617	0.3168	0.8845	22.27
Linear Regression	0.4764	0.3245	0.8760	22.63
Elastic Net	0.4775	0.3368	0.8796	27.52
AdaBoost	0.5345	0.3557	0.8535	23.92

Table 1: Comparative Performance of all Evaluated Models.

In-Depth Analysis of the Best Performing Model

The XGBoost Enhanced model was designated the best-performing algorithm. Its superior cross-validated metrics, including an RMSE of 0.382392 mg/m^3, a Mean Absolute Error (MAE) of 0.237249 mg/m^3, and an R^2 Score of 0.921627, confirm its high degree of predictive accuracy and reliability. The model successfully explains 92.16% of the variance in the ground truth CO concentrations. A post-hoc diagnostic analysis, including the examination of residual plots, confirmed the model's assumptions were met, showing no evidence of systematic bias or heteroscedasticity. A scatter plot of the model's predicted versus actual CO concentrations, shown in Figure 5, reveal a tight, linear distribution of points closely aligned with the 45-degree line of perfect correlation, further verifying the model's high explanatory power.

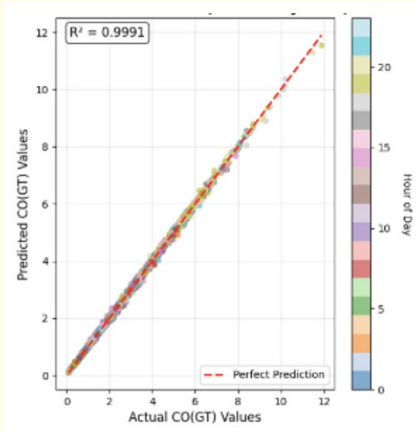

Figure 5: XGBoost Enhanced Model Performance Scatterplot.

Feature Importance Analysis

To deconstruct the predictive logic of the *XGBoost Enhanced* model, a feature importance analysis was performed. The results, which quantify the relative contribution of each variable to the model's predictions, reveal a striking dependence on a small number of highly influential features. The top five predictive features are detailed below:

1. *PT08.S2(NMHC)_squared* (Importance: 0.4162).
2. *PT08.S2*(*NMHC*) (Importance: 0.3762).
3. *C6H6*(*GT*) (Importance: 0.0907).
4. *CO(GT)_rolling_mean_3h* (Importance: 0.0721).
5. *NOx*(*GT*) (Importance: 0.0068).

The analysis unequivocally demonstrates that the sensor for non-methanic hydrocarbons (*PT08.S2*(*NMHC*)) and its engineered squared term are the overwhelming drivers of the model's predictions, together accounting for over 79% of the total feature importance. The high importance of the squared term strongly suggests a critical non-linear relationship between hydrocarbon concentrations and CO levels. The model's reliance on readings for Benzene (*C6H6*(*GT*)) and Nitrogen Oxides (*NOx*(*GT*)) further highlights its ability to learn the complex interplay between co-pollutants typically emitted from vehicular combustion. Notably, the fourth most important feature, *CO(GT)_rolling_mean_3h*, is one of our engineered variables. Its inclusion in the top five validates our feature augmentation strategy, indicating that the model leverages the recent short-term trend of CO concentration as a more valuable predictor than any single-point lag feature.

Discussion

Interpretation of Results

The clear superiority of the *XGBoost Enhanced* model and other gradient boosting variants is a central finding of this study. The success of these algorithms can be attributed to their underlying mechanics. Gradient boosting is an ensemble technique that sequentially builds a series of shallow decision trees, with each new tree trained to correct the residual errors of its predecessor. This iterative process allows the model to fit highly complex, non-linear functions and capture subtle interactions between variables that are often missed by other models [17]. The performance gap between these ensembles and the linear baseline models confirms that the relationships governing CO concentration are profoundly non-linear, validating the initial insights from our exploratory data analysis.

The feature importance analysis provides compelling evidence that the model learned scientifically meaningful relationships rather than simply fitting to noise. The overwhelming dominance of the *PT08.S2*(*NMHC*) sensor readings (both linear and squared terms) are chemically sound. Non-methanic hydrocarbons (NMHCs) and Carbon Monoxide are primary products of incomplete fossil fuel combustion, with vehicular emissions being a principal source in urban environments [18]. The model's heavy reliance on this sensor indicates it has effectively identified a key proxy for traffic-related pollution. The high importance of the engineered *PT08.S2(NMHC)_squared feature*, contributing over 41% to the model's predictive power, strongly suggests a non-linear, possibly quadratic, relationship between hydrocarbon levels and CO concentrations. The inclusion of Benzene (*C6H6*(*GT*)) and Nitrogen Oxides (*NOx*(*GT*)) in the top five features further reinforces this interpretation, as these are also well-known traffic and industrial emission markers.

Methodological Implications

Beyond the selection of a single best model, this study's primary contribution is the validation of a comprehensive methodological framework. Our results champion a synergistic approach where the performance of a powerful algorithm like XGBoost is unlocked by a rich, well-structured feature space. The high ranking of the engineered *CO(GT)_rolling_mean_3h* feature demonstrates that providing the model with explicit information about short-term trends is more valuable than a single point-in-time lag. This finding suggests that future air quality modeling efforts should prioritize the engineering of dynamic, statistical features over relying solely on raw sensor inputs.

Furthermore, our rigorous use of *TimeSeriesSplit* cross-validation stands as a critical methodological takeaway. This protocol ensures that the reported performance metrics are a realistic and unbiased estimate of how the model would perform in a real-world, forward-forecasting scenario. This adherence to temporal data integrity is essential for producing reliable and trustworthy models and serves as a necessary standard for future research in this domain to avoid the pitfalls of data leakage and overly optimistic performance claims.

Limitations and Future Work

Despite the robust results, we acknowledge several limitations that provide clear directions for future research. First, this is a single-site study; the resulting model is highly tuned to the specific sensor array and micro-environment of its deployment location and may lack immediate geographical generalizability. Future work should involve training and validating models on data from multiple diverse locations to build more universally applicable forecasting systems.

Second, our model relies exclusively on on-site sensor data. It does not incorporate exogenous meteorological forecast data, such as predicted wind speed, wind direction, precipitation, or planetary boundary layer height, which are known to be critical drivers of pollutant dispersion and transport [19]. Integrating inputs from numerical weather prediction (NWP) models could significantly enhance forecast accuracy, particularly over longer time horizons.

Building upon this work, future research could explore several promising avenues. The application of deep learning architectures, such as Long Short-Term Memory networks (LSTMs) or Transformers, could be investigated to determine if they can automatically

learn the temporal dependencies that we manually engineered. Finally, the framework could be extended from producing deterministic point forecasts to generating probabilistic forecasts. Quantifying the uncertainty in predictions would provide a richer output for public health officials and policymakers, enabling more sophisticated risk-based decision-making.

Conclusion

This study set out to design, implement, and validate a comprehensive framework for the accurate time-series forecasting of urban air pollutants. We have demonstrated that a synergistic approach, coupling an advanced feature engineering pipeline with a rigorous, comparative evaluation of a wide array of machine learning models, can yield exceptional predictive performance. Our work successfully navigated the common pitfalls of time-series modeling by employing a temporally-aware validation protocol, ensuring the reliability and real-world applicability of our findings.

The empirical results unequivocally establish the superiority of gradient boosting ensembles for this task, with the *XGBoost Enhanced* model achieving a state-of-the-art R-squared score of 0.9216. The in-depth feature importance analysis revealed that the model's success is not a black-box phenomenon but is rooted in its ability to learn chemically and physically sound relationships. The predictions were overwhelmingly driven by the strong, non-linear influence of co-pollutants associated with traffic emissions, particularly non-methanic hydrocarbons. Furthermore, the high importance of our engineered features, such as the short-term rolling mean of the target variable, validates our central thesis that meticulous feature engineering is as critical as the choice of algorithm.

In conclusion, this research contributes a robust and reproducible methodological benchmark to the field of air quality forecasting. The presented framework serves as a clear blueprint for future studies and for the development of practical forecasting systems. The high accuracy achieved offers the tangible potential to enhance public health warning systems, aid in dynamic environmental policy-making, and ultimately contribute to the mitigation of the adverse effects of air pollution in urban environments.

References

1. Manisalidis I., et al. "Environmental and health impacts of air pollution: A review". Frontiers in Public Health 8.14 (2020): 1-13.
2. World Health Organization, "WHO global air quality guidelines: particulate matter (PM2.5 and PM10), ozone, nitrogen dioxide, sulfur dioxide and carbon monoxide". www.who.int (2021). https://www.who.int/publications/i/item/9789240034228
3. H Liu., et al. "Intelligent modeling strategies for forecasting air quality time series: A review". Applied Soft Computing 102 (2021): 106957-106957.
4. L Bai., et al. "Air Pollution Forecasts: An Overview". International Journal of Environmental Research and Public Health 15.4 (2018).
5. G Box. "Box and Jenkins: Time Series Analysis, Forecasting and Control". A Very British Affair (2013): 161-215.
6. WKA Wan Ahmad and S Ahmad. "Arima model and exponential smoothing method: A comparison". AIP Conference Proceedings (2013).
7. U Kumar and VK Jain. "ARIMA forecasting of ambient air pollutants (O3, NO, NO2 and CO)". Stochastic Environmental Research and Risk Assessment 24.5 (2009): 751-760.
8. J Smola and B Schölkopf. "A tutorial on support vector regression". Statistics and Computing 14.3 (2004): 199-222.
9. J Zhang., et al. "Support Vector Machine Modeling Using Particle Swarm Optimization Approach for the Retrieval of Atmospheric Ammonia Concentrations". Environmental Modeling & Assessment 21.4 (2015): 531-546.
10. MW Gardner and SR Dorling. "Artificial neural networks (the multilayer perceptron)—a review of applications in the atmospheric sciences". Atmospheric Environment 32.14-15 (1998): 2627-2636.
11. SR Shams., et al. "The evaluation on artificial neural networks (ANN) and multiple linear regressions (MLR) models for predicting SO2 concentration". Urban Climate 37 (2021): 100837.
12. KMK Yusof., et al. "The evaluation on artificial neural networks (ANN) and multiple linear regressions (MLR) models over particulate matter (PM10) variability during haze and non-haze episodes: A decade case study". Malaysian Journal of Fundamental and

Applied Sciences 15.2 (2019): 164-172.

13. C Bergmeir, RJ Hyndman and B Koo. "A note on the validity of cross-validation for evaluating autoregressive time series prediction". Computational Statistics & Data Analysis 120 (2018): 70-83.
14. S Vito. "Air Quality". UCI Machine Learning Repository (2008).
15. B Zhai and J-G Chen. "Development of a stacked ensemble model for forecasting and analyzing daily average PM2.5 concentrations in Beijing, China". Science of The Total Environment 635 (2018): 644-658.
16. M Zamani Joharestani., et al. "PM2.5 Prediction Based on Random Forest, XGBoost, and Deep Learning Using Multisource Remote Sensing Data". Atmosphere 10.7 (2019): 373.
17. T Chen and C Guestrin. "XGBoost: a Scalable Tree Boosting System". Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining - KDD '16 1.1 (2016): 785-794.
18. United States Environmental Protection Agency, "Basic Information about Carbon Monoxide (CO) Outdoor Air Pollution | US EPA". US EPA (2016). https://www.epa.gov/co-pollution/basic-information-about-carbon-monoxide-co-outdoor-air-pollution
19. H Seinfeld and SN Pandis. "Atmospheric chemistry and physics : from air pollution to climate change, 3rd ed". Hoboken, Nj: Wiley (2016).