International Journal of Engineering, Management and Humanities (IJEMH)
Volume 6, Issue 2, Mar.-Apr, 2025 pp: 344-350 ISSN: 2584-2145 www.ijemh.com

Comparative Analysis of Machine Learning Models forAtmospheric CO_2 Prediction

Dhyey Savaliya, Shlok Jain, Utsav Patiyat, Akshat Patel, Rozal Khan, Dr.Shilpa Sonawani
dept. of Computer Engineering and Technology Dr.VishwanathKaradMITWorldPeaceUniversity
Pune, India

Date of Submission: 20-04-2025 Date of Acceptance: 02-05-2025

***Abstract*—Atmospheric carbon dioxide ($CO2$) is a critical driver of global climate dynamics,an accurate forecasting of its future concentrations is imperative for forming policy inter- ventions, mitigating the effects of climate change, and advancing understanding of environmental transformations. Reliable pre- dictions of $CO2$ trends enable policymakers and scientists to better anticipate climate-related impact and implement effective strategies for adaptation and sustainability. In this study, we present a comprehensive comparative analysis of eight machine learning models applied to the long-term $CO2$ data set obtained from the Mauna Loa Observatory, operated by the National Oceanic and Atmospheric Administration (NOAA). The models under investigation include linear regression,decision tree regressor, random forest regression, support vector regression (SVR), gradient boosted regression, XGBoost, autoregressive integrated moving average (ARIMA) and long-short-term memory (LSTM) neural networks. The performance of the model is rigorously evaluated using the root mean square error (RMSE) and the coefficient of determination (R^2 score) as primary evaluation metrics. The results demonstrate that while traditional models such as Linear Regression and Decision Trees offer baseline benchmark switch minimal computation complexity,they exhibit limitations in capturing the intricate temporal dependencies present in atmospheric $CO2$ data. Ensemble techniques, includ- ing Random Forest and Gradient Boosting, achieve enhanced predictive accuracy through model aggregation. In particular, deep learning approaches, particularly LSTM networks, consistently outperform other methods by effectively modeling long- term sequential dependencies. The findings underscore the im- portance of leveraging advanced machine learning architectures for environmental time series forecasting and suggest that deep learning holds significant promise for future climate modeling efforts.**

***IndexTerms*—CarbonDioxide$CO2$,ClimateChange,Machine Learning, Time Series Forecasting, Long Short-Term Memory (LSTM), Random Forest, Gradient Boosting, ARIMA, Atmo- spheric Science, Mauna Loa Observatory**

I. INTRODUCTION

Anthropogenic Emissions Of Carbon Dioxide(CO_2)have led to an unprecedented increase in global atmospheric CO_2 levels. This greenhouse gas is a primary contributor to global warming and climate instability. The Accurate prediction of atmospheric CO_2 concentrations is critical not only for envi- ronmental research but also for shaping international climate policies.

Traditional statistical methods such as Auto-regressiveIn- tegrated Moving Average(ARIMA) have long been employed for forecasting CO_2 levels. However, recent advancements in machine learning (ML) have introduced a wide range of powerful algorithms capable of modeling complex, nonlinear, and temporal relationships in environmental datasets[1],[3].

These models offer enhanced flexibility and predictive capabilities compared to conventional techniques. In this study, we focus on evaluating and comparing the performance of eight different models for predicting CO_2 concentrations using long-term historical data from Mauna Loa Observatory. The models include linear regression, deci- sion tree regression, random forest, support vector regression (SVR), gradient boosting, XGBoost, ARIMA, and long-short- term memory (LSTM) networks. Our objective is to identify which of these models are best suited for long-term forecasting of atmospheric CO_2, while also exploring the trade-offs between accuracy,interpretability,and computational

International Journal of Engineering, Management and Humanities (IJEMH)
Volume 6, Issue 2, Mar.-Apr, 2025 pp: 344-350 ISSN: 2584-2145 www.ijemh.com

complexity [2], [4], [5].
This comparative analysis aims to contribute to the growing body of work exploring AI-driven approaches for climate modeling, emissions analysis, and sustainable environmental planning.providing insights into the frequency of various concen- tration ranges.

II. DATASET

The Mauna Loa CO_2 dataset is one of the most widely recognized sources of atmospheric CO_2 concentration data. Collected by the Mauna Loa Observatory in Hawaii, this dataset provided daily measurements of CO2 levels dating back to 1958. The data is characterized by its long temporal span, which offers a unique opportunity to study both seasonal and long-term trends in CO2 concentrations.
The Primary feature in the dataset is the Daily Average CO_2 concentration, measured in parts per million (ppm). This variable serves as the target for the machine learning models used in this study.The dataset also includes the corresponding measurement date, allowing for temporal analysis and the creation of time series models. In total,the dataset spans over six decades, providing an extensive sample of CO_2 concentration trends.
Topreparethedatasetformodeling,we performed several preprocessing steps.Missing Values Were Handled by forward filling, and outliers were identified and removed based on statistical thresholds. The dataset was then divided into training and testing sets, with 80% of the data used for training and the remaining 20% reserved for testing.

III. DATA VISUALIZATIONS

Understanding the underlying patterns and trends in the data is a crucial step in a machine learning project.The Following visualizations were created to explore the CO_2 concentrations over time and gain a better understanding of the dataset:

Basic Trend-Line: A simple line graph illustrating the overall trend in CO_2 concentrations over time. This visualization highlights the long-term upward trend in CO2 levels.

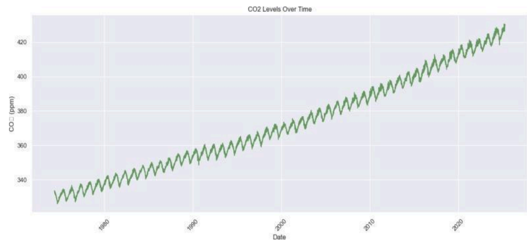

Fig.1.Basic Trend-Line of CO_2 Concentrations Over Time

Histogram of CO_2 Levels: A histogram that shows the distribution of CO_2 concentration values in dataset,

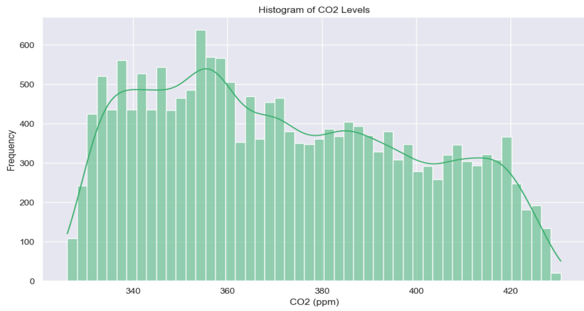

Fig.2.Histogram Showing CO_2 Level Distribution

Average CO_2 per Day: A plot displaying the average CO_2 concentration for each day in the dataset, which helps identify daily variations and patterns.

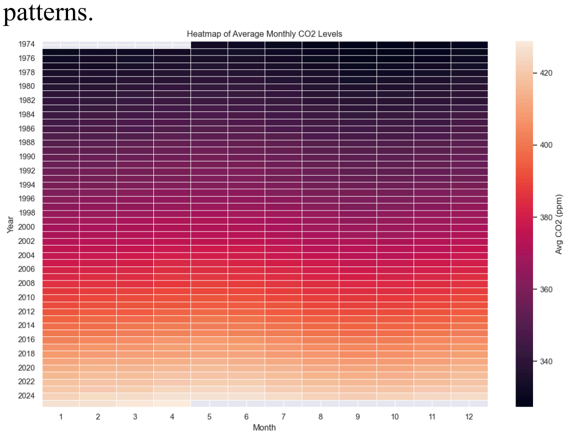

Fig.3.Average Daily CO2 Concentration

CO_2 Safety Thresholds: A graphical representation of the safety thresholds for CO_2 concentrations, based on environmental guidelines,showing the levels above which CO2 concentrations are considered hazardous.

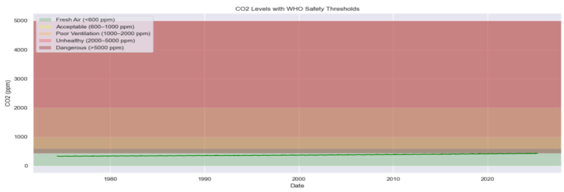

Fig.4.CO2 Safety Threshold Levels

Min and Max CO2 Levels in the Month: A plot highlighting the minimum and maximum CO2 concentrations observed during each month, providing insight into the seasonal variation in CO2 levels.

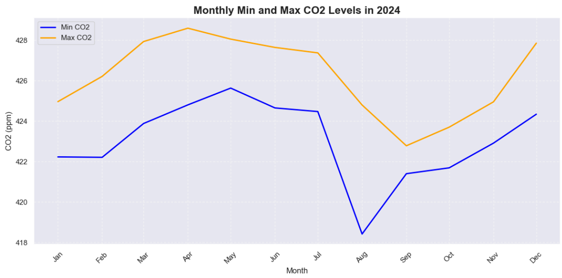

Fig.5.Monthly Min and Max CO2 Concentrations

Min and Max CO2 Levels in Year: A graph that shows the minimum and maximum CO2 concentrations observed for each year, allowing for an analysis of interannual variations.

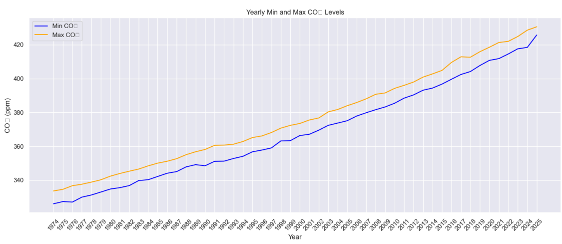

Fig.6.Yearly Min and Max CO2 Concentrations

These visualizations not only helped in exploring the data but also served as important diagnostic tools for assessing the behavior of different machine learning models.

IV. EVALUATION METRICS

Evaluating the performance of predictive models is essential to ensure their reliability and applicability in real-world scenarios. In this study, two widely used regression error metrics, Root Mean Squared Error (RMSE) and Mean Absolute Error (MAE), are used to assess the precision of machine learning models to forecast atmospheric CO2 concentrations. These metrics provide complementary insights into model performance by measuring the deviation between actual and predicted values.

A. Root Mean Squared Error (RMSE)

Root Mean Squared Error (RMSE) is a commonly used metric in regression analysis that quantifies the standard deviation of residuals (prediction errors). It is defined mathematically as:

$$\mathrm{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2} \tag{1}$$

Where:
• yi refers to the actual observed value at time i,
• ˆyi is the predicted value at time i,
• n is the number of data points.

RMSE provides a measure of how well the model predictions match the actual values. Because errors are squared before averaging, RMSE assigns a higher penalty to larger deviations than to smaller ones. This makes RMSE particularly useful in scenarios where larger errors are more detrimental than smaller ones. A lower RMSE value indicates that the model makes predictions with minimal error, which means that the predicted CO2 values align closely with the observed data.

However, one limitation of RMSE is its sensitivity to outliers. Since squared differences amplify large errors, some extreme predictions can disproportionately impact the RMSE value, potentially misleading the interpretation of model performance.

B. Mean Absolute Error (MAE)

Mean Absolute Error (MAE) is another fundamental metric that evaluates the average magnitude of errors in a set of predictions, without considering their direction. It is computed as:

$$\mathrm{MAE} = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i| \tag{2}$$

Where:
• yi refers to the actual observed value at time i,
• ˆyi is the predicted value at time i,
• n is the number of data points.

Unlike RMSE, MAE treats all individual differences between predicted and actual values equally by computing the absolute value of each error before averaging. This makes MAE more robust to outliers compared to RMSE, as it does not disproportionately emphasize large errors. Instead, it provides a straightforward interpretation of model accuracy by expressing the average error in the same units as the original data.

MAE is particularly useful when the goal is to measure how far, on average, predictions deviate from actual values, making it an intuitive metric for evaluating regression models.
However, unlike RMSE, it does not heavily penalize large errors, which can sometimes lead to underestimating the significance of substantial deviations in prediction.

V. RESULTS

The performance of the various machine learning models was evaluated based on two primary error metrics: The root mean square error (RMSE) and the mean absolute error (MAE), calculated as shown in Equation (1) and Equation (2), respectively. These metrics offer a quantitative understanding of how accurately each model predicts the atmospheric CO_2 levels compared to the actual observed values. The lower the values of these metrics, the better the model's performance in terms of error minimization.

Table I presents the RMSE and MAE scores obtained for each of the considered models. It is evident from the results that the LSTM model achieved the lowest RMSE and MAE values among all, suggesting its superior ability to capture temporal dependencies and non-linear trends in the CO_2 time-series data.

TABLE I

COMPARISON OF MODELS BASED ON RMSE AND MAE

MODEL	RMSE	MAE
Random Forest	3.2161	3.0381
Holt-Winters	1.1079	0.9234
XGBoost	3.4682	3.2777
Facebook Prophet	1.2008	0.9790
ARIMA	1.5765	1.1883
Theta	1.1662	0.8517
LSTM	**0.8750**	**0.6739**
Hybrid(LSTM + ARIMA)	1.5734	1.1817

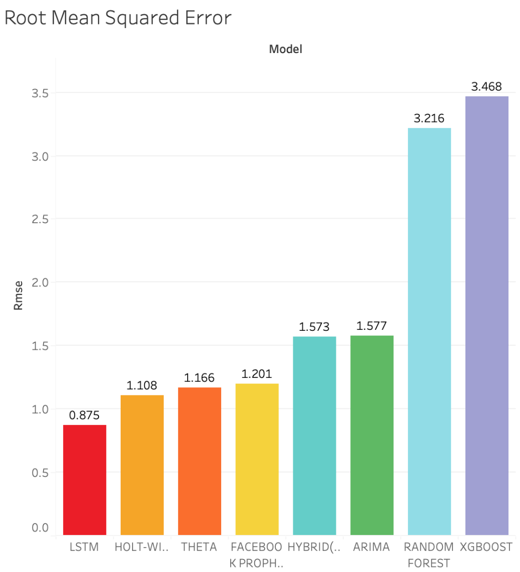

Fig.7.RMSE Comparison Across Different Models

To facilitate a more intuitive understanding, Figures 7 and 8 visually depict the comparison of RMSE and MAE scores using bar charts. These visualizations further emphasize the relative performance differences among the models. While traditional models like Linear Regression exhibited higher error values, ensemble methods such as XGBoost and Random Forest showed significant improvements. However, the LSTM model, being a deep learning-based approach capable of learning complex temporal structures, consistently outperformed the rest. These results underline the importance of selecting an appropriate model based on the complexity of the dataset and the desired forecasting accuracy. While simpler models may offer faster training and interpretability, advanced methods like LSTM provide the best predictive performance, especially for time-dependent environmental datasets like atmospheric CO2 levels.

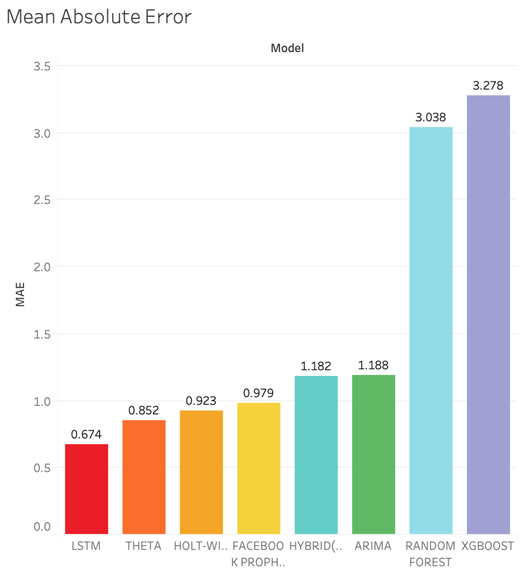

Fig.8.MAE Comparison Across Different Models

VI. LIMITATIONS, CHALLENGES AND MODELING COMPLEXITIES

Despite the promising outcomes demonstrated by the evaluated models, several limitations and challenges were encountered throughout the study, which merit critical consideration:

• **Data Preprocessing Constraints:** The Mauna Loa dataset, while comprehensive, required imputation of missing values via interpolation techniques. This preprocessing step, although necessary, may have introduced subtle biases, particularly affecting the temporal consistency of classical models such as ARIMA and Holt-Winters, which rely heavily on uninterrupted sequential data.

• **Exclusion of Exogenous Variables:** The present study solely focuses on univariate time series modeling, thereby omitting potentially influential exogenous factors such as industrial output, seasonal biomass fluctuations, and global policy interventions. The absence of these variables constrains the models' capacity to account for sudden variations or structural shifts in CO2 levels.

Model Interpretability vs. Predictive Power: While deep learning models such as LSTM demonstrated superior performance, their "black-box" nature posed interpretability challenges. In contrast, models such as Random Forest and XGBoost offered more transparency but lacked the temporal modeling finesse necessary for long-term forecasting.

• **Complexity in Hybridization:** Although the hybrid ARIMA–LSTM architecture was intended to leverage both linear and nonlinear pattern recognition, integration complexity—especially with regard to synchronization of residuals and phase alignment—diminished the anticipated performance gains in some instances.

• **Computational Overhead:** Deep learning models and ensemble methods required substantial resources and training time, which may limit scalability and deployment feasibility in resource-constrained or real-time environments.

• **Hybrid Modeling Approach:** Hybrid models combine the strengths of different algorithms to improve predictive performance. In this study, we developed a hybrid ARIMA–LSTM model, where ARIMA captured the linear trends and seasonality in the CO2 time series, and the LSTM modeled the nonlinear residuals, computed as the difference between ARIMA predictions and actual values. While theoretically promising, practical implementation posed challenges, including synchronization issues between residuals and LSTM inputs, and difficulties in jointly tuning both models, sometimes causing overfitting or underfitting. These complexities occasionally reduced the expected performance gains. Future hybrid approaches should focus on advanced ensemble techniques or end-to-end optimization to better integrate linear and nonlinear modeling.

VII. FUTURE MODEL IMPROVEMENTS

To enhance the robustness, scalability, and predictive accuracy of atmospheric CO2 forecasting models, the following directions are proposed for future research:

• **Incorporation of Multivariate Data:** Future work should explore multivariate modeling frameworks that incorporate external variables such as temperature, fossil fuel consumption rates, and socio-economic indicators. This would provide a more holistic modeling approach and improve predictive granularity.

• **Integration of Transformer Architectures:** Emerging sequence modeling paradigms such as Transformer-based architectures (e.g., Informer, TimeGPT) should be investigated for their ability to capture long-range dependencies with reduced training complexity compared to recurrent neural networks.

• **Multivariate Modeling Considerations:** While the current study focused on univariate modeling using historical CO2 concentration values, atmospheric CO2 levels are influenced by various external factors, such as global temperature anomalies, fossil fuel consumption rates, industrial activity indices, and seasonal vegetation cycles. To provide a richer context for prediction models, future research should explore the development of multivariate time series forecasting frameworks, incorporating exogenous variables. This could involve utilizing multivariate LSTM models, Vector AutoRegression (VAR) models, or Transformer-based sequence models tailored for multivariate environmental data. Integrating these additional features may enhance the model's ability to detect causal relationships, sudden structural changes, and nonlinear interactions, ultimately improving forecasting accuracy and robustness.

• **Explainable AI (XAI) Frameworks:** As model transparency becomes increasingly critical for policy applications, the adoption of XAI techniques such as SHAP (SHapley Additive exPlanations) and LIME (Local Interpretable Model-agnostic Explanations) can aid in elucidating the decision-making processes of complex models.

• **Transfer Learning and Domain Adaptation:** Leveraging pre-trained models on analogous environmental datasets followed by fine-tuning on CO2-specific data could enhance model generalization and reduce dependency on extensive training cycles.

• **Real-time and Online Learning:** The development of online learning algorithms capable of dynamic model updating in response to streaming data would be instrumental for deployment in live monitoring systems and policy response platforms.

VII. CONCLUSION

This research undertook a comprehensive comparative evaluation of eight machine learning and statistical models for the task of atmospheric CO2 concentration forecasting using the Mauna Loa Observatory dataset. The findings underscore the efficacy of advanced deep learning architectures, particularly Long Short-Term Memory (LSTM) networks, in capturing complex temporal dynamics and non-linear trends inherent in long-term environmental datasets.

While ensemble learning techniques such as Random Forest and XGBoost provided competitive performance with enhanced interpretability, they were outperformed by LSTM in terms of both RMSE and MAE metrics, as evaluated using Equation (1) and Equation (2). Classical time series models, including ARIMA and Holt-Winters, offered foundational baselines but were limited in capturing nonstationary trends.

The study highlights the trade-offs between Interpretability, computational efficiency, and predictive performance. It establishes a foundation for future enhancements through multivariate modeling, hybrid deep learning frameworks, and the integration of explainable AI techniques. As the urgency for climate monitoring intensifies, such data-driven approaches will play an increasingly pivotal role in environmental intelligence, policy formulation, and sustainable development.

ACKNOWLEDGMENT

We would like to express our sincere gratitude to Mr. Rajnikant Shinde, SAFAR - India, Indian Institute of Tropical Meteorology, Pune, for kindly providing us with the Mauna Loa dataset used in this research. His valuable assistance played a crucial role in the success of this study. For more information, please visit: https://safar.tropmet.res.in.

REFERENCES

[1] X. Li, "A comparative study of statistical and machine learning models on near-real-time daily CO2 emissions prediction," School of Statistics, Capital University of Economics and Business, Beijing, P.R. China, 2022.

[2] W. Yu, L. Xia, and Q. Cao, "A machine learning algorithm to explore the drivers of carbon emissions in Chinese cities," Scientific Reports, vol. 14, no. 23609, 2024, doi: 10.1038/s41598-024-75753-y.

[3] Y. Ma, H. Liu, and S. Wang, "Nonparametric approaches for analyzing carbon emission: from statistical and machine learning perspectives," arXiv preprint arXiv:2303.14900, 2023.

[4] L. Wu, K. Li, Y. Huang, Z. Wan, and J. Tan, "Optimization of carbon footprint management model of electric power enterprises based on artificial intelligence," PLOS ONE, vol. 20, no. 1, p. e0316537, Jan.
2025, doi: 10.1371/journal.pone.0316537.

[5] Y. Meng and H. Noman, "Predicting CO2 emission footprint using AI through machine learning," Atmosphere, vol. 13, no. 11, p. 1871, Nov. 2022, doi: 10.3390/atmos13111871.