NATURENGS
E-ISSN 2717-8013
https://dergipark.org.tr/en/pub/naturengs

Forecasting CO_2 Emissions with Machine Learning Methods: Türkiye Example and Future Trends

İbrahim AYAZ

Bitlis Eren University, Department of Computer Technologies, Bitlis Türkiye – 13200

ARTICLE INFO

Received 02.12.2024
Accepted 25.12.2024

Doi: 10.46572/naturengs.1595329

ABSTRACT

Climate change is a critical problem that causes global environmental and social issues due to increased greenhouse gas emissions caused by human activities. Carbon dioxide (CO_2) emissions, in particular, are one of the main elements of global warming and have devastating effects on ecosystems. Keeping track of carbon dioxide emissions resulting from human activities like burning fossil fuels, clearing forests, and farming, as well as forecasting their future patterns, is essential for creating effective sustainable environmental strategies. The study utilized machine-learning models to evaluate CO_2 emissions per individual, using the dataset from the Global Carbon Atlas that has released in 2023.

In the study, the traditional ARIMA model and deep learning-based LSTM networks were comparatively discussed. The models were trained with the aim of predicting Türkiye's future CO_2 emission levels by learning from past data, and their performances were evaluated with MAE, MSE, RMSE, and R^2 metrics. The LSTM model achieved an R^2 score of 90.4%, while the ARIMA model achieved an R^2 score of 94.3%. The findings show that machine learning techniques are a powerful tool in the fight against climate change and provide valuable insights for policymakers. The findings of the study guide more effective monitoring of CO_2 emissions and determination of strategies for sustainable development goals.

Keywords: Carbon Emission, Machine Learning, Time Series, Forecasting

1. Introduction

An international concern, climate change is defined by the rise in greenhouse gases in the atmosphere brought on by human activity, which has serious negative effects on the environment and society. Fossil fuel combustion, deforestation, and agricultural practices are a few examples of human-caused activities that contribute to the atmospheric buildup of greenhouse gases, primarily carbon dioxide (CO_2). The primary contributors to global warming and the long-term alterations in the climate system are these gases [1], [2]. CO_2 emissions create effects such as increasing temperature on the earth's surface, rising sea levels, extreme weather changes, and devastating consequences on ecosystems [3]. One of the regions where water stress and drought are most intense is the Middle East. This problem directly affects agricultural areas. CO_2 emission rates in Middle Eastern countries are expected to increase by 13.28% by 2026 [4]. According to Global Carbon Atlas data, the country with the highest average CO_2 emission per capita is Sint Maarten (Dutch part) [5]. Countries with more than 10 tonnes of CO_2 emissions per capita in the Global Carbon Atlas dataset are shown in Figure 1.

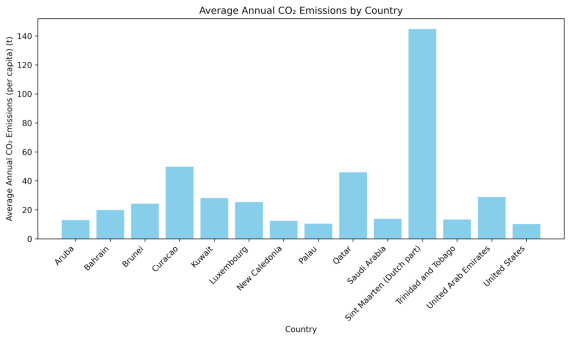

Figure 1. Average CO_2 Amount of Countries

Figure 1 shows the average tCO_2 emissions per capita for Sint Maarten (Dutch part) between 1926-2023, although it varies by country.

Carbon dioxide emissions, a significant contributor to climate change, have far-reaching consequences for Türkiye's environmental and economic systems. The rise in temperatures and alteration of rainfall patterns are intensifying the scarcity of water in regions such as

* Corresponding author. e-mail address: iayaz@beu.edu.tr
ORCID : 0000-0003-3519-1882

Southeast Anatolia, where agriculture is significantly reliant on irrigation [6]. The aforementioned water stressors result in diminished agricultural productivity and elevated water management costs, directly impacting food security and rural livelihoods [7]. Furthermore, urban areas are confronted with heightened risks of heat waves and flooding due to extreme weather events, necessitating costly adaptations to infrastructure [8]. These challenges underscore the imperative for targeted climate policies and emission reduction strategies tailored to Türkiye's distinctive socio-economic and environmental context.

In this context, monitoring CO_2 emissions and predicting their future trends are critical in combating climate change. Estimating emissions is vital for both environmental policies and economic planning. With conventional modeling techniques, it might be difficult to forecast complicated environmental variables, such as CO2 emissions. Machine learning approaches, on the other hand, have become a powerful tool for analyzing and forecasting time series data lately [9], [10].

The application of machine learning techniques has constituted a significant area of investigation within the domain of estimation and analysis of CO_2 emissions.

In particular, machine learning methods have been employed for the estimation of greenhouse gas emissions in Türkiye. Papucçu and Bayramoğlu (2016) employed an artificial neural networks method to estimate CO_2 emissions, utilizing energy production and consumption data, as well as the amount of energy used for industrial production and transportation [11]. Garip and Oktay (2018) employed random forest and support vector machine methods to estimate Türkiye's CO_2 emissions [12]. Pence et al. (2023) estimated the energy production and emissions from animal manure using machine learning methods such as support vector machines (SVM), multi-layer perceptron's (MLP), and linear regression (LR) [13].

Machine learning is a set of algorithms that can perform certain tasks by learning from data, and it offers models that can especially cope with large data sets and capture complex patterns. Autoregressive Integrated Moving Average (ARIMA) and Long Short-Term Memory (LSTM) networks are the two main machine-learning techniques for estimating CO2 emissions [14-17]. These models can predict future emission levels by analyzing past emission data and thus contribute to the shaping of environmental policies.

The Global Carbon Atlas open-access data set published in 2023 was used in the study. The selection of the Global Carbon Atlas dataset is based on a number of criteria that serve to highlight its importance and accuracy for the analysis of CO_2 emissions in Türkiye. Firstly, the dataset provides comprehensive and open-access data on per capita CO_2 emissions for countries worldwide, with annual updates that guarantee temporal reliability and relevance [5]. In comparison to other sources, such as those provided by the World Bank or the European Environment Agency, the Global Carbon Atlas dataset is specifically focused on carbon emissions, offering a more detailed perspective that facilitates comprehensive modeling and forecasting. Furthermore, its structured format is highly compatible with machine learning frameworks, including LSTM and ARIMA. This compatibility ensures greater accuracy and reliability in the results of the study, particularly in forecasting Türkiye's future CO_2 emission trends [18].

LSTM and ARIMA models were trained with tCO_2 (ton-carbon dioxide) emission data per capita. When the models were evaluated with performance metrics, LSTM and ARIMA models achieved success with 90.4% and 94.3% R^2 scores, respectively. The results obtained are promising in estimating Türkiye's CO_2 emission amounts in the coming years.

In the introduction section of the study, information about CO_2 emission amounts and literature research is given. In the second section, the data set used in the study, the methods used, and performance metrics are explained. In the third section, the application results obtained are evaluated and discussed. Finally, the study was completed with the conclusion section.

2. Materials and Methods

The dataset used in the study is given in section 2.1, and the literature review of LSTM, SVM, and RF models is given in sections 2.2, 2.3, and 2.4, respectively. The performance metrics used in the models are given in section 2.5. In addition, a 4x3090 RTX graphics processor and a GPU server with 120 GB memory were used in the application phase of machine learning methods.

2.1. Dataset

In this study, the Global Carbon Atlas open-access dataset published in 2024 was used [5]. The dataset provides annual CO_2 emission amounts of countries and per capita emission amounts (tons). The data used in the study includes annual CO_2 emission amounts of Türkiye, and estimates were made to make future emission estimates. The data for Türkiye covers the period 1865-2023.

The utilization of data from this historical range, comprising 158 years of data, enabled the development of artificial intelligence models with the objective of enhancing the precision of CO_2 emission estimates, thereby facilitating the formulation of efficacious climate policies. Figure 2 illustrates the per capita tCO_2 (tonnes of carbon dioxide) in Türkiye between 1865 and 2023.

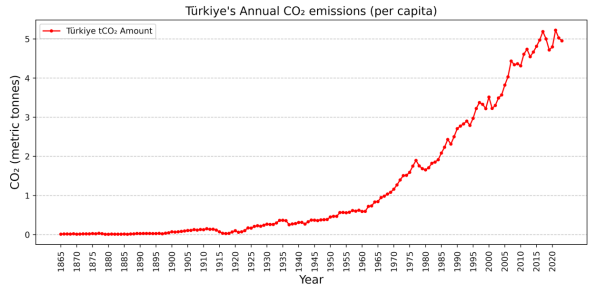

Figure 2. Türkiye's CO_2 emission levels over the years

As seen in Figure 2, there is a gradual increase in the amount of CO_2 emissions per capita in Türkiye from 1865 to 2023.

2.2. LSTM

LSTM is a recurrent neural network (RNN) model. This model works very effectively on time series and sequential data. Unlike traditional RNN networks, effective results can be obtained in learning data with long-term dependencies [19], [20], [21]. The architecture of the RNN network is shown in Figure 3.

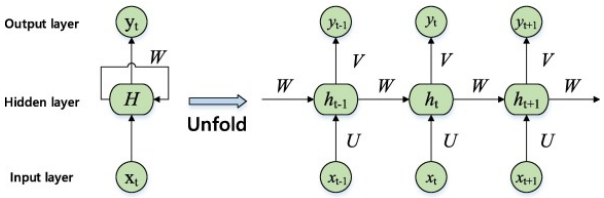

Figure 3. The architecture of the RNN network [20].

Its internal loop structure, as shown in Figure 3, demonstrates dynamic behavior by creating a temporal link between the most recent and earlier states. The input at time and the output at time define the output at time in the figure. Equation 1 and Equation 2 represent the RNN computation procedure:

$$h_t = tanh(Ux_t + Wh_{t-1} + b_h) \quad (1)$$

$$y_t = softmax(Vh_t + b_y) \quad (2)$$

In the case of x_t, h_t, and y_t, the input vector, hidden cell state, and output at time t are denoted, respectively. W, U, and V stand for the respective RNN layer's weight matrices. These parameters are bias vectors, $b_{h,}$ and b_y [20].

2.3. ARIMA

ARIMA is a potent technique for forecasting and modeling time series data. This model works based on past values (autoregressive), differenced states (integrated), and moving average components of the data. ARIMA models developed by Box and Jenkins (1976) [22] provide a systematic approach that allows both stationarization and forecasting of time series data. However, studies by Zhang [23] indicated that the linear assumptions of ARIMA models may limit the performance in time series with more complex patterns and proposed hybrid models. ARIMA is widely used in many areas such as finance, climate forecasting, and emission forecasting.

2.4. Performance Evaluation Metrics

The trained models' performance in forecasting CO_2 emission regression is measured utilizing a variety of performance assessment criteria. R Square Error (R^2), Mean Squared Error (MSE), Mean Absolute Error (MAE), and Root Mean Squared Error ($RMSE$) performance evaluation metrics are used in the study. The following equations provide the metrics' mathematical expressions. Within the metrics, y represents the actual value, $\bar{y}$ represents the average value of y, and $\hat{y}$ represents the value that was forecasted [24].

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 \quad (3)$$

$$RMSE = \sqrt{\sum_{i=1}^{n}\frac{(\hat{y}_i - y_i)^2}{n}} \quad (4)$$

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i| \quad (5)$$

The convergence of MSE, $RMSE$, and MAE metrics to zero indicates that the error rate of the model is low.

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(\hat{y}_i - y_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y}_i)^2} \quad (6)$$

The convergence of the expression to one indicates that the error rate of the model is low. These metrics are frequently employed in regression analysis to evaluate the precision and calibration of predictive models.

The MAE is a measure of the average magnitude of errors in a set of predictions, without consideration of their direction. It offers a straightforward means of understanding the accuracy of predictions.

The MSE is a statistical measure that calculates the average of the squared differences between the predicted and actual values. It assigns greater weight to larger errors. This sensitivity to outliers can be advantageous when large errors are particularly undesirable. The $RMSE$ is the square root of the MSE, which returns the error metric to the original units of the target variable, thus facilitating interpretation. The coefficient of determination R^2 demonstrates the proportion of the variance in the dependent variable that can be predicted from the independent variables, thereby providing insight into the explanatory power of the model. Collectively, these metrics offer a comprehensive view of model performance, addressing different aspects such as error size, sensitivity to outliers, and explanatory power.

3. Result and Discussion

In this study, the dataset was partitioned into training and test data at 70%-30%, 75%-25%, and 80%-20%, respectively, with the objective of predicting future annual per capita CO_2 emissions. The highest level of success was achieved with a training and testing ratio of 80:20. Two models, LSTM and ARIMA, were employed to predict the emission data. The results of the model predictions were visualized. The performance of each model was evaluated by comparing the RMSE, MAE, MSE, and R^2 performance metrics.

The hyperparameters of the LSTM model exert a direct influence on the model's performance. The time step determines the extent to which the model evaluates historical data from a previous point in time. This value is selected in accordance with the characteristics of the dataset and the learning capacity of the model. The number of epochs indicates the number of times the model passes over the training data; the number of epochs that usually yields optimal results in the validation set is determined. The choice of optimizer affects the learning speed and overall performance of the model; adaptive optimization algorithms such as Adam usually yield faster and more stable results. The choice of LSTM hyperparameters was determined by an empirical method. Table 1 shows the hyperparameters of the LSTM model.

Table 1. LSTM Hyper Parameters..

Parameters Name	Value
Time-Step	10
Epoch	200
Optimizer	Adam
Learning Rate	0.001
Loss	MSE
Batch Size	64

The LSTM model was trained with the hyperparameters in Table 1. The training and test prediction performance, as well as the CO_2 emission estimates between 2014 and 2040, are shown in Figure 4.

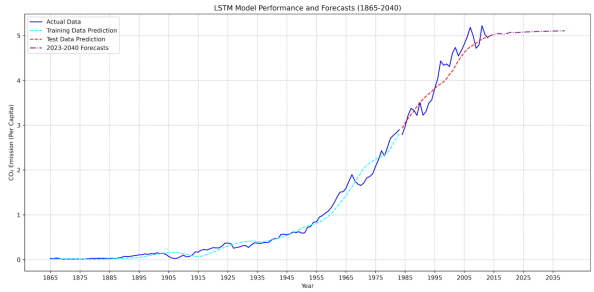

Figure 4. Estimation of the LSTM model on real data

The forecasts obtained from the LSTM model on the training and test data are rather consistent, as shown in Figure 4. Based on past data, the model correctly forecasted future CO2 emissions for 2024–2040.

The hyper-parameters of the ARIMA model, namely p (number of autoregressive terms), d (number of differences required to make the time series stationary), and q (number of moving average terms), are empirically determined as 5, 1, and 0, respectively.

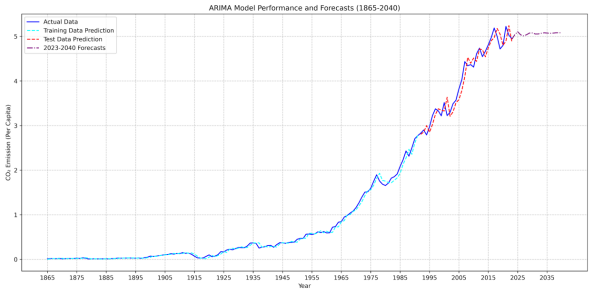

Figure 5. Estimation of the ARIMA model on real data.

Figure 5 displays the ARIMA model's performance graph. As the graphic illustrates, the ARIMA model is shown to fit actual data more accurately than the LSTM model. This evaluation means that the ARIMA model is more successful for the current problem.

Table 2. LSTM and ARIMA performance metric results.

Performance Metrics and Models	MAE		MSE		RMSE		R2	
	Train	Test	Train	Test	Train	Test	Train	Test
LSTM	7.36%	18.98%	1.11%	5.25%	10.56%	22.92%	98.07%	90.43%
ARIMA	**3.15%**	**15.67%**	**0.25%**	**3.48%**	**5.05%**	**18.66%**	**99.46%**	**94.3%**

The LSTM and ARIMA models' training and testing performance outcomes are displayed in Table 2.

The findings demonstrate that the ARIMA model exhibits the most optimal performance and effectively approximates the annual CO_2 emissions estimation. The ARIMA model yielded more precise forecasts than the other models, exhibiting lower error metrics and a high R^2 score. By the year 2040, the estimated amount of tCO_2 emissions (tonnes) per capita is 5.077. The 18-year average forecast of the ARIMA model predicts a figure of 5.056 tCO_2; this is a more accurate estimation than that provided by other models. The effectiveness of the ARIMA model is due to the linear ranges and seasonality of CO_2 emission data, which are better represented by this model. The superior performance of the ARIMA model with limited data is a result of its ability to handle larger datasets and more programming resources, which are required by LSTM models.

It is of paramount importance to be able to make accurate forecasts of CO_2 emissions if we are to develop effective environmental policies. Policymakers can set realistic and achievable emission reduction targets using ARIMA and LSTM models. These predictive models assist in the assessment of the potential impacts of proposed policies on future emissions, thereby facilitating informed decision-making. Furthermore, an understanding of future emission trends allows for more effective planning and resource allocation to mitigate

climate change. For instance, a study on forecasting CO_2 emissions in India using ARIMA demonstrated the model's efficacy in providing evidence-based information to assist in the implementation of sustainable climate policies.

The utilization of emission forecasts by industries enables the optimization of operational processes through the prediction of emission trends and the subsequent adjustment of production procedures to minimize environmental impact. Furthermore, predictive models assist in ensuring compliance with environmental regulations by forecasting future emission levels. Furthermore, precise forecasts facilitate the formulation of long-term sustainability strategies and investments. A study on generation decarbonization utilizing ARIMA-LSTM models emphasized their significance in industrial decision-making, exemplifying their deployment in forecasting coal-generated electricity and carbon dioxide emissions.

In conclusion, integrating ARIMA and LSTM model forecasts into decision-making processes offers valuable insights for policymakers and industries alike, enabling the implementation of proactive measures to address climate change and promote environmental sustainability.

4. Conclusions

In this study, LSTM and ARIMA models were trained to estimate CO_2 emissions in the coming years (2024-2040) using Global Carbon Atlas Türkiye data. The models' ability to effectively estimate CO_2 emissions was assessed using success indicators such as RMSE, MAE, MSE, and R^2. In addition, the amount of CO_2 emissions in the coming years (2024-2040) was estimated.

The LSTM model showed good performance in general with MAE, MSE, RMSE, and R^2 test performance metric values of 18.98%, 5.25%, 22.92%, and 90.43%, respectively. The ARIMA model stood out by exhibiting better performance than the LSTM model with MAE, MSE, RMSE, and R^2 test performance metric values of 15.67%, 3.48%, 18.66%, and 94.3%, respectively.

In the present study, we employ the ARIMA and LSTM models to forecast CO_2 emissions up to 2040. To evaluate the precision of these forecasts, we initially establish a baseline by examining the performance of our model on historical data. To assess the accuracy of future forecasts, we intend to utilize recent emissions data from official sources, including Türkiye's Informative Inventory Report (IIR). These data will enable us to assess the accuracy of our forecasts by providing detailed information on emission sources and quantities [25].

Future studies could investigate increasing the prediction accuracy by incorporating hybrid models or additional features into the model.

References

[1] **Intergovernmental Panel on Climate Change (IPCC)**, *Climate Change 2013 – The Physical Science Basis: Working Group I Contribution to the Fifth Assessment Report of the Intergovernmental Panel on Climate Change*. Cambridge: Cambridge University Press, 2014. doi: 10.1017/CBO9781107415324.

[2] **V. Masson-Delmotte *et al.***, Eds., *Climate Change 2021: The Physical Science Basis. Contribution of Working Group I to the Sixth Assessment Report of the Intergovernmental Panel on Climate Change*. Cambridge, United Kingdom and New York, NY, USA: Cambridge University Press, 2021. doi: 10.1017/9781009157896.

[3] **N. R. Council, D. on E. and L. Studies, B. on A. S. and Climate, and A. C. C. P. on A. the S. of C. Change**, *Advancing the Science of Climate Change*. National Academies Press, 2011.

[4] **N. Rajabi Kouyakhi**, "CO2 emissions in the Middle East: Decoupling and decomposition analysis of carbon emissions, and projection of its future trajectory," *Sci. Total Environ.*, vol. 845, p. 157182, Nov. 2022, doi: 10.1016/j.scitotenv.2022.157182.

[5] **P. Friedlingstein *et al.***, "Global Carbon Budget 2024," *Earth Syst. Sci. Data Discuss.*, pp. 1–133, Nov. 2024, doi: 10.5194/essd-2024-519.

[6] **H.-O. Pörtner *et al.***, Eds., *Climate Change 2022: Impacts, Adaptation and Vulnerability. Contribution of Working Group II to the Sixth Assessment Report of the Intergovernmental Panel on Climate Change.* 2022.

[7] **D. L. Koç, B. Kapur, M. Ünlü, and R. Kanber**, "The Situation of Water Resources and Agricultural Irrigation in Turkey," *Çukurova Tarım Ve Gıda Bilim. Derg.*, vol. 37, no. 2, Art. no. 2, Dec. 2022, doi: 10.36846/CJAFS.2022.80.

[8] **J. Xu, D. Cai, and J. Zhu**, "Navigating the green wave: Urban climate adaptation and firms' investment decisions-evidence from China," *Energy Econ.*, vol. 141, p. 108087, Jan. 2025, doi: 10.1016/j.eneco.2024.108087.

[9] **N. V. Chawla**, "Data Mining for Imbalanced Datasets: An Overview," in *Data Mining and Knowledge Discovery Handbook*, O. Maimon and L. Rokach, Eds., Boston, MA: Springer US, 2009, pp. 875–886. doi: 10.1007/978-0-387-09823-4_45.

[10] **L. Breiman**, "Random Forests," *Mach. Learn.*, vol. 45, no. 1, pp. 5–32, Oct. 2001, doi: 10.1023/A:1010933404324.

[11] **H. Pabuçcu and T. Bayramoğlu**, "Yapay sinir ağlari ile CO2 emisyonu tahmini: Türkiye örneği," *Gazi Üniversitesi İktisadi Ve İdari Bilim. Fakültesi Derg.*, vol. 18, no. 3, pp. 762–778, 2016.

[12] **E. GARİP and A. B. OKTAY**, "Forecasting CO2 Emission with Machine Learning Methods," in *2018 International Conference on Artificial Intelligence and Data Processing (IDAP)*, Sep. 2018, pp. 1–4. doi: 10.1109/IDAP.2018.8620767.

[13] **I. Pence, K. Kumaş, M. C. Siseci, and A. Akyüz**, "Modeling of energy and emissions from animal manure using machine learning methods: the case of the Western Mediterranean Region, Turkey," *Environ. Sci. Pollut. Res.*, vol. 30, no. 9, pp. 22631–22652, Feb. 2023, doi: 10.1007/s11356-022-23780-5.

[14] **X. Li and X. Zhang**, "A comparative study of statistical and machine learning models on carbon dioxide emissions prediction of China," *Environ. Sci. Pollut. Res.*, vol. 30, no. 55, pp. 117485–117502, Nov. 2023, doi: 10.1007/s11356-023-30428-5.

[15] **Y. Jin, A. Sharifi, Z. Li, S. Chen, S. Zeng, and S. Zhao**, "Carbon emission prediction models: A review," *Sci. Total Environ.*, vol. 927, p. 172319, Jun. 2024, doi: 10.1016/j.scitotenv.2024.172319.

[16] **M. Yildirim, E. Cengil, Y. Eroglu, and A. Cinar**, "Detection and classification of glioma, meningioma, pituitary tumor, and normal in brain magnetic resonance imaging using deep learning-based hybrid model," *Iran J. Comput. Sci.*, vol. 6, no. 4, pp. 455–464, Dec. 2023, doi: 10.1007/s42044-023-00139-8.

[17] **M. Yildirim, H. Bingol, E. Cengil, S. Aslan, and M. Baykara**, "Automatic Classification of Particles in the Urine Sediment Test with the Developed Artificial Intelligence-Based Hybrid Model," *Diagnostics*, vol. 13, no. 7, Art. no. 7, Jan. 2023, doi: 10.3390/diagnostics13071299.

[18] **"Glossary | DataBank."** Accessed: Dec. 15, 2024. [Online]. Available: https://databank.worldbank.org/metadataglossary/world-development-indicators/series/EN.ATM.CO2E.PC

[19] **Z. Han, B. Cui, L. Xu, J. Wang, and Z. Guo**, "Coupling LSTM and CNN Neural Networks for Accurate Carbon Emission Prediction in 30 Chinese Provinces," *Sustainability*, vol. 15, no. 18, Art. no. 18, Jan. 2023, doi: 10.3390/su151813934.

[20] **X. Wang, W. Liu, Y. Wang, and G. Yang**, "A hybrid NOx emission prediction model based on CEEMDAN and AM-LSTM," *Fuel*, vol. 310, p. 122486, Feb. 2022, doi: 10.1016/j.fuel.2021.122486.

[21] **B. Kocaman and V. Tümen**, "Detection of electricity theft using data processing and LSTM method in distribution systems," *Sādhanā*, vol. 45, no. 1, p. 286, Dec. 2020, doi: 10.1007/s12046-020-01512-0.

[22] **G. Tunnicliffe Wilson**, "Time Series Analysis: Forecasting and Control,5th Edition, by George E. P. Box, Gwilym M. Jenkins, Gregory C. Reinsel and Greta M. Ljung, 2015. Published by John Wiley and Sons Inc., Hoboken, New Jersey, pp. 712. ISBN: 978-1-118-67502-1," *J. Time Ser. Anal.*, vol. 37, p. n/a-n/a, Mar. 2016, doi: 10.1111/jtsa.12194.

[23] **G. P. Zhang**, "Time series forecasting using a hybrid ARIMA and neural network model," *Neurocomputing*, vol. 50, pp. 159–175, Jan. 2003, doi: 10.1016/S0925-2312(01)00702-0.

[24] **S. Kumari and S. K. Singh**, "Machine learning-based time series models for effective CO2 emission prediction in India," *Environ. Sci. Pollut. Res.*, vol. 30, no. 55, pp. 116601–116616, Nov. 2023, doi: 10.1007/s11356-022-21723-8.

[25] **Çevre ve Şehircilik Bakanlığı, "Türkiye'nin Bilgilendirici Envanter Raporu (IIR) 2021."** Çevre ve Şehircilik Bakanlığı, 2021. Accessed: Dec. 15, 2024. [Online]. Available: https://webdosya.csb.gov.tr/db/cygm/menu/turkey-s-irr-2021_tr_20211101034946.pdf