Machine learning-based Time Series Models for Effective CO2 Emission prediction in India

Sunil Kumar Singh (sksingh@mgcub.ac.in)
MGCUB: Mahatma Gandhi Central University https://orcid.org/0000-0001-8954-6648
Surbhi Kumari
Mahatma Gandhi Central University

Research Article

Keywords: Time Series Forecasting, Linear Regression, Random Forest Regressor, Air Pollution, CO2 Emissions, Holt-Winter, LSTM

Posted Date: February 7th, 2022

DOI: https://doi.org/10.21203/rs.3.rs-1265771/v1

Machine learning-based Time Series Models for Effective CO_2 Emission prediction in India

Surbhi Kumari[1], Sunil Kumar Singh[2*]
[1,2]Dept. of Computer Science and Information Technology
Mahatma Gandhi Central University, Motihari, Bihar, India
[*]Email: sksingh@mgcub.ac.in; sunilsingh.jnu@gmail.com
*corresponding author

Abstract

As per the report of datacommons.org, CO_2 emission in India is 1.80 metric tons per capita, which is very harmful for living beings. So, this paper presents India's detrimental CO_2 emission effect with the prediction of CO_2 emission for the next ten years based on univariate time series data. We used three statistical models i.e., AutoRegressive Integrated Moving Average (ARIMA) model, Seasonal AutoRegressive Integrated Moving Average with eXogenous factors (SARIMAX) model, and Holt-Winter model, two machine learning models, i.e., Linear Regression and Random Forest model and one deep learning model, i.e. Long Short Term Memory (LSTM) model. The performance analysis shows that LSTM, SARIMAX, and Holt-Winter methods are accurate models and have the Least Mean Squared Error(MSE), Root Mean Square Error (RMSE), and Median Absolute Error (MedAE) for this kind of univariate data distribution.

Keyword – Time Series Forecasting, Linear Regression, Random Forest Regressor, Air Pollution, CO_2 Emissions, Holt-Winter, LSTM

1. Introduction

According to Ministry of Statistics and Programme Implementation, UN (World Population Prospects 2019) the current population of India is 1,400,517,328 as of January 2022 based on interpolation of the latest United Nations data which is the second-largest, just falling behind China and standing at second in the world. It would catch China and even surpass it shortly if it continues to grow at the same rate. With this, environmental consequences which may arise are many but CO_2 emission remains the topmost concern because of the problems which ensue due to its increased rate (Bonga & Chirowa, 2014). According to UN data, India's CO_2 emission rose faster than the world average of 0.7 %. Increased CO_2 will accentuate the world's food and water crisis and increase the incidences of natural disasters.

The increased flooding, landslide, cloud bursts, etc. are already evident and would further increase if we continue to go the same way. Forecasting CO_2 emissions are also one important key to creating awareness among the public on solving environmental problems(Abdullah & Pauzi, 2015). So, the proposed work has carefully examined the three statistical models, two machine learning models, and a deep learning model for time series CO_2 emission forecasting and also did performance analysis based on nine metrics to get the best performing model for the emission of CO_2 till 2030. By carefully examining the outcomes of different models, we found some models to be fairly viable forecasts. The data used is of the past 40 years, which plots the increase in CO_2 emission against time.

The forecast we have come up with will help to understand our pace of emission and further correction we need to do to keep the temperature rise under control. This would help future policy actions necessary to develop India's Nationally Determined Contributions Pledge which India has taken at Paris agreement. This paper has also focused on finding which time series models i.e. statistical models, machine learning models, and deep learning models, are best suited for this kind of CO_2 emissions data.

And it gives an idea for future papers which could use these models and refine the future forecast by taking into account other exogenous variables such as increasing population, technology advancement, and several other future technologies and policy actions that may positively or negatively affect the CO_2 emission. This paper has made a comparative analysis of different models and their accuracy in forecasting CO_2 emission, which would be helpful for future researchers to ascertain the forecasting models suitable for the purpose.

2. Related works

Many works have been done in time series data using machine learning, deep learning, and statistical model. Let us know about some of them. In the first paper (Masini et al., 2021), the author survey the most recent advances in supervised machine learning (ML) and high dimensional models by considering both linear and non-linear methods for time-series forecasting and considering ensemble and hybrid models by combining ingredients from different alternatives. They also apply time series forecasting in the economic and financial fields. In (Crespo Cuaresma et al., 2004), the author studies the forecasting abilities of a battery of univariate models and includes AutoRegressive (AR) models.

In (Elsworth & Güttel, 2020), the author proposed an RNN model with a dimension reducing symbolic representation to deal with the sensitivity of hyperparameters in any time series model and solves the limitation of other models, that is the initialization of random weights. Also, his model is faster without affecting the performance of forecasting. Also, in (Zuo et al., 2020) author collected the CO2 emission data from the different provinces of China and proposed a model named LSTM-STRIPAT, an integrated model to predict emission in 2020, and in (Amarpuri et al., 2019)the research is aimed at predicting the CO2 levels in the year 2020 to make the Government of India understand the challenges. A deep learning hybrid model of Convolution Neural Network and Long Short-Term Memory Network (CNN-LSTM) is used as a forecasting model.

A prediction based work was done in (R. Kumar et al., 2020), in which they have collected the data from Delhi and National Capital Region (NCE), India to predict the Air quality index. For the same, they have applied the machine learning models and measured the performances in terms of MAE, MSE, RMSE, and MAPE metrics. Similar work was done in (S. Kumar et al., 2020) to measure the PM2.5 pollutant particles in the Delhi atmosphere. In this work, Extra-Trees regression and AdaBoost based regression model is applied to predict PM2.5 concentrations effectively. They have also considered additional atmospheric and surface factors such as wind speed, atmospheric temperature, pressure, etc.

The work (Ahmed et al., 2010) applied various eight machine learning models on famous M3 time series competition data and compared them. The models' multilayer perceptron and the Gaussian process regression performed the best. In (Nyoni & Bonga, 2019) author used the Box- Jenkins ARIMA approach on time series data of CO_2 emission in India from 1960 to 2017, and based on forecast, they suggested five policy prescriptions to improve the environmental conditions.

ARIMA model has also been used in the analysis of air pollutants and to predict them based on historical data in (Gopu et al., 2021) and also said that this is an efficient way by which we can find out the values of the pollutants when exceeding the limits prescribed by the World Health Organization (WHO). SARIMA is an ARIMA capable of dealing with the seasonality of the dataset, and SARIMAX is SARIMA with X factor, which is nothing but exogenous factors that affect the data. The paper (Nontapa et al., 2020) analysed and forecasted using the SARIMA and SARIMAX model with decomposition method and compared them on evaluation metric MAPE where SARIMAX with decomposition model outperforms traditional ones. There is information regarding time series forecasting and limitation with the SARIMAX model in paper (Özmen, 2021)with the application of retail store data. In this work of CO2 emission, a Random Forest regressor has been used for forecasting. In the paper (Fang et al., 2020) author proposed an optimal random forest model for the accurate prediction of an infectious diarrhoea epidemic, considering meteorological factors. He compared his proposed model to ARIMA and ARIMAX, and the RF model outperforms with MAPE of approximately 20% to the ARIMA model with MAPE reaching 30%. In the paper (Lepore et al., 2017) author describes that with the introduction of the EU's new CO2 emissions regulation, ships' operators are required to implement systems that will monitor and report their CO2 emissions. These systems will help them make informed decisions regarding their operations.Due to the complexity of the navigation information that ships provide, there is no widely available standard method or solution that can be used in real environments. This paper presents an extensive analysis of the various regression techniques that can be used to exploit the navigation information that ships provide.The study is made based on the data collected by the Grimaldi Group's Ro-Pax cruise ship during its operation. It aims to identify possible methods and models that can be used to analyze the ship's CO2 emissions and develop a predictive model.

There is a project which aims to develop models and artificial neural networks to predict the energy consumption of office buildings in Chile and CO2 emissions(Pino-Mejías et al., 2017). Eight fundamental variables have been used to analyze the design parameters of commercial air-cooled cooling and heating systems. The results of the study show that the linear regression models with higher accuracy and better performance are those with the least amount of predictive errors. It is expected that the models will help estimate the energy savings that different design concepts would produce during the

construction phases.This procedure was developed to generate training data for ANN using a statistical method. The resulting database contains case files that are representative of the office buildings.The statistical models that rely on the multi-perceptron method are more accurate than those that rely on the standard linear regression model. They can reproduce the results of ISO 17000:2008 with high accuracy.This study aimed to develop an ANN model that could predict office buildings' energy consumption and greenhouse gas emissions in Santiago, Chile. The model was tested on a non-standard set of factors.This method will help designers and developers to develop more energy-efficient and accurate methods for calculating greenhouse gas and energy consumption. The framework proposed in this study can be used to develop energy-efficient building standards that are realistic and sustainable. A paper (Wang et al., 2020) is the study, that the data on carbon emissions be accurate for developing effective carbon mitigation strategies. For China, the US, and India, the data on their carbon emissions are different.The US carbon emissions data shows volatile growth and decline. This paper Author proposes a method to improve the accuracy of the data by developing a combination of the ARIMA and the MNGM models. This method could reduce the residual error of the model MNGM by applying ARIMA and BPNN. It can also decrease the forecasting error of the model. As for the US' carbon emissions, it is expected to keep a downward trend during the next couple of decades.

Meanwhile, China and India's carbon emissions will keep growing. The proposed two forecasting techniques have shown that they can improve the forecasting models' accuracy by correcting previous errors. They also exhibited sound performance for the three different types of data.There is a statistical Model called Holt-Winter(Chatfield, 1978) is there for dealing with trends and seasonal variation. According to this paper, some available idiosyncratic modifications can improve the Holt-Winter's Forecasts.There is an argument that a fairer comparison would be that between Box-Jenkins and a non-automatic version of Holt-Winters. There is a suggested method for predicting and estimating trends in specific datasets. If there are multiple trends in data and the forecasting model dimensions increase, this trend estimation becomes a more tedious task. So paper (Sbrana, 2021) has a closed-form result for simple prediction and estimation of the multivariate smooth-trend model, a state-space representation of Holt-Winters celebrated recursions.

3. The Dataset

We have CO_2 and greenhouse gas emission data[1] of 40 years in this work. We have collected the data from the year 1980 to 2019 by using the CAIT data source. The CAIT dataset is the most comprehensive and includes all the sectors along with gases. Greenhouse gas emission data indicates that 60% of the GHG emissions are from the top 10 emitting countries. In this work, we have used the univariate time series data of India, having an increasing CO_2 emission trend.

Before splitting the data into training and testing samples, the data preprocessing is done. Test data is used to evaluate the performance of the models, and better performing models are used for forecasting the CO_2 emissions.

4. Proposed Model

As we talked about in an earlier section that CO2 emission in India is growing very faster and is very dangerous for the environment and ultimately to all living beings, the reduction of this emission should be the highest priority for government and industries. An accurate forecast of emissions would direct help in policy-making and implementing them. So for this purpose, we have a CO2 emission's time series data with some attributes for the last 40 years, from 1980 to 2019. The requirement for this work was a univariate dataset, so it needed some data pre-processing and also data cleaning to deal with not available values in the dataset. The dataset contains two columns, one for years and the second for Co2 emission in metric tons.

Based on these previous data, we have used many time series models to forecast India's ten years CO2 emission. The dataset has an increasing trend and is also stationary and for this particular kind of dataset, we have used three statistical models say ARIMA, SARIMAX, and Holt-Winter model, two machine learning models i.e. Linear Regression model and Random Forest regression model, and in last a deep learning model LSTM (S. Kumar et al., 2021).before using dataset in these models we need to split them into two parts i.e training data and testing data to train and test the models.

Different models have different ways to deal and work with data. Let discuss these six time series models briefly :

[1] https://www.climatewatchdata.org/ghg-emissions?end_year=2018&gases=all-ghg%2Cco2&start_year=1990

Auto-Regressive Integrated Moving Average (ARIMA): We start with a basic stationary model like AR and MA in time series analysis. But to handle non-stationary models, we go with ARIMA. MA, AR, and ARMA are specific cases of the ARIMA model.

A model that uses the dependency between an observation and a residual error from a moving average model applied to lagged observations. ARIMA is an acronym that stands for Auto-Regressive Integrated Moving Average. Specifically,AR (Autoregression), which is a model that uses the dependent relationship between an observation and some number of lagged observations, I (Integrated) is the use of differencing of raw observations in order to make the time series stationary and MA (Moving Average).

A standard notation is used of ARIMA(p, d, q) where the parameters are substituted with integer values to quickly indicate the specific ARIMA model being used. Where p is the number of lag observations included in the model, also called the lag order, d is the number of times that the raw observations are differenced, also called the degree of differencing and q is the size of the moving average window, also called the order of moving average. In the work, we have applied ARIMA (1,2,1) to predict the CO_2 emissions(Wellington, 2019).

Seasonal Autoregressive Integrated Moving Average with an Exogenous Variable (SARIMAX): Before knowing about SARIMAX, we will try to understand SARIMA. SARIMA is Seasonal Autoregressive Integrated Moving Average in which along with the trend there is seasonality as well. This is ARIMA, but having the effect of seasonality is something with those time series that have both trends and seasonality. Now, we have to understand the X factor of SARIMAX, which is an Exogenous Variable that means there is some external factor that is impacting it. It is saying that there are some factors outside of the overall factor that we are studying (Özmen, 2021).

Holt-Winter's Model: The Holt Winter model is a traditional model dealing with time-series data behaviour. These behaviors here mean the average value, the trend increasing or decreasing trend, and the seasonality which is nothing but the repetitive pattern in a cycle. Based on the seasonal component of data, this model has two variations. The first is the Additive Holt-Winter Model, and the second is the Multiplicative Holt-Winter. In this work, the multiplicative model has been used to forecast the next 10 years' data. The multiplicative Holt-Winters is described as a method that calculates the value of the level, trend, and seasonal adjustment that are exponentially smooth. Since this method is best suited for data with increasing trends and seasonality, this paper considers this model for CO_2 emission forecasting (Chatfield, 1978).

Linear Regression: Linear regression is a Machine learning model to solve regression problems. This model solves the problem by assuming a linear relationship between the given input attributes and the output as shown in equation (1).

$$Target\ Output = Input_1 * weight_1 + Input_2 * weight_2 \dots\dots\dots\dots Input_n * weight_n + Bias \quad (1)$$

Here n is a total number of attribute/input features. Weights are also called regression coefficients learned while training the model based on train data. The bias here is nothing but the intercept as linear regression is based on the linear equation. Here time series forecasting is also a regression problem where inputs are previous data. In this work, CO_2 emission data is univariate data, In this paper CO_2 emission data is univariate data, and for each year ,prediction the last three year's data considered as 3 input features for the whole training, testing, and forecasting.Like CO_2 emission of year 1980, 1981, 1982 are used to predict the emission in the year 1983 and so on (Huang & Hsieh, 2020).

Random Forest Regressor: The Random Forest model is a supervised technique for both classification regression and non-linear problems. This method uses the ensemble learning method for regression and is a bagging technique as it uses individual decision trees together to give better results. This can also be used in time series forecasting, but results are not sure to be expected. In this model, data should be in a proper way before fitting into the model. Data splitting has been done as we have univariate CO_2 emission data. One of the advantages of the random forest model is that it handles the missing values and maintains accuracy (Fang et al., 2020).

Long Short-Term Memory (LSTM): LSTM is a deep learning model which is based on a recurrent neural network (RNN). This model is considered to be the best model for processing, classifying, and forecasting time series data. Traditional RNN has the problem of exploding and vanishing gradient descent, so LSTMs were developed to deal with it. For univariate time series forecasting, there is a number of variations of LSTM. Some of them are:Vanilla LSTM, Stacked LSTM, Bidirectional LSTM,CNN LSTM and ConvLSTM (Elsworth & Güttel, 2020).

Flow chart

The flowchart shown in the figure 1, presents the framework of the proposed model. In which, it can be seen that we have applied the time-series and machine learning models as stated above.

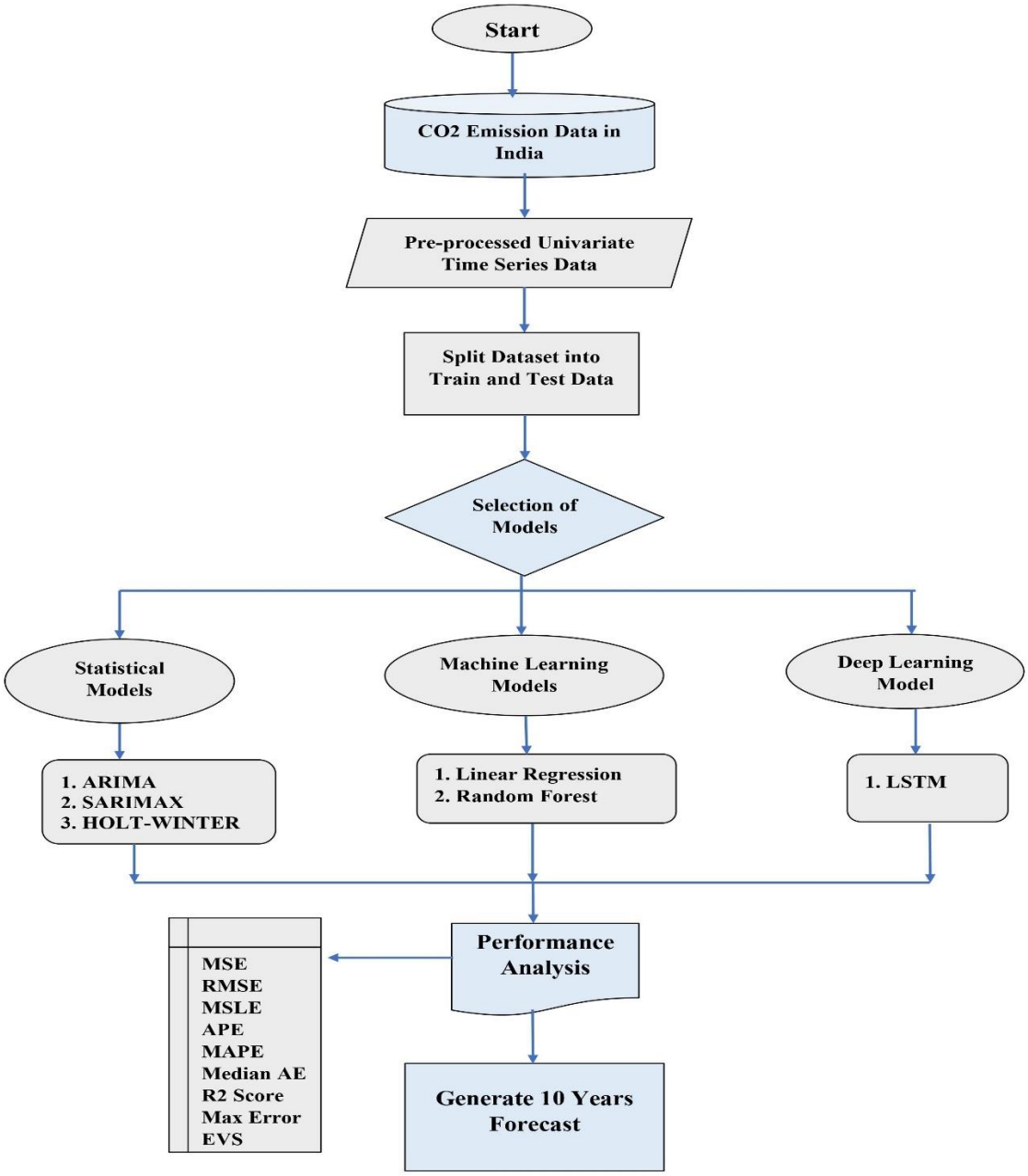

Figure 1: Proposed framework for CO2 Emission forecasting

In this framework, we have applied Statistical, Machine learning and deep learning models to forecast the CO_2 emissions.

For the same, we have used nine performance metrics to analyze the effectiveness of the applied models. Based on the efficiancy of the models, best performing models are used to forecast the CO_2 emissions for the next ten years i.e. from 2020 to 2030.

Before applying the models, we have pre-processed the data to cope up with the missing values. We have split the pre-processed data into two parts; training and testing. Performance of the models on test data shows the efficacy; and accordingly, it is used to forecast the CO_2 emission for the next 10 years till 2030.

5. Performance metrics

In this section, we have discussed the performance metrics used to determine the effectiveness of the models. These models forecast CO_2 emission, and forecasting, which is a kind of regression problem. There are many evaluation metrics in general; we have used herer nine to evaluate the models' effectiveness. Before using all these metrics, we must know about the residual error, i.e. $(y - \hat{y})$. Here y and $\hat{y}$ Indicates the actual and predicted values. The performance metrics used to evaluate the models are as follows.

***Mean Squared Error (MSE)*:** In MSE, first calculate the squares of the residual error as defined above for each data point and then calculate the average of that(R. Kumar et al., 2020). The formula for Mean squared error is shown in equation (2).

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

Where, y_i and $\hat{y}_i$ indicates the actual and predicted values respectively. n indicates the number of data points.

***Root Mean Squared Error (RMSE)*:** It is the same as MSE, the only addition is a square root sign to it(R. Kumar et al., 2020). The formula for Mean absolute error is represented in equation (3).

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

***Mean squared Log Error*:** In MSLE,the residual error is calculated with the logarithm of the original and predicated value.It is an extension on Mean Squared Error (MSE) mainly used when predictions have large deviations.The formula for Mean absolute log error is represented in equation (4).

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

***Mean Absolute Error**:* MAE is the sum of the absolute residual error. That means it does not matter about negative and positive(R. Kumar et al., 2020)i. The formula for Mean absolute error is indicated in equation (5).

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

Mean Absolute Percentage Error: With the addition of percentage to MAE, MAPE is defined(R. Kumar et al., 2020). The formula for Mean absolute percentage error is shown in equation (6).

$$MAPE = \frac{100\%}{n}\sum_{i=1}^{n}\left|\frac{y_i - \hat{y}_i}{y_i}\right| \quad (6)$$

***Median Absolute Error*:** The median absolute error is robust to outliers, making it intresting and capable enough to deal with impacts of outliers on whole predication (Yin & Xie, 2021). The formula for Median absolute error is represented by

$$MedAE(y, \hat{y}) = median(|y_1 - \hat{y}_1|, \ldots\ldots\ldots |y_n - \hat{y}_n|) \quad (7)$$

***Max Error*:** The max error metric calculates the maximum residual error. The formula for Max error is shown in equation (8).

$$Max\ Error = Max|y_i - \hat{y}_i| \quad (8)$$

***R2 Score Error*:** The formula for R2 Square error is also having it significance to measure the effectiveness of the models (Shaikh et al., 2021), it is defined as shown in equation (9).

$$R2 = 1 - \frac{\sum_{i=1}^{n}(\hat{y}_i - y_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y}_i)^2} \quad (9)$$

***Explained Variance Score Error*:** The Explained variance score error is calculated to measure the proportion of the variability of the predictions of the applied machine learning models. Variance is a measure of how far observed values differ from the average of predicted values, i.e., their difference from the predicted value mean (dos Santos et al., 2021). It can be seen as indicated in equation (10).

$$EVS(y, \hat{y}) = 1 - \frac{var(y - \hat{y})}{var(y)} \quad (10)$$

6. Performance Analysis

The performance of the proposed framework is analyzed by writing a program in a python 3.6 programming environment. We have also used Keras and Tensor flow libraries to implement the machine learning and deep learning-based models. The applied models for predictions are analyzed as follows.

6.1 Experiment: Performance observation of the models

We have applied the models on CO_2 emission dataset, after splitting the dataset into training and test samples. Trained models are applied on the test dataset to observe the models' performances. Table 1, shows the performances of the applied models, here, in Table 1, we can see that MSLE value for ARIMA model is not available because it can not be applied.

Table 1: The error rate of models on given CO2 emission dataset

	ARIMA	SARIMAX	Holt-Winter	Random Forest Regressor	Linear Regression	LSTM
MSE	1654252.446	4654.124	20771.376	975659.575	58016.926	3676.646
RMSE	1286.177	68.221	144.123	987.755	240.867	60.635
MSLE	NA	0.830	0.004	0.787	0.0180	0.001
MAE	1099.065	44.155	113.441	771.631	196.571	45.524
MAPE	98.969	6.554	5.043	137.238	12.023	3.101
MEDIAN AE	898.047	30.823	115.816	550.206	155.177	28.898
MAX ERROR	2639.0572	314.016	243.078	2057.741	557.092	135.933
R2 SCORE	-2.744	0.989	0.737	-7.55E+31	0.894	0.990
EVS	-0.010	0.990	0.895	-2.94E+31	0.965	0.990

Further, we have also plotted the bar graph of the applied models to observe the relative performances, as shown in figure 2. Observations from fig-2(a), 2(b), and 2(f), it can be seen that LSTM, SARIMAX, and Holt-Winter models have lowest MSE, RMSE, and MAE values in comparison to other models.
Similarly, thery are having lower MAPE and Max Error values in decreasing order of w.r.t. models SARIMAX, Hot-Winter, and LSTM.
Observations form fig-2(d) SARIMAX, and LSTM have the lowest MAE values. Similarly, when we look to the R2 Score Error and EVS Error in fig- 2(g) and fig-2(i) the models Holt-Winter, Linear Regression, and SARIMAX have the least error value in increasing order. In last we have remaining MSLE referring fig -2(c) LSTM, Holt-Winter, and Linear

Regression have the least error value in ascending order. Also, ARIMA is not participating in it negative predictions, which MSLE cannot handle.

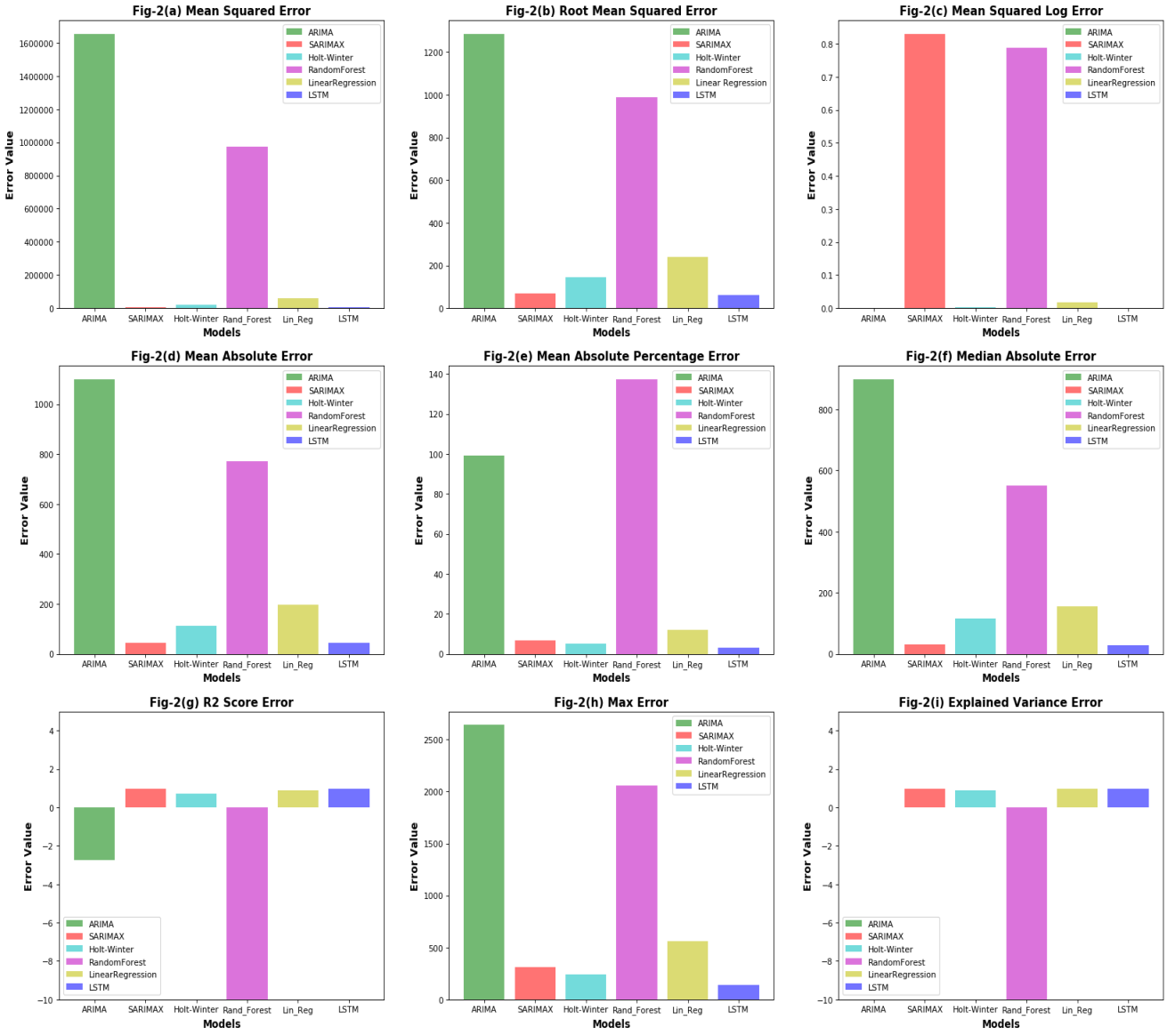

Figure 2: Comparative error plot w.r.t models

Overall observation from figure 2, can be drawn that LSTM, SARIMAX, and Hot-Winter are better performing models for time series forecasting data.

These three models can be applied for effective CO_2 emissions predictions. When looking at all nine performance metric values, we can say that LSTM is on the best performing model.

6.2 Experiment: 10 years forecasting of CO_2 emissions

In this section of the experiment, we have applied the models to forecast the CO_2 emissions for next 10 years i.e. from 2020 to 2030. Table 2, shows the summarized results of forecasted values of the models.

Table 2: 10 Years Forecast of All Models

Year	ARIMA Model	SARIMAX Model	Linear Regression	Random forest Regressor	LSTM Model	Holt-Winter Model
2020	-66.235	2744.336	3313.565	562.259	2798.554	2903.261
2021	-0.280	2932.130	3771.178	562.259	2921.785	3037.139
2022	2.780	3060.027	4582.197	562.259	3102.743	3110.804
2023	2.921	3168.172	5388.920	562.259	3302.388	3281.392
2024	2.921	3355.428	6460.649	562.259	3477.855	3394.070
2025	2.921	3429.468	7679.896	562.259	3691.419	3529.761
2026	2.921	3646.887	9183.267	562.259	3912.318	3595.711
2027	2.921	3655.132	10959.754	562.259	4135.085	3773.705
2028	2.921	3675.008	13106.276	562.259	4383.752	3884.879
2029	2.921	3816.568	15669.673	562.259	4641.486	4022.382
2030	2.921	3969.219	18749.809	562.259	4912.014	4080.618

Further, we have shown the graphical representation of the models to observe the performance w.r.t. actual and predicted values from 1980 to 2020. Then further forecasted for next ten years.

In this sections, we have plotted the graph for actual CO_2 emissions for the entire dataset and predicted the values from 2010 to 2020 by applying the aforesaid models. Further, we have forecasted the values from 2020 to 2030.

6.2.1 ARIMA's Forecasting

From figure 3, it can be observed that ARIMA predicted values are too far from the actual CO_2 emissions. Therefore, it can also be seen that forecasting for 2020 to 2030 will not be appropriate.

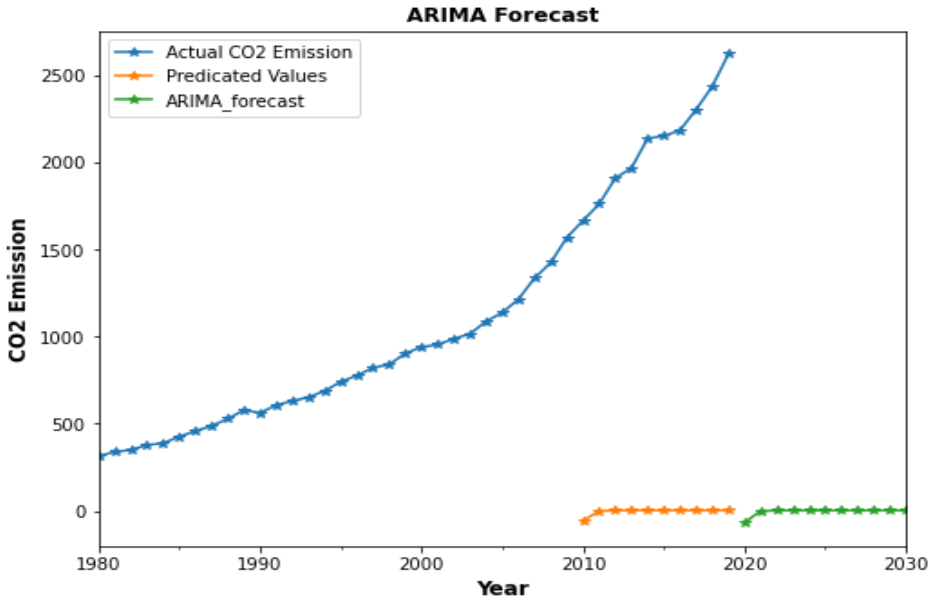

Figure 3 ARIMA forecasting vs. Actual CO_2 emission

6.2.2 SARIMAX Forecasting

SARIMAX forecasting is almost in the same pattern as the actual CO_2 emission can be seen from figure 4. It can be used as one of the appropriate models for forecasting the CO_2 emissions.

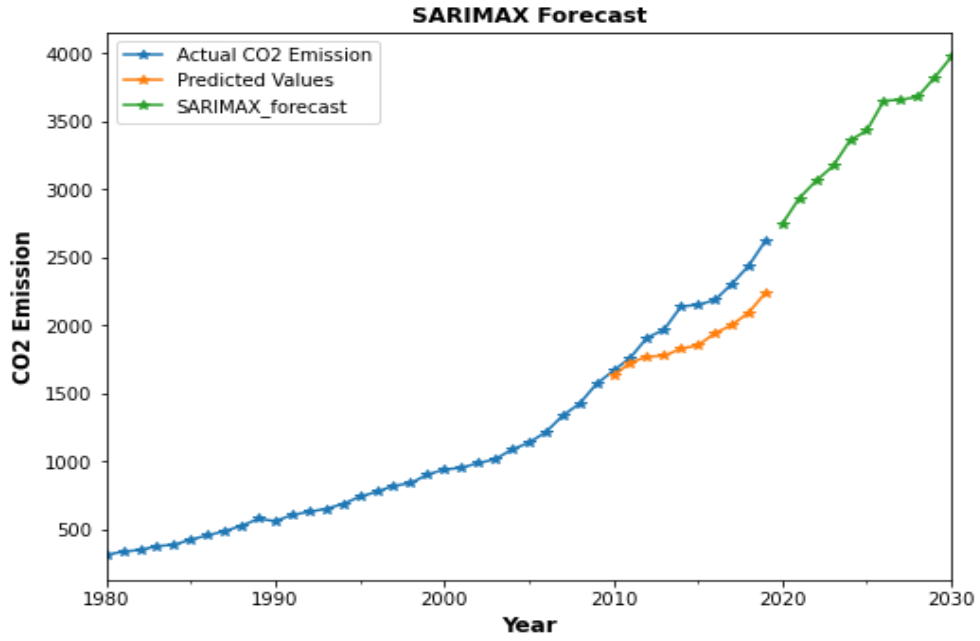

Figure 4. SARIMAX forecasting vs. Actual CO_2 emission

6.2.3 Linear Regression Forecasting

Figure 5, shows that CO_2 predicted values by linear regression and actual emission, which is quite similar. Forecasting values are also in the same pattern as we can see the actual CO_2 emissions.

But because of performance metric values of linear regression on test data, its hard to rely on forecasting.

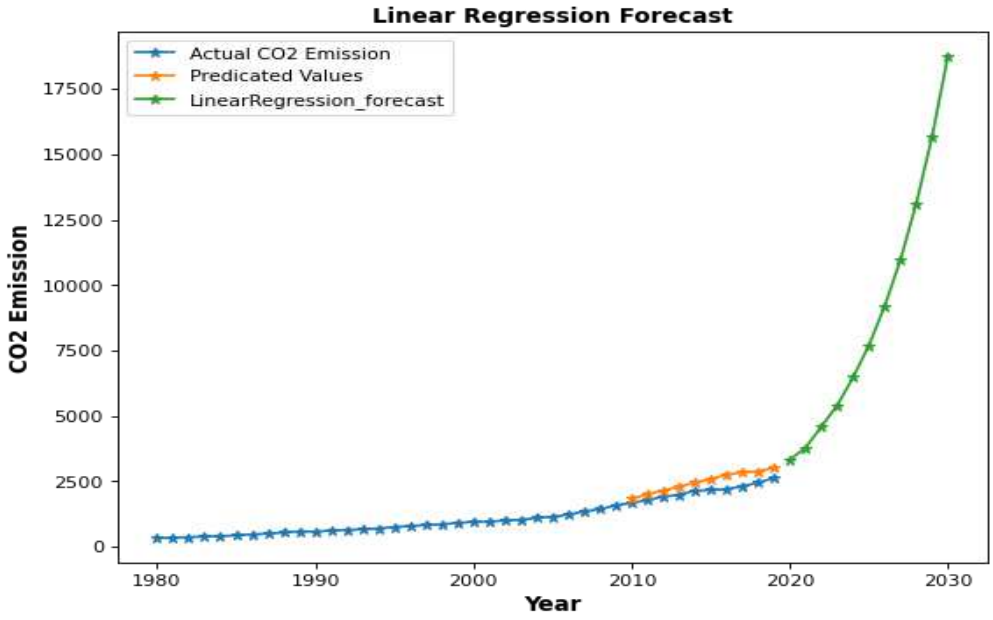

Figure 5. Linear Regression forecasting vs. Actual CO_2 emissions

6.2.4 Random Forest Forecasting (RF)

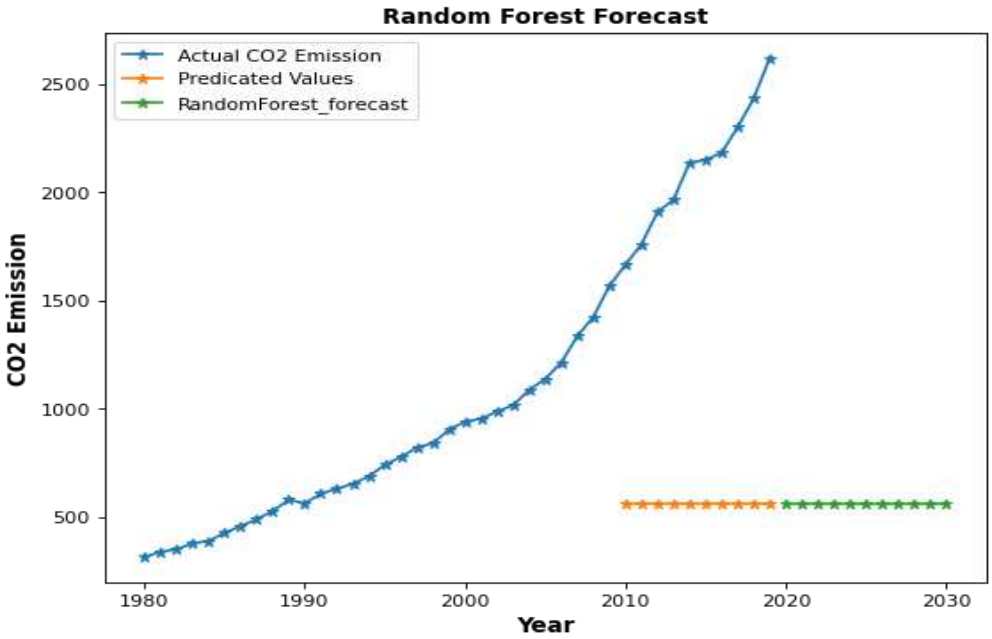

Figure 6. RF forecasting vs Actual CO_2 emission

Figure 2, shows the CO_2 predicted values which is quite different for Actual emissions. Therefore, we can conclude that Random forest model is not appropriate for CO_2 emissions forecasting.

6.2.5 Holt-Winter model Forecasting

We can see that Holt-Winter predicted values are quite similar to actual CO_2 emissions. This model can also be considered one of the effective forecasting models if we compromise a few performance metric values.

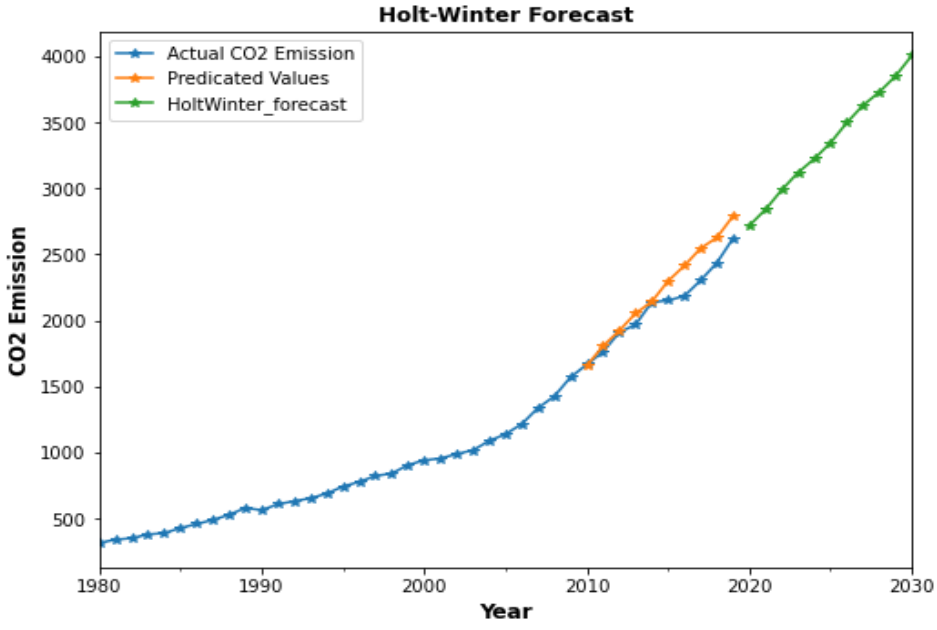

Figure 7. Holt-Winter forecasting vs. Actual CO_2 emissions

6.2.6 LSTM model Forecasting

Figure 8, shows, LSTM behaviour for prediction and forecasting of the CO_2 emissions.

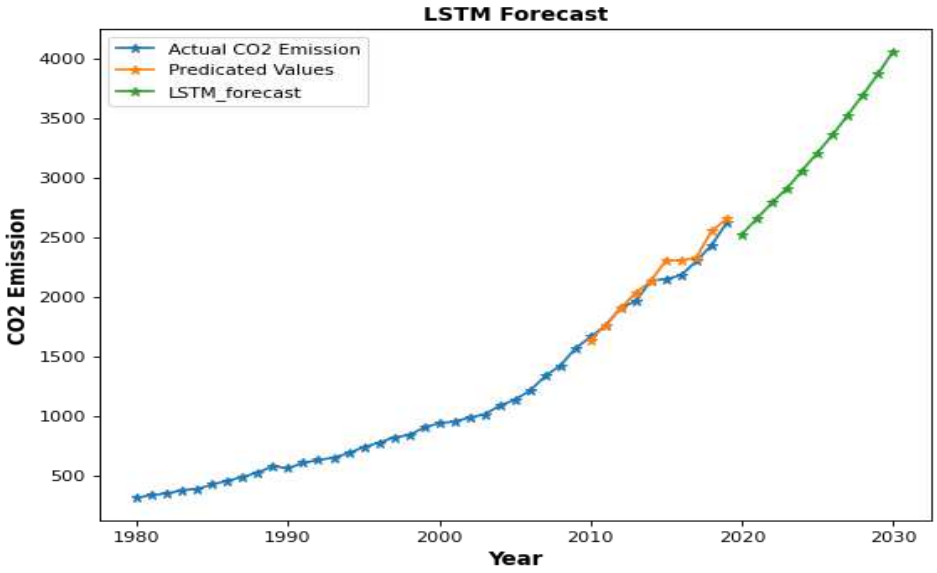

Figure 8. LSTM forecasting vs. Actual CO_2 emissions

Observations from figure 8 shows that when we compare the actual CO_2 emissions and predicted value (as indicated from 2010 to 2020), it is almost same. We have already analysed from figure 2, that LSTM is best performing model in terms of performace metric values. Therefore, we can conclude that LSTM is best peforming model to forecast the CO_2 emission.

6.3 Comparative Forecasting

In this section, we have shown the comparative forecasting values for CO_2 emissions from 2020 to 2030.

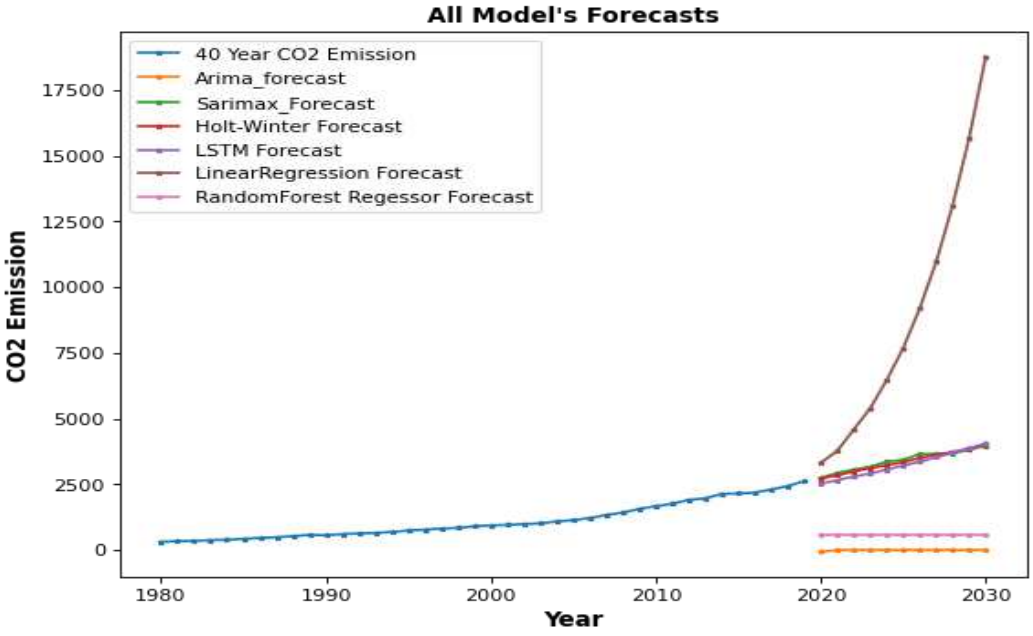

Figure 9 Comparative forecasting of CO_2 emissions from the year 2020 to 2030

Observations from figure 9 show that LSTM, SARIMAX, and Holt-Winter are appropriate forecasting models. Further, when we look at the performance metric values, then LSTM is the best performing model for forecasting the CO_2 emissions.

Here the contribution of this work, can also be seen as to control the CO_2 emission to save human lives.

7. Conclusion and Future Research Directions

In this work, we have applied the machine learning, deep learning and statistical models to forecast the CO_2 emissions in India. For the same, we have considered nine performance metrics to evaluate the effectiveness of the applied models. Observations from the results show that LSTM, SARIMAX, and Holt-Winter are better performing models to forecast CO2 emission.

To analyze the performance of the applied models, we have used the 40 years of CO_2 emissions data. Section 6.2.6 and performance metric values conclude that LSTM is the best forecasting model. Appropriate CO_2 emission forecasting would be fruitful for the upcoming governments to frame their policies to abide by the UN-specific reduction measures, i.e. reducing CO2 emission by 58% by 2030.

This work has not inculcated many factors such as the growth of population, advancing technology, switch to renewable energy sources, and future government actions. These are some exogenous factors that future researchers can explore. They may find a way to deal with such deviations due to these factors so that the new models could forecast data with more accurate CO_2 emissions.

Declarations

Ethical Approval Not applicable

Consent to Participate Not applicable

Consent to Publish Permitted for publication as per journal guidelines

Funding Not applicable

Authors' contributions

Surbhi Kumari: Introduction, literature review, and performed the experiments.

Sunil Kumar Singh: conceptualised and designed the experiments and analyzed and interpreted the results.

Conflicts of interest/Competing interests There is no conflict of interest whatsoever with any other work by any of the authors of this work

Data Availability The data that support the findings of this study are openly available and it will be provided on request.

References

Abdullah, L., & Pauzi, H. M. (2015). Methods in forecasting carbon dioxide emissions: A decade review. *Jurnal Teknologi*, *75*(1).

Ahmed, N. K., Atiya, A. F., El Gayar, N., & El-Shishiny, H. (2010). An empirical comparison of machine learning models for time series forecasting. *Econometric Reviews*, *29*(5), 594–621. https://doi.org/10.1080/07474938.2010.481556

Amarpuri, L., Yadav, N., Kumar, G., & Agrawal, S. (2019). Prediction of CO 2 emissions using deep learning hybrid approach: A Case Study in Indian Context. *2019 Twelfth International Conference on Contemporary Computing (IC3)*, 1–6.

Bonga, W. G., & Chirowa, F. (2014). Level of Cooperativeness of Individuals to Issues of Energy Conservation. *Available at SSRN 2412639*.

Chatfield, C. (1978). The Holt-winters forecasting procedure. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, *27*(3), 264–279.

Crespo Cuaresma, J., Hlouskova, J., Kossmeier, S., & Obersteiner, M. (2004). Forecasting electricity spot-prices using linear univariate time-series models. *Applied Energy*, *77*(1), 87–106. https://doi.org/10.1016/S0306-2619(03)00096-5

dos Santos, P. R. S., de Souza, L. B. M., Lélis, S. P. B. D., Ribeiro, H. B., Borges, F. A. S., Silva, R. R. V, Carvalho Filho, A. O., Araujo, F. H. D., Rabêlo, R. de A. L., & Rodrigues, J. J. P. C. (2021). Prediction of COVID-19 using time-sliding window: the case of Piauí state-Brazil. *2020 IEEE International Conference on E-Health Networking, Application & Services (HEALTHCOM)*, 1–6.

Elsworth, S., & Güttel, S. (2020). *Time Series Forecasting Using LSTM Networks: A Symbolic Approach*. http://arxiv.org/abs/2003.05672

Fang, X., Liu, W., Ai, J., He, M., Wu, Y., Shi, Y., Shen, W., & Bao, C. (2020). Forecasting incidence of infectious diarrhea using random forest in Jiangsu Province, China. *BMC Infectious Diseases*, *20*(1), 1–8.

Gopu, P., Panda, R. R., & Nagwani, N. K. (2021). Time Series Analysis Using ARIMA Model for Air Pollution Prediction in Hyderabad City of India. In *Soft Computing and Signal Processing* (pp. 47–56). Springer.

Huang, C.-H., & Hsieh, S.-H. (2020). Predicting BIM labor cost with random forest and simple linear regression. *Automation in Construction*, *118*, 103280.

Kumar, R., Kumar, P., & Kumar, Y. (2020). Time Series Data Prediction using IoT and Machine Learning Technique. *Procedia Computer Science*, *167*, 373–381. https://doi.org/10.1016/j.procs.2020.03.240

Kumar, S., Mishra, S., & Singh, S. K. (2020). A machine learning-based model to estimate PM2. 5 concentration levels in Delhi's atmosphere. *Heliyon*, *6*(11), e05618.

Kumar, S., Mishra, S., & Singh, S. K. (2021). Deep Transfer Learning-based COVID-19 prediction using Chest X-rays. *Journal of Health Management*, *23*(4), 730–746.

Lepore, A., dos Reis, M. S., Palumbo, B., Rendall, R., & Capezza, C. (2017). A comparison of advanced regression techniques for predicting ship CO2 emissions. *Quality and Reliability Engineering International*, *33*(6), 1281–1292.

Masini, R. P., Medeiros, M. C., & Mendes, E. F. (2021). Machine Learning Advances for Time Series Forecasting. *Journal of Economic Surveys*. https://doi.org/10.1111/joes.12429

Nontapa, C., Kesamoon, C., Kaewhawong, N., & Intrapaiboon, P. (2020). A New Time Series Forecasting Using Decomposition Method with SARIMAX Model. *International Conference on Neural Information Processing*, 743–751.

Nyoni, T., & Bonga, W. G. (2019). Prediction of co2 emissions in india using arima models. *DRJ-Journal of Economics & Finance*, *4*(2), 1–10.

Özmen, E. S. (2021). Time Series Performance and Limitations with SARIMAX: An Application with Retail Store Data. *Electronic Turkish Studies*, *16*(5).

Pino-Mejías, R., Pérez-Fargallo, A., Rubio-Bellido, C., & Pulido-Arcas, J. A. (2017). Comparison of linear regression and artificial neural networks models to predict heating and cooling energy demand, energy consumption and CO2 emissions. *Energy*, *118*, 24–36.

Sbrana, G. (2021). High-dimensional Holt-Winters trend model: Fast estimation and prediction. *Journal of the Operational Research Society*, *72*(3), 701–713.

Shaikh, S., Gala, J., Jain, A., Advani, S., Jaidhara, S., & Edinburgh, M. R. (2021). Analysis and Prediction of COVID-19 using Regression Models and Time Series Forecasting. *2021 11th International Conference on Cloud Computing, Data Science & Engineering (Confluence)*, 989–995.

Wang, Q., Li, S., & Pisarenko, Z. (2020). Modeling carbon emission trajectory of China, US and India. *Journal of Cleaner Production*, *258*, 120723.

Wellington, G. (2019). Emissions in India using ARIMA Models 2 . Determine Stationarity of Time Series 4 . Diagnostic Checking 3 . Model Identification and Estimation 5 . Forecasting and Forecast Evaluation. *Dynamic Research Journals*, *4*(2), 1–10.

Yin, L., & Xie, J. (2021). Multi-temporal-spatial-scale temporal convolution network for short-term load forecasting of power systems. *Applied Energy*, *283*, 116328.

Zuo, Z., Guo, H., & Cheng, J. (2020). An LSTM-STRIPAT model analysis of China's 2030 CO2 emissions peak. *Carbon Management*, *11*(6), 577–592. https://doi.org/10.1080/17583004.2020.1840869