Environmental Science and Pollution Research (2023) 30:116601–116616
https://doi.org/10.1007/s11356-022-21723-8

GIS APPLIED TO SOIL-AGRICULTURAL HEALTH FOR ENVIRONMENTAL SUSTAINABILITY

Machine learning-based time series models for effective CO_2 emission prediction in India

Surbhi Kumari[1] · Sunil Kumar Singh[1]

Received: 20 January 2022 / Accepted: 25 June 2022 / Published online: 2 July 2022

Abstract

China, India, and the USA are the countries with the highest energy consumption and CO_2 emissions globally. As per the report of datacommons.org, CO_2 emission in India is 1.80 metric tons per capita, which is harmful to living beings, so this paper presents India's detrimental CO_2 emission effect with the prediction of CO_2 emission for the next 10 years based on univariate time-series data from 1980 to 2019. We have used three statistical models; autoregressive-integrated moving average (ARIMA) model, seasonal autoregressive-integrated moving average with exogenous factors (SARIMAX) model, and the Holt-Winters model, two machine learning models, i.e., linear regression and random forest model and a deep learning-based long short-term memory (LSTM) model. This paper brings together a variety of models and allows us to work on data prediction. The performance analysis shows that LSTM, SARIMAX, and Holt-Winters are the three most accurate models among the six models based on nine performance metrics. Results conclude that LSTM is the best model for CO_2 emission prediction with the 3.101% *MAPE* value, 60.635 *RMSE* value, 28.898 *MedAE* value, and along with other performance metrics. A comparative study also concludes the same. Therefore, the deep learning-based LSTM model is suggested as one of the most appropriate models for CO_2 emission prediction.

Keywords Time series forecasting · Linear regression · Random forest regressor · Air pollution · CO_2 emissions · Holt-Winters · LSTM

Introduction

According to the Ministry of Statistics and Programme Implementation, UN (United Nations, Department of Economic and Social Affairs, Population Division 2019), the current population of India is 1,400,517,328 as of January 2022; based on interpolation of the latest United Nations data, India is 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 and 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.

Increasing CO_2 emissions can affect human health in two ways; directly and indirectly. It affects directly when inhaled in high dosage and can be the cause of serious diseases such as breathlessness, blindness, dizziness, and even delirium (Ağbulut 2022). Global problems such as climate change, acid rain, and global warming can also be seen in the indirect form of high CO_2 emissions (Ağbulut 2022; Bakay and Ağbulut 2021; Liu et al. 2020). All these forms of emissions are highly hazardous for human beings and the environment. The increased flooding, landslide, cloud bursts, etc., are already evident and would further increase if we continue to go in the same way.

As per the (The Lancet 2016) report, approximately 6.5 million peoples die annually due to severe diseases caused by air pollution worldwide. And this number is greater than

Responsible Editor: V.V.S.S. Sarma

✉ Sunil Kumar Singh
sksingh@mgcub.ac.in; sunilsingh.jnu@gmail.com

Surbhi Kumari
surbhigupta387@gmail.com

[1] Dept. of Computer Science and Information Technology, Mahatma Gandhi Central University, Motihari, Bihar, India

the number of deaths caused by tuberculosis, AIDS/HIV, and road accidents combinedly. A study by Mele and Magazzino et al. (2021) and Solgi and Keramaty (2016) indicated that around three-quarters of the population of India are exposed to air pollution, far higher than the minimum standard set by the Indian government. And it is four times greater than the standard set by the WHO (World Health Organization).

A rapid increase in the CO_2 emission is also caused by the exponential rise in the number of vehicles on the road. It is because of the increment in population all across the world. A study by Ağbulut (2022) and Ahmadi (2019) reveals that the number of vehicles will be three times more than that of today by 2050. Therefore, to mitigate the adverse effect of rising CO_2 emissions, a few concrete steps are required to be taken by the decision and policymakers in the country.

Forecasting CO_2 emissions are also one important key to creating awareness among the public on solving environmental problems (Abdullah and 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 until 2030. By carefully examining the outcomes of different models, we have found a few models to be fairly viable forecasts. The data used is from the past 40 years, which plots the increase in CO_2 emission against time.

The forecast we have come up with will help us understand our emission pace and further correction we need 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 the 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 the best suited for this kind of CO_2 emission data.

And it gives an idea for future papers which could use these models and refine the future forecast by taking into account the other exogenous variables such as increasing population, technological 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.

The remaining part of the paper is organized as follows; the "Literature review" section talks about the work done in the area of CO_2 emission forecasting and a few similar works. Dataset source and its descriptive analysis are presented in the "The dataset" section. The "Proposed model" and the "Performance metrics" sections describe the proposed methodology and performance metrics used to evaluate the model. Finally, a performance analysis of the model and a comparative study with recent works is done in the "Performance analysis" section. The "Conclusion and future research directions" section concludes the work along with policy implications, limitations, and future research directions.

Literature review

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

In Elsworth and 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 solved the limitation of other models in the initialization of random weights. Also, their model is faster without affecting the performance of forecasting. Also, Zuo et al. (2020) collected the CO_2 emission data from the different provinces of China and proposed a model named LSTM-STRIPAT, an integrated model to predict emissions in 2020, and in Amarpuri et al. (2019), the research is aimed at predicting the CO_2 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 by Kumar et al. (2020), in which they have collected the data from Delhi and National Capital Region (NCR), India, to predict the Air quality index. They have applied the machine learning models and measured the performances in terms of *MAE*, *MSE*, *RMSE*, and *MAPE* metrics. Similar work by Kumar et al. (2020) measured the $PM_{2.5}$ pollutant particles in the Delhi atmosphere. In this work, Extra-Trees regression and AdaBoost-based regression models are applied to predict $PM_{2.5}$ concentrations effectively. They have also considered additional atmospheric and surface factors such as wind speed, atmospheric temperature, and pressure.

The work (Ahmed et al. 2010) applied eight machine learning models to famous M3 time series competition data and compared them. The models' multilayer perceptron and the Gaussian process regression performed the best. In Nyoni and Bonga (2019), authors used the Box-Jenkins

ARIMA approach on time series data of CO_2 emission in India from 1960 to 2017, and based on the forecast, they suggested five policy prescriptions to improve the environmental conditions.

ARIMA model has also been used to analyse air pollutants and to predict them based on historical data (Gopu et al. 2021). And it is 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 work (Nontapa et al. 2020) analysed and forecasted time-series 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 the paper (Özmen 2021) with the application of retail store data. In this work on CO_2 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 diarrhea 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), the author describes that with the introduction of the EU's new CO_2 emissions regulation, ships' operators are required to implement systems that will monitor and report their CO_2 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 extensively analyzes the various regression techniques that can exploit ships' navigation information. 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 CO_2 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 CO_2 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 show that the linear regression model with higher accuracy and better performance are those with the least value of predictive errors. It is expected that the models will help to 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 to predict office buildings' energy consumption and greenhouse gas emissions in Santiago, Chile. 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) shows that the data on carbon emissions are accurate for developing effective carbon mitigation strategies. For China, the US, and India, the data pattern on their carbon emission are different. The US carbon emission data shows volatile growth and decline. In this paper, the author proposes a method to improve the accuracy of the data by developing a combination of the ARIMA and the Metabolic Nonlinear Grey Model (MNGM). 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 are expected to keep a downward trend during the next couple of decades.

In Mele and Magazzino (2020), the author studies the monthly data of iron and steel industries, air pollution, and economic growth for China from 2000 to 2017 to find the relationship between them. The approach used for this work is the long short-term memory (LSTM) model. As per the paper, the iron and steel industries are the most energy-consuming and air-polluting sectors, which became the primary concern of the Chinese government. For energy supply, the iron and steel industry depend on coal, which produces a high percentage of direct CO_2 emissions. In this work, authors did not choose standard econometric techniques, rather they have used the machine learning method; LSTM comes under a recurrent neural network to analyze the relationships. And lastly, they successfully showed the strong relationship between the Iron and steel industries, and CO_2 emissions, which may help the government to make managerial policy implications.

In Magazzino et al. (2021), the author's interest is to examine the link between ICT, economic growth, urbanization, environmental pollution, and energy consumption using the ML models. In this work, time series, panel data of 25 OECD countries from 1990 to 2017 has been used. The author used the D2C algorithm, which predicts the linkage between two variables in a multivariate setting by (i) constructing a set of relationship features based on asymmetric descriptors of the multivariate dependency and (ii) learning a mapping between the features and the presence of a causal link using a classifier. The same algorithm is used in work (Mele and Magazzino 2021) over the annual data from yeras1980 to 2018 to analyse the relationship between pollution emission, economic growth, and COVID-19. For COVID-19, data was collected from 29th January 2020 to 18th May 2020.

In Bakay and Ağbulut (2021), SVM, ANN, and deep learning models have been used for forecasting five gases that are CO_2, CH_4, N_2O, F-gases, and total GHG. This work has used Turkey's energy production data from 1990 to 2014. Findings indicate that deep learning-based models perform better with *RMSE*, *rRMSE*, *MBE*, and *MAPE* performance metrics.

In MK (2020), the author has worked on the role of energy use in the prediction of CO_2 emission and economic growth in India using artificial neural networks. The author uses two optimizers as Adam and Stochastic Gradient Descent for prediction, and it is found that Adam optimizer works as a better optimizer for their work. There are some other suggested models too for forecasting (Hewamalage et al. 2021) author talked about the recurrent neural network model for competition data. In this work, they have also presented a detailed discussion on the current status of the RNN model and its future directions.

In 1978, a statistical model named Holt-Winters (Chatfield 1978) was introduced, which deals 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 Box-Jenkins and a non-automatic version of Holt-Winters would be a fairer comparison. There is a suggested method for predicting and estimating trends in specific datasets. If multiple trends in data and the forecasting model dimensions increase, this trend estimation becomes a more tedious task, so the 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.

In Magazzino (2017), author investigated the relation among the CO_2 emission, economic growth, and energy used in the South Caucasus area and Turkey using three variable Vector Autoregressive (VAR) technique over the panel data. The empirical analysis uses yearly data of real per capita GDP, per capita CO_2 emissions, and per capita energy uses in 1992–2013 for Armenia, Azerbaijan, Georgia, and Turkey. Analysis shows the reaction of CO_2 emissions to energy use is negative and statistically significant in both the estimated coefficients and impulse responses.

Later in Magazzino et al. (2020), authors investigate the casual relationship among solar and wind energy production, coal consumption, economic growth, and CO_2 emissions for three most significant contributors of CO_2 emission and energy consumers i.e. China, India, and the USA. They have used an advanced methodology in machine learning to verify the predictive causal linkages among variables. The Casual Direction from Dependency (D2C) algorithm set CO_2 emissions as the target variables.

The authors (Magazzino et al. 2021) investigated the casual relation between renewable energy technologies, biomass energy consumption, per capita GDP, and CO_2 emissions for Germany. In this work, they devised an innovative model, the Quantum model, combined with machine learning models to find the relation. As a result, they find biomass energy uses powerful for the reduction of CO_2 emission and use of renewable energy more potent for the CO_2 reduction.

After exploring the literature, we have observed that several methods, including various algorithms, theories, and mathematical models, have been studied to track trends, and forecast different harmful gases such as CO_2, CH_4 etc., with energy consumption, the death rate in COVID-19. Furthermore, we have observed that statistical, machine learning and deep learning models are not applied together. To fill this gap, we have applied all the three classes of models for effective prediction of CO_2 emission, which tells about the best-performing models and analyses the reasons behind their performance.

The dataset

We have used 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, which 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.

The data pre-processing is done before splitting the data into training and testing samples. Test data is used to evaluate the performance of the models, and better performing models are used for forecasting the CO_2 emissions.

The descriptive analysis of the dataset CO_2 emission data is shown below in Table 1. After observing the min, median, and max values, we can see the increasing behaviour of CO_2 emission.

The dataset used in the work has an increasing trend as shown in Fig. 1. From Fig. 1, we can observe that trend indicates the increasing or decreasing behaviour of the data w.r.t. time. Furthermore, seasonality is analysed to observe the dataset's characteristics in terms of regular and predictable changes that recur every calendar year. Any predictable fluctuation or pattern that recurs or repeats over a one-year period is said to be seasonal.

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

Table 1. Descriptive analysis

Variables→	Year	CO_2 emission
Count	40	40
Mean	1999.500	1106.632
Median	1999.50	922.13
Standard deviation	11.690	673.169
Standard error	1.848	106.438
Kurtosis	−1.200	−0.622
Skewness	0.000	0.772
Minimum	1980.000	314.016
Maximum	2019.000000	2620.000

Furthermore, we have analysed the time-series variable by observing the original flow of the dataset, shown (in top of Fig. 1) from the year 1980 to 2019. We have also analysed the data's behaviour by observing the data's Trend, Seasonality, and Residual.

We have also analysed the residual behaviour of the dataset, and removed which were left over observing the trends and seasonality of the data. In time-series models, we assume that data is stationary and only the residual components satisfy the stationarity condition.

Stationarity behaviour of the data is tested using augmented Dickey-Fuller (ADF) (Ajewole et al. 2020) test and Kwiatkowski-Phillips-Schmidt-Shin (KPSS) (Baum 2018) and converted into stationary data by applying the shifting and differencing approaches.

Proposed model

As we have discussed in section 1, CO_2 emission in India is growing significantly faster and is quite dangerous for the environment and ultimately to all living beings, so the reduction of CO_2 emission should be the highest priority for the government and industries. An accurate forecast of emissions would directly help in policy-making and implementing them, so for this purpose, we have a CO_2 emission 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 data cleaning to deal with not available values in the dataset. The dataset contains two columns, one for years and the second for CO_2 emission in metric tons.

We have used a few time series-based models to forecast India's 10-year CO_2 emission. The dataset has an increasing trend and is also stationary. For this particular kind of dataset, we have used three statistical models: ARIMA, SARIMAX, and Holt-Winters 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 applying the models to the dataset, we have split it into two parts i.e. training data and testing data to train and test the models.

Different models have different ways of dealing and working with the data. Here, we have discussed these six time-series models briefly:

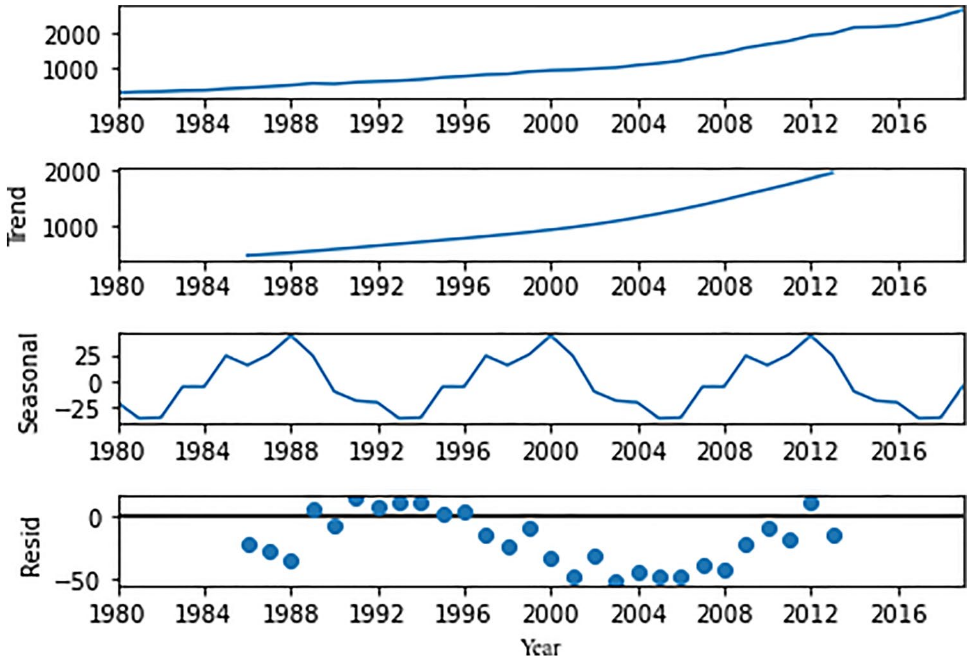

Fig. 1 Time series properties of variable

Autoregressive-integrated moving average

We start with a basic stationary model like AR and MA in time series analysis. But to handle non-stationary models, we go with utoregressive-integrated moving average (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 applies to lagged observations. ARIMA is an acronym that stands for autoregressive-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 to make the time series stationary and MA (moving average).

A standard notation is used for ARIMA (p, d, q), where the parameters are substituted with integer values to indicate the specific ARIMA model being used quickly. 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 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 some external factor is impacting it. It is said that there are some factors outside the overall factor we are studying (Özmen 2021).

Holt-Winters model

The Holt-Winters model is a traditional model dealing with time-series data behavior. These behaviors mean the average value, the 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-Winters model, and the second is the multiplicative Holt-Winters. 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 Eq. (1).

$$\begin{aligned}\text{Target output} = \text{Input}_1 \times \text{weight}_1 + \text{Input}_2 \\ \times \text{weight}_2 \ldots\ldots\ldots\ldots\ldots \text{Input}_n \times \text{weight}_n + \text{Bias}\end{aligned} \tag{1}$$

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 is nothing but the intercept as linear regression is based on the linear equation. Time series forecasting is also a regression problem where inputs are previous data. In this work, CO_2 emission data is univariate data, and for each year's prediction, the last 3 year's data are considered 3 input features for the whole training, testing, and forecasting. Like CO_2 emission of the years 1980, 1981, 1982 are used to predict the emission in the year 1983 and so on (Huang and 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 combines individual decision trees 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

Long short-term memory (LSTM) is a deep learning model based on a recurrent neural network (RNN) with the addition of a small memory element. A common LSTM unit is made of a cell, an input gate, an output gate, and a forget gate. The cell remembers values over arbitrary time intervals, and the three gates regulate the flow of information into and out of the cell. This model is considered to be the best model for processing, classifying, and forecasting time

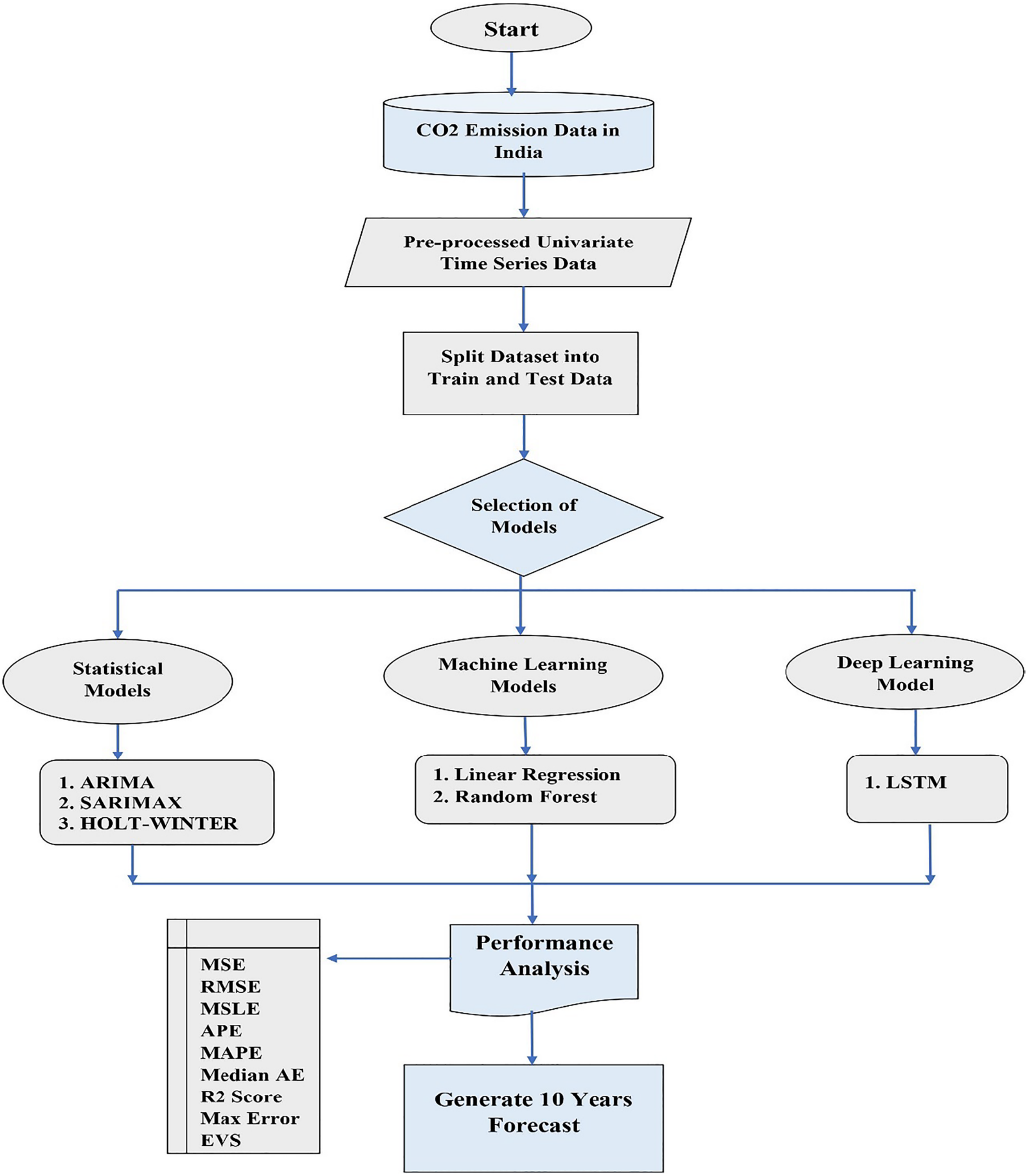

Fig. 2 Proposed framework for CO_2 emission forecasting

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 are several variations of LSTM. Some of them are Vanilla LSTM, Stacked LSTM, Bidirectional LSTM, CNN LSTM, and ConvLSTM (Elsworth and Güttel 2020).

Flow chart

The flowchart shown in Fig. 2 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.

This framework initially pre-processed the dataset collected from CAIT data source. It contains 40 years of data starting from 1980 to 2019. In pre-processing, we have filled the missing values for the smooth functioning of the applied models.

We have split the pre-processed data into two parts; training and testing and then after that, we selected the six models that are from a statistical, machine learning and deep learning models to forecast the CO_2 emissions. The reason for selecting these modes is that these are suggested models for time series forecasting in various fields like business, economics, environmental science, and in science and technology.

After model selection, we trained the model by fitting the training data and then applied the trained models over test data to observe the performance. Furthermore, we have used nine performance metrics to analyse the effectiveness of the applied models. Before applying the models, the performance of the models on test data shows the efficacy; and accordingly, it is used to forecast the CO_2 emission for the next ten (10) years until 2030.

Performance metrics

In this section, we have discussed the performance metrics used to determine the effectiveness of the models. These models are used to analyze CO_2 emission, and forecasting, which is a kind of regression problem. There are many evaluation metrics; we have used 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 $\hat{y}$ indicate the actual and predicted values. The performance metrics used to evaluate the models are as follows.

Mean squared error

In mean squared error (MSE), first, calculate the squares of the residual error as defined above for each data point and then calculate the average of that(Ağbulut et al. 2021a; R. Kumar et al. 2020)

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

where y_i and $\hat{y}_i$ indicate the actual and predicted values, respectively. n indicates the number of data points. The study of (Bakay and Ağbulut 2021) suggests that MSE values can vary from 0 to ∞ where smaller values are preferable.

Root mean squared error

It is the same as MSE, the only addition is a square root sign (R. Kumar et al. 2020) The formula for mean absolute error is represented in Eq. (3).

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

Root mean squared error (*RMSE*) takes the values from 0 to ∞, and the smaller *RMSE* values are desirable (Bakay and Ağbulut 2021).

Mean squared log error

In mean squared log error (*MSLE*), the residual error is calculated with the logarithm of the original and predicated values. 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 Eq. (4).

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

Also, this metric does not deal with negative values, so if there is any negative error value for any data point in the dataset, then this metric will not be applicable.

Mean absolute error

Mean absolute error (*MAE*) is the sum of the absolute residual error. That means it does not matter about negative or positive (R. Kumar et al. 2020). The formula for mean absolute error is indicated in Eq. (5).

$$\mathrm{MAE} = \frac{\sum_{i=1}^{n}\left|\hat{y}_i - y_i\right|}{n} \tag{5}$$

Mean absolute percentage error

Mean absolute percentage error (*MAPE*) is defined with the addition of percentage to *MAE* (R. Kumar et al. 2020). For a good model, a smaller *MAPE* value is desirable. The formula for mean absolute percentage error is shown in Eq. (6).

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

Table 2: The error rate of models on given CO_2 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
MedianAE	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
R^2 *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

If the *MAPE* value < 10%, then the prediction will be classified as the "high prediction accuracy."

If the *MAPE* value ranges between 10 and 20%, the prediction will be classified as "good prediction accuracy."

If the *MAPE* value ranges between 20 and to 50%, predictions will be classified as "reasonable predication accuracy."

And if *MAPE* value > 50%, then predication will be classified as "inaccurate prediction accuracy"(Ağbulut et al. 2021b).

Median absolute error

The median absolute error is robust to outliers, making it interesting and capable enough to deal with the impacts of outliers on whole prediction (Yin and Xie 2021). The formula for median absolute error is represented by

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

The median absolute error (*MedAE*) is a robust measure of the variability of a univariate sample of quantitative data. Its value could be anything between 0 and infinity. Therefore, the lesser the value, the higher the accuracy of the model will be.

Max error

The max error metric calculates the maximum residual error. The formula for max error is shown in Eq. (8).

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

R^2 score error

The formula for R^2 square error also has it significance in measuring the effectiveness of the models (Shaikh et al. 2021; Ağbulut et al. 2020); it is defined as shown in Eq. (9).

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

R^2 score error, also known as the determination coefficient, ranges from 0 to 1. It gives a clue on how well the trends of the model results is able to track the trends of actual data (Ağbulut et al. 2021b). If the value is closer to 1, the model will be considered more accurate, so the more significant values is desirable.

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 measures 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 Eq. (10).

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

It is a fundamental concept for the R^2 score error; also, that model will be the best fit with the higher EVS value. It also ranges from 0 to 1(Good and Fletcher 1981).

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.

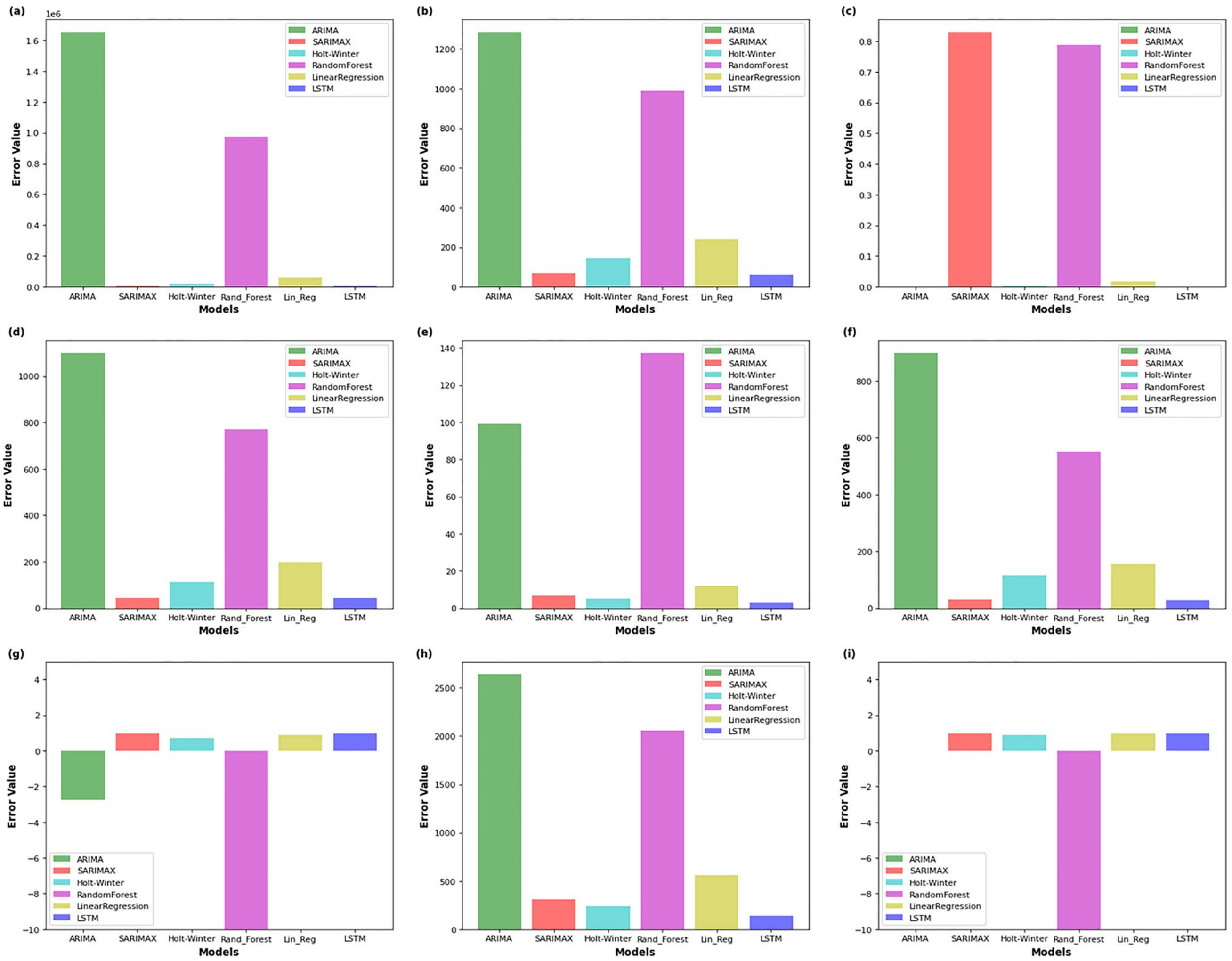

Fig. 3 Comparative error plot w.r.t. models

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 to the test dataset to observe the models' performances. Table 2 shows the performances of the applied models; here, in Table 2, we can see that the *MSLE* value for the ARIMA model is not available because it cannot be applied to a negative error value.

Furthermore, we have also plotted the bar graph of the applied models to observe the relative performances, as shown in Fig. 3. Observations from Fig. 3(a), 3(b), and 3(f), it can be seen that LSTM, SARIMAX, and Holt-Winters models have the lowest *MSE*, *RMSE*, and *MAE* values in comparison to other models.

Similarly, they have lower *MAPE* and max error values in decreasing order of w.r.t. models SARIMAX, Hot-Winter, and LSTM.

Observations from Fig. 3(d), SARIMAX and LSTM have the lowest *MAE* values. Similarly, when we look at the R^2 score error and EVS error in Fig. 3(g) and (i), the models Holt-Winters, linear regression, and SARIMAX have the least error value in increasing order. Last, we have the remaining *MSLE* referring to Fig. 3(c) LSTM, Holt-Winters, and linear regression with the least error value in ascending order. Also, ARIMA is not participating in it negative predictions, which *MSLE* cannot handle.

Overall observation from Fig. 3 shows 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 emission predictions. When looking at all nine performance metric values, we can say that LSTM is one of the best-performing models.

Table 3 10 years forecast of all model

Year	ARIMA model	SARIMAX model	Linear regression	Random forest regressor	LSTM model	Holt-Winters 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

Furthermore, we have shown the graphical representation of the models to observe the performance w.r.t. actual and predicted values from 1980 to 2019. Then, further forecasted CO_2 emission for the next 10 years.

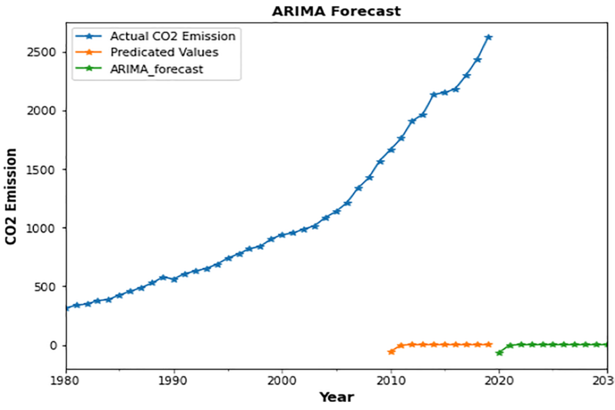

Fig. 4 ARIMA forecasting vs. actual CO_2 emission

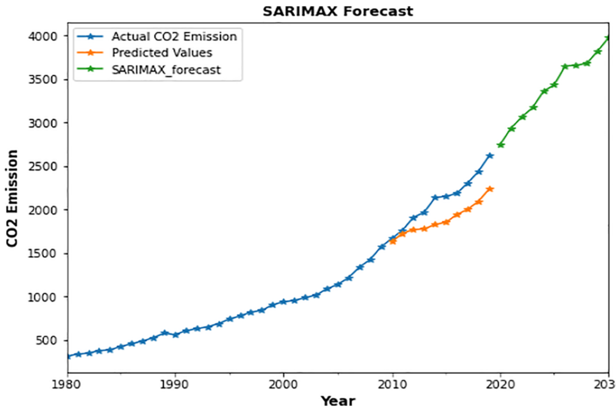

Fig. 5 SARIMAX forecasting vs. actual CO_2 emission

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 the next 10 years i.e. from 2020 to 2030. Table 3 shows the summarized results of the forecasted values of the models.

In this section, we have plotted the graph for the actual CO_2 emissions for the entire dataset and predicted the values from 2010 to 2019 by applying the models as mentioned earlier. Furthermore, we have forecasted the values from 2020 to 2030.

ARIMA's forecasting

From Fig. 4, it can be observed that ARIMA's predicted values are too far from the actual CO_2 emissions. Our experiment tested the dataset we have for prediction with an increasing trend only, which means non-stationay in nature, also tested by ADF test (Ajewole et al. 2020). ARIMA model work on stationary data, so we have to convert data into stationary one. In the conversion process, we shifted the data and differentiated it, which caused the loss of information. This reason makes the model to give an inaccurate result. Here, in our experiment, we used Arima (1,2,1) for prediction. From Table 2, it can be seen that the ARIMA model has a *MAPE* value of 98.969%, which is inaccurate. Therefore, we can conclude that forecasting for 2020 to 2030 will not be appropriate.

SARIMAX forecasting

SARIMAX forecasting is almost in the same pattern as the actual CO_2 emission can be seen from Fig. 5. It can be used as one of the appropriate models for forecasting the CO_2 emission. This better performance because it can capture the information from its seasonality. It also takes care

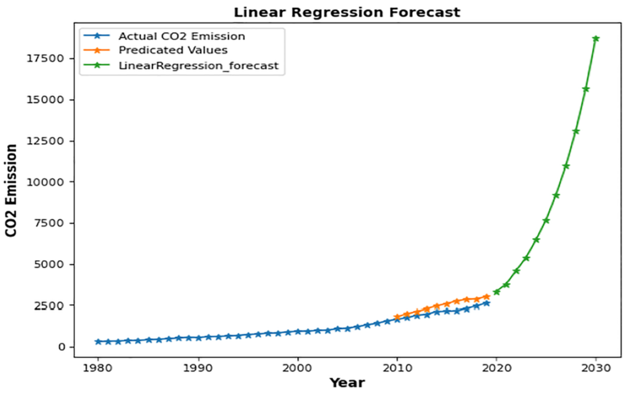

Fig. 6 Linear regression forecasting vs. actual CO_2 emissions

of exogenous factors in model training and makes a good prediction.

Also, we can see from Table 2, the *MAPE* value is the third smallest value at 6.554%, which clearly makes SARIMAX a better model.

Linear regression forecasting

Figure 6 shows that CO_2 predicted values by linear regression and actual emission, which is quite similar. We know that the linear regression model determines a mathematical equation under the black box while training the data with train data, so the linear relation assumes the output used as a predictor in the testing phase and then in forecasting.

But because of performance metric values of linear regression on test data, it is hard to rely on forecasting. As we can see from Table 2, the *MAPE* value is the fourth smallest value with 12.023%, which clearly makes the Linear model a good model, but not the best.

Random forest forecasting (RF)

A random forest is an ensemble machine learning technique that constructs multiple trees while training data and gives class labels for classification problems or mean/average prediction for regression. It can also be used in both univariate and multivariate time series forecasts by manually creating lag and seasonal component variables. According to the nature of data, different algorithms react differently, so we have to try it over our dataset. We tried our dataset into a random forest model in our work, which did not perform well.

Figure 7 shows the CO_2 predicted values, which is quite different from actual emissions. Also, the *MAPE* value is also maximum with 137.239%, which gives the worst prediction. Therefore, we can conclude that the random forest model is not appropriate for CO_2 emission forecasting.

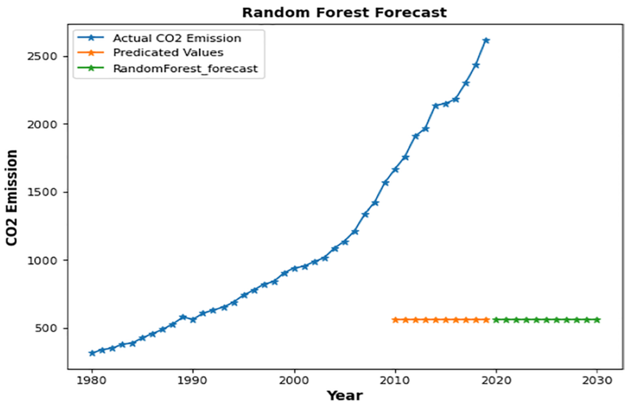

Fig. 7 Random forest forecasting vs actual CO_2 emission

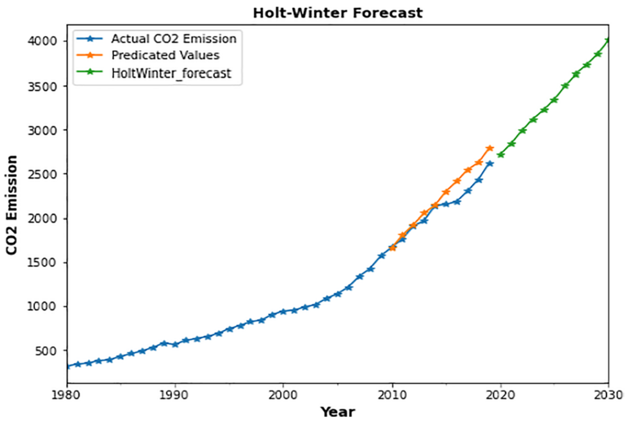

Fig. 8 Holt-Winters forecasting vs. actual CO_2 emissions

Holt-Winters model forecasting

We can see from Fig. 8 that Holt-Winters-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 performances metric values. The Holt-Winters model is used for exponential smoothing using "additive" or "multiplicative" models with increasing or decreasing trends and seasonality to make short-term forecasts. In our work, we have a multiplicative variant of the model, which took care of all three aspects of time series data i.e average, trend, and seasonality, and gave satisfying results. Also, from Table 2, the *MAPE* value is the second smallest value at 5.043%, which clearly makes Holt-Winters a suitable prediction model.

LSTM model forecasting

Figure 9 shows the LSTM behavior for prediction and forecasting of the CO_2 emissions. The LSTM is a model that uses the recurrent neural network concept but has a short

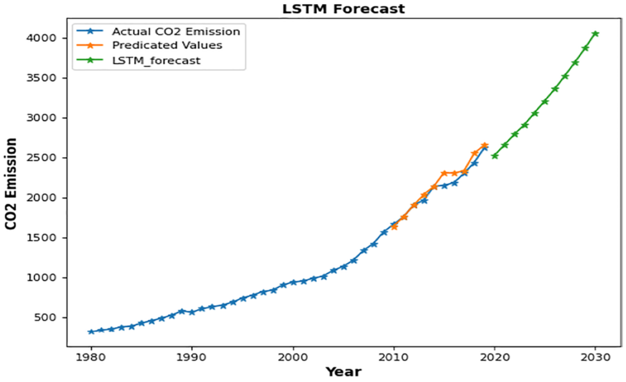

Fig. 9 LSTM forecasting vs. actual CO_2 emissions

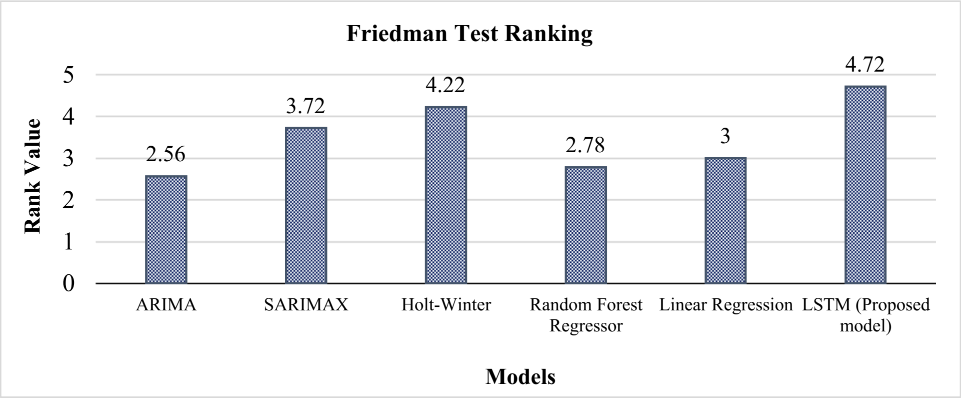

Fig. 10 Comparative rank of the models

memory element in the form of a series of logic gates to train the model and make the predictions. This model also destroys the problem of exploding and vanishing gradient. This is the reason this model is performing well.

Observations from Fig. 9 show that when we compare the actual CO_2 emissions and predicted value (as indicated from 2010 to 2020) are almost identical. We have already analysed from Fig. 3 that LSTM is the best performing model in terms of performance metric values. As we can see from Table 2, the LSTM model has the lowest value of 3676.646 and 60.635 for *MSE* and *RMSE*, respectively, and for these metrics, smaller values are desirable. Hence, this is the best forecasting model. *MSLE* metric values for the LSTM model are also lowest with 0.001 than in other models.

The *MAE* value for the LSTM model is 45.524, the second smallest among all models, which seems very appropriate for forecasting. *MAPE* is again the smallest value with 3.101%, which makes LSTM an excellent model. *MedAE* and MaxError have the lowest value with 28.898 and 135.933, respectively, and obviously, the LSTM outperformed here. The last R^2 score error is 0.990, most desirable, and EVS error is also 0.990, tied up with the SARIMAX model. Therefore, we can conclude that LSTM is the best-performing model to forecast the CO_2 emission.

Statistical analysis

In this section, we have applied the Friedman test (García et al. 2010) to verify the best-performing model statistically.

Observations from Fig. 10 show that LSTM is the best performing model with the highest rank value of 4.72. Therefore, the LSTM model can be used to forecast the CO_2 emission effectively.

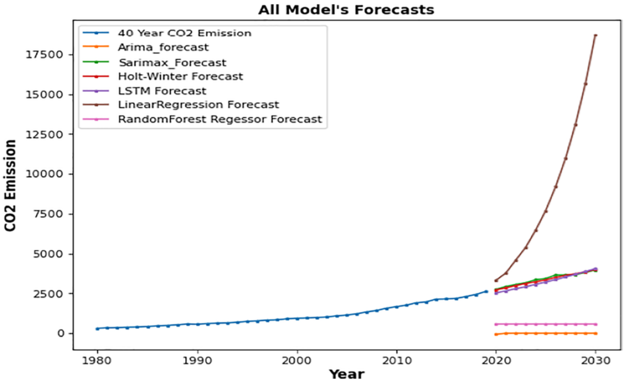

Fig. 11 Comparative forecasting of CO_2 emissions from the year 2020 to 2030

Comparative forecasting

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

Observations from Fig. 11 show that LSTM, SARIMAX, and Holt-Winters are appropriate forecasting models. Furthermore, when we look at the performance metric

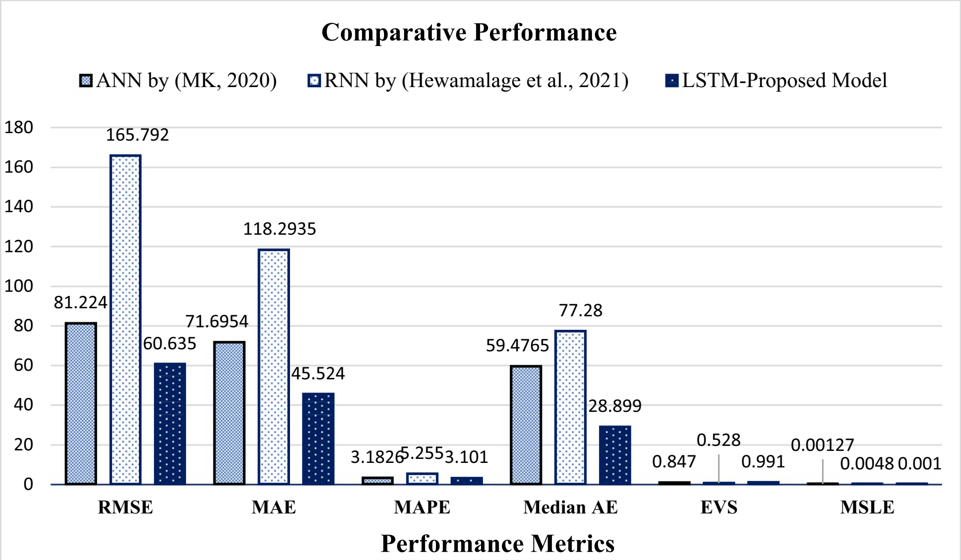

Fig. 12 Comparative analysis of the proposed model

values, then the LSTM is the best performing model for forecasting the CO_2 emissions.

A comparative study with recent work

In this section, we have analyzed the comparative performance of the proposed model with the recent works (MK 2020) and (Hewamalage et al. 2021). The work by (MK 2020) proposed the ANN model as a best forecasting model for CO_2 emission along with economic growth. Similarly, (Hewamalage et al. 2021) proposed that RNN (Recurrent Neural Network) as the best forecasting model over the competitive M4 dataset.

Now we have applied both the ANN and RNN model to our univariate CO_2 emission dataset, it is observed that ANN is performing better than RNN but both are not better than our proposed model LSTM.

From Fig. 12, it can be observed that the proposed LSTM model is the best-performing model for univariate CO_2 emission forecasting w.r.t. the performance metrics *RMSE*, *MAE*, *MAPE*, *MedianAE*, *EVS*, and *MSLE*. Therefore, we can conclude that the proposed LSTM model is one the best model for CO_2 emission forecasting.

Conclusion and future research directions

The accurate prediction of CO_2 emission in India in the coming decades is crucial for people's lives and the government. As per the available data, India is in the second position in the largest CO_2 emitting countries. The contribution of our work is to control the CO_2 emission to save human lives. In this work, we have applied the statistical, machine learning, and deep learning-based time series model to observe the CO_2 emission pattern. The performance of these models is evaluated based on nine appropriate performance metrics to choose a suitable one for future forecasting.

Findings indicate that out of applied models LSTM, SARIMAX, and Holt-Winters are better performing models for effective CO_2 emission prediction. To analyze the performance of the applied models, we have used the 40 years of CO_2 emission data. The result performance metrics-based observations conclude that LSTM is the best performing model with a 3.101% *MAPE* value, 60.635 *RMSE* value, and 28.898 *MedAE* values. Only the *MAE* value predicts SARIMAX as best. And EVS error value is tied between both LSTM and SARIMAX. We have also performed the Friedman test to show the statistical significance of the proposed model. In which, we have observed that LSTM has the highest Friedman ranking value 0f 4.72, indicates that it is one of the most suitable models for CO_2 emission forecasting.

Furthermore, a comparative study done with the recent works using the performance metrics *RMSE*, *MAE*, *MAPE* *Median AE*, *EVS*, and *MSLE* to conclude the suitability of the proposed model. The comparative results conclude that LSTM is the most suitable model along with the lowest errors in *RMSE*, *MAE*, *MAPE*, *Media AE*, and *MSLE*. Therefore, the LSTM model is proposed to be the best model and is used to forecast the next 10 years (2020 to 2030) years CO_2 emissions.

According to the LSTM model, CO_2 emission will be approximately five thousand metric tons which is two times of current CO_2 emission if it will keep going at the same pace, so 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 CO_2 emission by 58% by 2030. As a result, it is strongly advised that governments implement various policies, rules, norms, limits, and legislations to minimise CO_2 emissions and fossil-fuel usage in the various sectors.

- A few suggestions for policy-making could be increasing taxes on ecologically damaging uses.
- Imposing carbon taxes, cap-and-trade systems, carbon offsets, carbon caps, and eco-friendly technology standards.
- To educate people about the issues that create pollution and to raise community consciousness in society.
- Introducing free bus policies and promoting the electric vehicles would reduce the national fuel usage and the country's carbon impact.
- India should use low carbon-emitting sources like carbon-free hydrogen and sustainable biofuels for its power supply and reduce its dependence on coal, as it is an emerging renewable energy leader.
- Voluntary approaches should be adopted as a tool to reduce industrial emissions.

This work is limited only to univariate data prediction for CO_2 emission. We have not inculcated many factors such as population growth, economic growth, advancing technology, switching to renewable energy sources, and future government actions. These are a few 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. After including the factors above, we can find the correlation among them, and multivariate analysis can be done for effective CO_2 emission prediction.

Author contribution Surbhi Kumari: introduction, literature review, and performed the experiments. Sunil Kumar Singh conceptualised and designed the experiments, and analysed and interpreted the results.

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

Declarations

Ethics approval Not applicable.

Consent to participate Not applicable.

Consent for publication Permitted for publication as per journal guidelines.

Competing interests The authors declare no competing interests.

References

Abdullah L, Pauzi HM (2015) Methods in forecasting carbon dioxide emissions: a decade review. Jurnal Teknologi 75(1):67–82

Ağbulut Ü (2022) Forecasting of transportation-related energy demand and CO2 emissions in Turkey with different machine learning algorithms. Sustain Prod Consump 29:141–157

Ağbulut Ü, Gürel AE, Ergün A, Ceylan İ (2020) Performance assessment of a V-Trough photovoltaic system and prediction of power output with different machine learning algorithms. J Clean Prod 268:122269

Ağbulut Ü, Gürel AE, Biçen Y (2021a) Prediction of daily global solar radiation using different machine learning algorithms: evaluation and comparison. Renew Sust Energ Rev 135:110114

Ağbulut Ü, Gürel AE, Sarıdemir S (2021b) Experimental investigation and prediction of performance and emission responses of a CI engine fuelled with different metal-oxide based nanoparticles–diesel blends using different machine learning algorithms. Energy 215:119076

Ahmadi P (2019) Environmental impacts and behavioral drivers of deep decarbonization for transportation through electric vehicles. J Clean Prod 225:1209–1219

Ahmed NK, Atiya AF, El Gayar N, El-Shishiny H (2010) An empirical comparison of machine learning models for time series forecasting. Econ Rev 29(5):594–621. https://doi.org/10.1080/07474938.2010.481556

Ajewole KP, Adejuwon SO, Jemilohun VG (2020) Test for stationarity on inflation rates in Nigeria using augmented dickey fuller test and Phillips-persons test. J Undergrad Math 16:11–14

Amarpuri L, Yadav N, Kumar G, Agrawal S (2019) Prediction of CO2 emissions using deep learning hybrid approach: a case study in indian context. In: 2019 twelfth international conference on contemporary computing (IC3). IEEE, pp 1–6

Bakay MS, Ağbulut Ü (2021) Electricity production based forecasting of greenhouse gas emissions in Turkey with deep learning, support vector machine and artificial neural network algorithms. J Clean Prod 285:125324

Baum C (2018) KPSS: Stata module to compute Kwiatkowski-Phillips-Schmidt-Shin test for stationarity

Bonga WG, 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. J R Stat Soc: Ser C: Appl Stat 27(3):264–279

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

Dos Santos PR, De Souza LB, Lélis SP, Ribeiro HB, Borges FA, Silva RR, ... Rodrigues JJ (2021) Prediction of COVID-19 using time-sliding window: the case of Piauí state-Brazil. In: 2020 IEEE international conference on e-health networking, application & services (HEALTHCOM). IEEE, pp 1–6

Elsworth S, Güttel S (2020) Time series forecasting using lSTM networks: a symbolic approach. arXiv preprint arXiv: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

García S, Fernández A, Luengo J, Herrera F (2010) Advanced nonparametric tests for multiple comparisons in the design of experiments in computational intelligence and data mining: experimental analysis of power. Inf Sci 180(10):2044–2064

Good R, Fletcher HJ (1981) Reporting explained variance. J Res Sci Teach 18(1):1–7

Gopu P, Panda RR, Nagwani NK (2021) Time series analysis using ARIMA model for air pollution prediction in Hyderabad city of India. In: In Soft Computing and Signal Processing. Springer, Singapore, pp 47–56

Hewamalage H, Bergmeir C, Bandara K (2021) Recurrent neural networks for time series forecasting: current status and future directions. Int J Forecast 37(1):388–427

Huang C-H, Hsieh S-H (2020) Predicting BIM labor cost with random forest and simple linear regression. Autom Constr 118:103280

Kumar R, Kumar P, Kumar Y (2020) Time series data prediction using IoT and machine learning technique. Procedia Comp Sci 167(2019):373–381. https://doi.org/10.1016/j.procs.2020.03.240

Kumar S, Mishra S, Singh SK (2021) Deep Transfer Learning-based COVID-19 prediction using Chest X-rays. J Health Manag 23(4):730–746

The Lancet (2016) Air pollution—crossing borders. Lancet 388:103. https://doi.org/10.1016/S0140-6736(16)31019-4

Lepore A, dos Reis MS, Palumbo B, Rendall R, Capezza C (2017) A comparison of advanced regression techniques for predicting ship CO2 emissions. Qual Reliab Eng Int 33(6):1281–1292

Liu Z, Li D, Zhang J, Saleem M, Zhang Y, Ma R, He Y, Yang J, Xiang H, Wei H (2020) Effect of simulated acid rain on soil CO2, CH4 and N2O emissions and microbial communities in an agricultural soil. Geoderma 366:114222

Magazzino C (2017) Economic growth, CO2 emissions and energy use in the South Caucasus and Turkey: a PVAR analyses. Int Energy J 16(4)

Magazzino C, Mele M, Schneider N (2020) A Machine Learning approach on the relationship among solar and wind energy production, coal consumption, GDP, and CO2 emissions. Renew Energy 151:829–836

Magazzino C, Mele M, Morelli G, Schneider N (2021) The nexus between information technology and environmental pollution: application of a new machine learning algorithm to OECD countries. Util Policy 72:101256

Masini RP, Medeiros MC, Mendes EF (2021) Machine learning advances for time series forecasting. J Econ Surv. https://doi.org/10.1111/joes.12429

Mele M, Magazzino C (2020) A machine learning analysis of the relationship among iron and steel industries, air pollution, and economic growth in China. J Clean Prod 277:123293

Mele M, Magazzino C (2021) Pollution, economic growth, and COVID-19 deaths in India: a machine learning evidence. Environ Sci Pollut Res 28(3):2669–2677

MK AN (2020) Role of energy use in the prediction of CO2 emissions and economic growth in India: evidence from artificial neural networks (ANN). Environ Sci Pollut Res 27(19):23631–23642

Nontapa C, Kesamoon C, Kaewhawong N, Intrapaiboon P (2020) A new time series forecasting using decomposition method with SARIMAX model. In: International Conference on Neural Information Processing. Springer, Cham, pp 743–751

Nyoni T, Bonga WG (2019) Prediction of CO2 emissions in india using arima models. DRJ-J Econ Finance 4(2):1–10

Özmen ES (2021) Time series performance and limitations with SARIMAX: an application with retail store data. Electron Turk Stud 16(5)

Pino-Mejías R, Pérez-Fargallo A, Rubio-Bellido C, Pulido-Arcas JA (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. J Oper Res Soc 72(3):701–713

Shaikh S, Gala J, Jain A, Advani S, Jaidhara S, Edinburgh MR (2021) Analysis and prediction of covid-19 using regression models and time series forecasting. In: 2021 11th international conference on cloud computing, data science & engineering (Confluence). IEEE, pp 989–995

Solgi E, Keramaty M (2016) Assessment of Health Risks of urban soils contaminated by heavy metals (Bojnourd city). J North Khorasan Univ Med Sci 7(4):813–827

United Nations, Department of Economic and Social Affairs, Population Division (2019) World Population Prospects 2019: Data Booklet (ST/ESA/SER. A/424)

Wang Q, Li S, Pisarenko Z (2020) Modeling carbon emission trajectory of China, US and India. J Clean Prod 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. Forecast Forecast Eval Dyn Res J 4(2):1–10

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

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

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.