JETIR.ORG ISSN: 2349-5162 | ESTD Year : 2014 | Monthly Issue
JOURNAL OF EMERGING TECHNOLOGIES AND INNOVATIVE RESEARCH (JETIR)
An International Scholarly Open Access, Peer-reviewed, Refereed Journal

PREDICTION OF AIR POLLUTION USING DEEP LEARNING MODEL

Name- Darsh Pandya
College- K.J Somaiya College of Engineering
Name- Harsh Shah
College- K.J Somaiya College of Engineering

ABSTRACT

People all across the world are aware of the disadvantages of air pollution. Polluted air can trigger throat discomfort, itchy eyes and nose, and other significant problems in people who are hypersensitive. The number of people who have died as a result of polluted air has increased dramatically in recent years. Sensors are employed in this article to monitor air pollutants such as carbon monoxide, methane or natural gas, LPG, and air quality in various regions of the city. Furthermore, the values discovered are then used to produce forecasts regarding future values. Deep neural networks and the Internet of Things have made it possible to detect and estimate the amount of pollution in the air

INTRODUCTION

Heart attacks, chronic bronchitis, lung cancer, significant respiratory diseases, and asthma are all significantly exacerbated by air pollution. The prediction of air pollution is necessary due to the worsening of human health caused by an increase in air pollutants, particularly CO and methane.

The amount of toxins in the air is influenced by a variety of climatic variables, including wind direction, temperature, atmospheric pressure, and moisture. Furthermore, using historical data, the level of pollution can be predicted. Deep learning neural networks are skilled at extracting features from spontaneous data and learning on their own.

This neural network feature can be used to transform time series prediction problems in which models can be created from raw data without directly changing the data using standardisation, normalisation, or differencing to make the data stationary. We presented a model in this research that can anticipate contaminants in the air.

According to WHO data, approximately 42 lakh people died in India as a result of the effects of dirty air. Outside air contamination is blamed for approximately 3.8 million unexpected deaths each year. At least 130 million individuals breathe air that is multiple times as polluted as WHO standards. Every year, India has 12 of the 24 cities with the most polluted air in the value. In India, air pollution has resulted in the unanticipated death of 2 million people.

Several modelling techniques were utilised for this goal. Artificial neural networks are one of the most widely used machine learning-based approaches. Deep learning falls under the umbrella of machine learning. It advances artificial neural networks by leveraging large data sets, solving issues without dividing them, employing additional layers, processing concurrently with sequential levels, and producing more accurate results. Deep learning has a lot of advantages for modelling air pollution. To create a good model, several procedures must be followed. This paper describes in detail modelling with deep learning architecture real-world air pollution data. This was accomplished by examining how this field can get sick in a variety of ways, including through your heart and lungs. As a result, it is critical to be able to predict PM2.5 concentrations with high precision so that people can avoid the negative impacts of air pollution. PM2.5 is caused by a variety of contaminants in urban areas, as well as weather. We created a spatial-temporal feature and a CNN-LSTM model to predict the hourly PM2.5 air in Beijing, China. This was accomplished by merging historical pollution data, weather data, and PM2.5 concentrations from adjacent sites. has lately changed and how it differs from other AI models. In addition, the step-by-step procedure of creating the deep learning model was demonstrated, and the outcomes of various research were provided so that they could be compared.

LITERATURE REVIEW

Tao, Q., Liu, F et al,(2019) Air pollution forecasting can provide us with reliable information about how polluted the air will be in the future. This improves air pollution control and allows us to plan for how to stop it. Many factors influence air pollution, including temperature, humidity, wind direction, wind speed, snowfall, rain, and so on, making it difficult to understand how the concentration of air pollutants changes.

Liao, Q., Zhu, M. et al,(2020) Air pollution is one of the most important environmental problems of the 21st century. This is because the world is becoming more industrialised and urbanised. To stop it, we need accurate forecasts of the air quality. But current state-of-the-art air quality forecasts are limited by highly uncertain chemistry-transport models (CTMs), shallow statistical methods, and

Abdellatif, B., Badr, H. et al,(2021) Due to human activities, industrialization, and urbanisation, air pollution has become a threat to life in many places around the world over the past few decades. Particulate Matter with a diameter of less than 2.5m (PM2.5) is one of the air pollutants that is very bad for your health. it

Heydari, A., Majidi Nezhad, M. et al,(2022) Air pollution monitoring is constantly improving, and the effects on people's health are receiving increased attention. Because nitrogen dioxide (NO2) and sulphur dioxide (SO2) are two of the most dangerous pollutants, numerous models have been developed to predict how they may harm the environment. Still, making accurate predictions is nearly impossible. The long short-term memory model is used as a forecaster engine in the proposed model to predict the amount of NO2 and SO2 produced by the Combined Cycle Power Plant. To reduce forecasting error, the MVO algorithm is used to optimise the LSTM parameters.

Xayasouk, T., & Lee, H. (2018) Deep learning techniques are being used by an increasing number of people every day. Deep learning provides quick and accurate results, especially when large amounts of data are analysed. We presented a deep-learning-based method for predicting fine dust in this paper. We also used the deep-learning algorithm to create a spatiotemporal prediction framework, which incorporates the dataset's temporal and spatial relationships into the modelling process. We demonstrated how well and how accurately our deep learning model could predict.

Dairi, A., Harrou, F. et al,(2021) Air pollution, which causes many chronic diseases and premature deaths, has become a constant threat to people's health all over the world in recent years. Poor air quality is not only harmful to people's health and the environment, but it also has a significant negative impact on politics, society, and the economy. As a result, more effort should be put into making accurate predictions of ambient air pollution in order to develop practical and useful solutions, improve air quality, and plan for prevention. In this paper, we propose a flexible and effective deep learning-based model for predicting pollutant levels in the air.

Jujjavarapu, G. , Duggirala, S. et al,(2020) Air pollution is a big problem that people all over the world are aware of. Some of the bad effects of polluted air are allergic reactions like sore throats, itchy eyes and noses, and other serious problems. In recent years, the number of people who died because of dirty air has gone up by a huge pollution. In this paper, sensors are used to measure air pollutants like Carbon Monoxide, Methane, or natural gas, LPG, and the quality of the air in different parts of the city. Also, the values that are found are then used to predict values for the future.

Methodology

“The long short-term memory (LSTM) model, the spatio-temporal deep learning (STDL)-based air quality prediction method, deep air learning (DAL), and the convolutional neural network are all prediction algorithms that can be used to predict air pollution (CNN). The LSTM method is commonly employed for this purpose. LSTM models are a subset of recurrent neural networks (RNN), which use time series data to forecast the future and learn about pollution and weather. Instead of neurons, memory blocks are used in the RNN's hidden layer in the LSTM model. There are three gates in the LSTM block system: an input gate, a forget gate, and an output gate. These three gates regulate information transmission between the cell and the outside world. The LSTM block system is seen in Figure 1”.

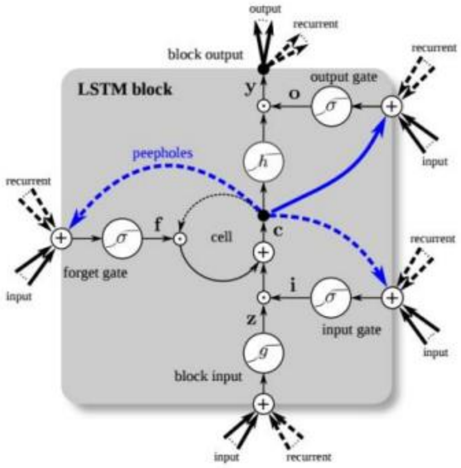


Figure 1. Block system of LSTM models.

The STDL method, which makes predictions utilised on both spatial and temporal variations, is the other method most commonly used in this situation. Stacking autoencoder models are used as an introduction model to remove things that are built into the air quality. A stacked autoencoder works on the basis that the output layer of the autoencoder in the layer below it is linked to the input layer of the layer above it.

Furthermore, spatiotemporal data is used to improve the performance of DAL models, which make predictions mostly by feature selection and semi-supervised learning. DAL is a powerful technique that integrates spatiotemporal semi-supervised learning and feature selection in the input and output layers.

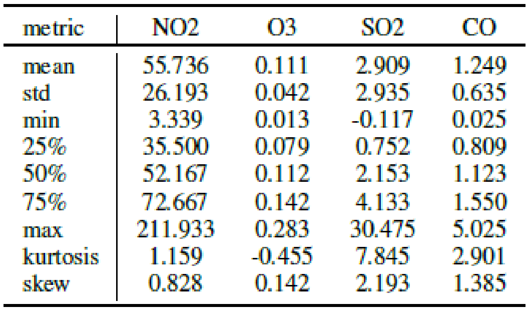
Experiments and Results

This section discusses the data used in this study and provides full details on how the experiment was set up and carried out. We also provide an analysis and discussion of the results.

Data description

The United States Environmental Protection Agency collected the dataset used in this study, which includes daily concentration readings of a number of ambient pollutants in various states. On the website "https://www.epa.gov/outdoor-air-qualitydata/download-daily-data," the datasets are open to the public. The four important pollutants that this study focuses on are ozone (O3), carbon monoxide (CO), sulphur dioxide (SO2), and nitrogen dioxide (NO2) (O3). We chose measurements from four different locations for our study: California, Arizona, Texas, and Pennsylvania. It should be noted that each state has numerous air quality monitoring stations; however, for measurement purposes, we chose one station from each state.Table I shows the numbers for the ambient air pollution datasets in Arizona. As with any perfectly symmetric distribution, the normal distribution has zero skewness and three kurtosis. Table I shows that the time-series datasets for ambient air pollution in Arizona are not Gaussian with positive support and have different variability intervals.

TABLE I: Statistics summary of the Arizona ambient air pollution datasets.



metric	NO2	O3	SO2	CO
mean	55.736	0.111	2.909	1.249
std	26.193	0.042	2.935	0.635
min	3.339	0.013	-0.117	0.025
25%	35.500	0.079	0.752	0.809
50%	52.167	0.112	2.153	1.123
75%	72.667	0.142	4.133	1.550
max	211.933	0.283	30.475	5.025
kurtosis	1.159	-0.455	7.845	2.901
skew	0.828	0.142	2.193	1.385

RESULTS AND DISCUSSION

The outcomes and a comprehensive explanation of the model are presented here. The whole experiment is scripted in Python. Pandas is a module for the Python programming language that enables the use of multidimensional arrays, which can then be used for data analysis. A number of Python libraries, including NumPy, Pandas, Matplotlib, Keras, TensorFlow, and seaborn, were included in the accompanying Jupyter notebook, which served as a simulation environment. Multiple specialised performance matrices were applied (described below). The utilisation of a dataset, which will be elaborated upon below, allowed for the findings of this study.

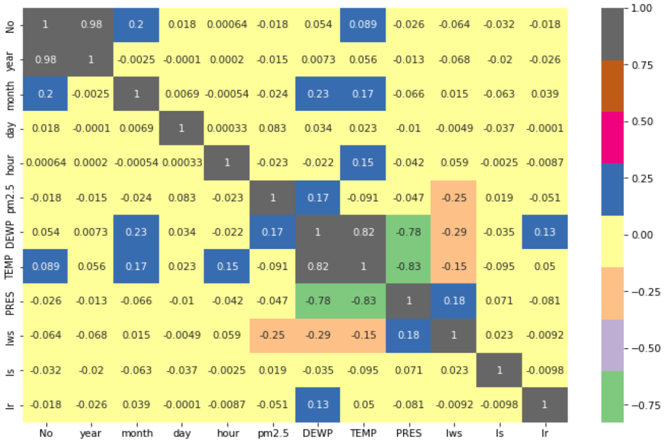


Figure 2: Correlation matrix of pollution dataset

In the figure above, number 2, we can see the correlation matrix for the air pollution dataset. Broadly put, The level of association between various pieces of data can be displayed in a table format called a correlation matrix. All of the potential permutations of data in a table are represented by the matrix. It's a great method for quickly comprehending the key points of a massive dataset as well as discovering interesting trends or patterns in the information at hand.

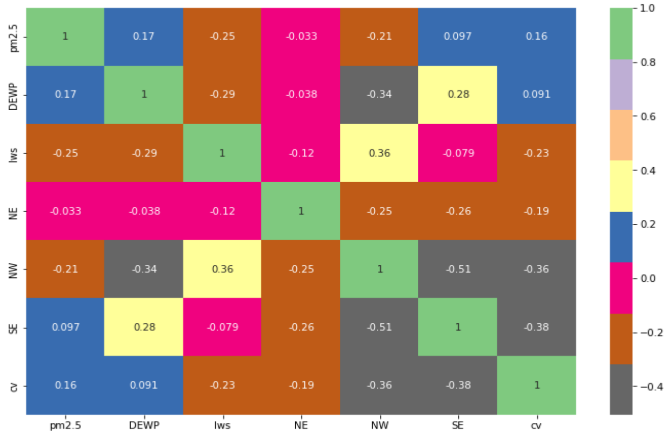


Figure 3: Heatmap of pollution dataset

The pollution dataset heatmap is shown in Figure 3. The heatmap is a graphical tool for displaying visitor activity as warm and cool areas. The warm colours indicate areas with the most visitor contact, with red being the region with the greatest interaction, while the cold colours denote areas with the lowest interaction.

Performance Evaluation

Root mean squared error (RMSE) and mean squared error (MSE) are the two most popular metrics, and they are used to measure how far off the prediction result is from the actual value.

1) Mean Absolute Error

This article introduces the mean absolute error as a statistical measure of the consistency between two independent descriptions of the same phenomenon. Y against X may be used to compare predicted values to actual values, end times to beginning times, or one measuring technique to another.

$$\mathrm{MAE}(y, \hat{y}) = \frac{1}{n} \sum_{i=0}^{n-1} |y_i - \hat{y}_i| \quad ..(2)$$

2) RMSE (Root Mean Square Error)

The The root-mean-squared error (RMSE) is another popular metric used in errors in estimating. It is used to determine how far off an estimate is from the true number. One alternative name for this method of error measurement is root-mean-square error. It evaluates how much of an error there really is. To compare how well different estimators, predict a given variable, a ratio of their relative errors is calculated. That is to say, it's the gold standard of precision.

$$\mathrm{RMSE}\left(\hat{\theta}\right) = \sqrt{\mathrm{MSE}\left(\hat{\theta}\right)} \quad ....(3)$$

Experimented Results

For this research, we have implemented several widely-used deep learning classification algorithms, including CNN and LSTM. And in contrast to many other methods that are being used at the present time. A number of graphs, metrics, and tables are used to present the findings that were obtained from the experiment. Since the experiment was carried out, we have conducted a thorough investigation of the data gathered from it and drawn our conclusions. The scope of this work included the provision of a deep learning model for classification and feature selection.

Table II: The proposed model is compared to the baseline model using performance metrics.

Model	MAE	RMSE
Base (CBGRU)	10.4	14.5
Propose CNN	8.06	10.90
Propose LSTM	7.3	10.4

Table II displays the three performance parameters using the proposed and base models. The mean absolute error (MAE) of the baseline model is 10.4, and the root mean squared error is 14.5 % while the proposed LSTM mode 7.3 MAE, and RMSE is 10.4 and the second proposed CNN model get MSE 8.06 and RMSE is 10.90 respectively.

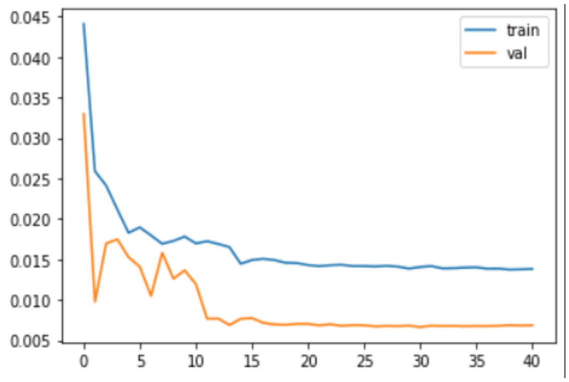


Figure 4: Plot graph of training and validation loss

The above figure 9 shows the loss curve of training and validation data. The number of training epochs is shown along the X-axis, and the level of loss that is attained by the model after each round of training is shown along the Y-axis. As can be seen in the diagram that is located previous paragraph, the ratio of loss in both training and validation at each time is represented by the blue and orange line, which indicates the loss of the training each time the model is being validated, as well as the constant line, which indicates the loss of the model during training..

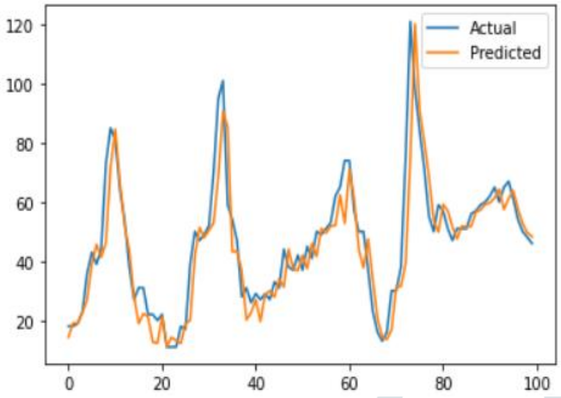


Figure 5: Plot graph of actual and predicted predication

The above figure 5 shows the plot graph of actual predication and predicated predication. The number of epochs used in this model is 0 to 100.

CONCLUSION

Air pollution is a worldwide issue, with concentrations in most places known to be harmful to health. Because of how quickly industrial technology has changed, several negative effects on the environment have occurred. It is critical to monitor the air quality to ensure that it is satisfactory. This study demonstrated a new deep hybrid model that incorporates an attention mechanism into the variational autoencoder to improve air pollution forecasting (dubbed IMDA-VAE). We demonstrated that the proposed technique works well for forecasting both univariate and multivariate time-series data on air pollution in this study. The proposed IMDA-VAE model predicted the concentrations of four important pollutants using VAE, Gated GRUs, LSTM, BiGRU, BiLSTM, ConvLSTM, LSTM-A, and GRU-A. (NO2, O3, SO2, and CO). Forecast accuracy has been tested using statistical measures such as R2, RMSE, MAE, MAPE, EV, MBE, and RMBE. Metrics revealed that the deep hybrid model performed well in modelling temporal dependencies in unsupervised learning without the need for sophisticated gating and memory processes found in recurrent models. In time-dependent modelling, the variational inference approximation showed well. Furthermore, univariate forecasts were more accurate than multivariate forecasts in this case. This is primarily due to the lack of a high link between the four pollutants studied.

REFERENCES

1. Tao, Q., Liu, F., Li, Y., & Sidorov, D. (2019). Air pollution forecasting using a deep learning model based on 1D convnets and bidirectional GRU. *IEEE access*, *7*, 76690-76698.
2. Liao, Q., Zhu, M., Wu, L., Pan, X., Tang, X., & Wang, Z. (2020). Deep learning for air quality forecasts: a review. Current Pollution Reports, 6(4), 399-409.

3. Abdellatif, B., Badr, H., Samira, D., & Khadija, D. (2021). Air-pollution prediction in smart city, deep learning approach. Journal of Big Data, 8(1).
4. Heydari, A., Majidi Nezhad, M., Astiaso Garcia, D., Keynia, F., & De Santoli, L. (2022). Air pollution forecasting application based on deep learning model and optimization algorithm. Clean Technologies and Environmental Policy, 24(2), 607-621.
5. Xayasouk, T., & Lee, H. (2018). Air pollution prediction system using deep learning. WIT Trans. Ecol. Environ, 230, 71-79.
6. Dairi, A., Harrou, F., Khadraoui, S., & Sun, Y. (2021). Integrated multiple directed attention-based deep learning for improved air pollution forecasting. IEEE Transactions on Instrumentation and Measurement, 70, 1-15.
7. Jujjavarapu, G. , Duggirala, S., Kavutarapu, A., Surapaneni, R. (2020) Ambient Air Pollution Forecasting System using Deep Neural Networks. International Journal of Recent Technology and Engineering (IJRTE) ISSN: 2277-3878, Volume-8 Issue-6.
8. M. Madhuri, G. H. Samyama Gunjal, and S. Kamalapurkar, "Air pollution prediction using machine learning supervised learning approach," Int. J. Sci. Technol. Res., 2020.