Published by :
http://www.ijert.org

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 11 Issue 10, October-2022

Estimation of Pollutant Levels using LSTM and SVR in New Delhi

Advait Yadav
Bhatnagar International School
New Delhi, India

Abstract **- Air pollution has rapidly risen in the past decade and has been a detriment to human health. Poor air quality is directly correlated to severe health issues, especially in the elderly and children. There is a need for accurate air quality prediction so that the government can put safety measures in place and prevent further degradation. In this paper, we use Long Short-Term Memory (LSTM) and Support Vector Regression (SVR) models to forecast the levels of SO_2, NO_2 and PM2.5. The models have been trained on archived pollution concentration data made publicly available by the Central Pollution Control Board. The model uses data exclusively from New Delhi, one the most polluted cities in the world, in order to achieve higher accuracy. The models have been evaluated using Root Mean Squared Error (RMSE), Mean Absolute Error (MAE) and R-Squared (R^2) which indicate LSTM to be best suited for the task of predicting pollutant levels in New Delhi.**

Keywords - recurrent neural networks; support vector regression; air quality prediction; machine learning; forecasting

I. INTRODUCTION

Air pollution has seen exponential growth in the past few years due to a variety of factors like industrial development, fossil fuel burning, agricultural activities and vehicular emissions. This degradation of air quality has significantly impacted human health and has been shown to have serious short- and long-term implications.

Air pollution has been shown to induce respiratory and cardiovascular diseases, neuropsychiatric complication, eye irritation, skin diseases and chronic conditions like lung cancer. It is also considered to be a major risk factor in progression of conditions like asthma, Alzheimer and Parkinson's diseases, autism, retinopathy, male infertility, ventricular hypertrophy, stunted fetal growth, and low birth weight [1]-[5]. Air pollution has also shown to be responsible for 1 in 10 deaths of children under 5 years of age.

New Delhi, the capital of India, is known to be one of the most polluted cities in the world. Various studies have been conducted to study the air quality levels in Delhi and have shown that the pollutant levels are much higher than the recommended limits. Studies have also seen a steady increase in mortality and morbidity rates over the past decade, which has been directly linked to the rise of pollutant levels [6]. Further development of the city has resulted in the rise of emission producing vehicles, which have been shown to damage human health [7]. A study noted that policemen who were exposed to high levels of pollutants were at higher risk of developing significant respiratory impairment and urged the use of masks to filter out pollutants [8].

These findings demonstrate the need for accurate forecasting of air quality, which will help the government in taking corrective measures and prevent further degradation. Accurate early detection of spikes in air quality can also help in shielding vulnerable sections of society like the elderly and children. This study focuses on building such a prediction model, which uses past data to accurately predict future pollutant levels. For building the model, we will be using LSTM (Long Short-Term Memory) and SVR (Support Vector Regression) and comparing the results between the two. These algorithms will be employed to predict values of 3 different pollutants – NO2, SO2 and PM2.5. The past data has been archived from 38 measuring stations throughout Delhi, and has been merged to create a larger dataset, which helps improve performance of the model.

The paper is structured as follows. We conduct critical revision of previous studies done in similar fields in section 2. In section 3 and 4, we examine how Long Short-Term Memory (LSTM) and Support Vector Regression (SVR) work respectively. Section 5 is used to describe the dataset and explore data preprocessing and normalization in order to produce better input for the model. In section 6 we present the results of the study, comparing the models using various error metrics. In section 7, we conclude the paper and suggest suitable changes and ideas for future work.

II. RELATED WORK

Various time series prediction models have been used to evaluate air quality and pollution particle levels using past data. Traditionally, statistical models have been the ideal choice for prediction. However, with increase in data, machine learning models have been shown to have higher accuracy.

A. Statistical Models

Statistical models use past data to learn and apply this experience to predict future values. The best statistical model for time series prediction is ARIMA (Autoregressive integrated moving average). ARIMA was applied to predict air pollution index and was shown to reach an accuracy of 95% [9].

Simple and multiple regression models were used to predict PM10 concentrations to achieve a percentage error of <30% [10]. Reference [11] compares the performance of ARIMA with an exponential smoothing model, proving ARIMA to be superior in predicting AQI values.

However, due to their inability to learn from dynamic parameters, statistical models are outperformed by machine learning models in most cases.

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 11 Issue 10, October-2022

B. Machine Learning Models

Machine learning models are nonlinear and use multiple parameters, which helps them in learning from complex parameters like air pollutant levels. ML models also benefit from large quantities of data, and can be trained with minimum human intervention. They have been shown to outperform traditional models like ARIMA or multivariate regression, which tend to perform well with simpler data [12].

ANN (Artificial Neural Network) models appear to be the most used for air quality forecasting [13]-[14]. Reference [15] uses ANN to predict PM10 levels in Kota. A back propagation neural network, modified using wavelet-transform technique, has been used to forecast air pollutants like PM10, SO2 and NO2 [16]. Reference [17] forecasts AQI in Delhi using a neural network based on principal component analysis.

Other machine learning models like SVR (Support Vector Regression), a variant of SVM (Support Vector Machines) [18] have been shown to have high accuracy in regression tasks. SVR is a popular choice in air pollution particle prediction [19]-[20]. Standard SVM and naïve bayes classification were applied to predict air quality in Beijing [21]. A hybrid of ANN and SVR was shown to have high accuracy in time series predictions, where ANN was used for partitioning the input space and SVR to model the portions [22]. Reference [23] compares machine learning algorithms for air pollution prediction, and concluded that SVR and Neural Networks (MLP) showcased the best accuracy.

III. LSTM ALGORITHM

LSTM is an advanced Recurrent Neural Network (RNN), a sequential network that allows information to stay for longer periods of time. RNN work by remembering the previous information and use it to process the current input.

However, they cannot remember long term dependencies due to vanishing gradient. LSTM are designed to overcome this by enabling the network to store information for thousands of steps, hence the name – "long short-term memory".

LSTM consists of 3 gates – Forget gate, Input gate and the Output gate, with each gate having an individual function. The first gate filters out the irrelevant information and decided on the information to be remembered. In the second gate, the network tries to learn new information from the input given to the cell. Finally, in the third gate, the cell passes on the updated information from the current step to the next one.

To explore these gates in more detail:

Forget Gate:

$$f_t = \sigma(x_t * U_f + H_{t-1} * W_f)$$

X_t: input to the current timestamp.
U_f: weight associated with the input
H_{t-1}: The hidden state of the previous timestamp
W_f: weight matrix associated with hidden state

Later, a sigmoid function converts f_t into a number between 0 and 1. This f_t is then multiplied with cell states of previous timestamps. If $f_t = 0$, the network will forget everything and if $f_t = 1$, the network will forget nothing.

Input Gate: The main job of the input gate is to quantify the new information carried by the input.

$$i_t = \sigma(x_t * U_i + H_{t-1} * W_i)$$

X_t: Input at the current timestamp t
U_i: weight matrix of input
H_{t-1}: A hidden state at the previous time stamp
W_i: Weight matrix of input associated with hidden state

Again, a sigmoid function is applied, resulting in a value between 0 and 1.

Output Gate:

$$o_t = \sigma(x_t * U_o + H_{t-1} * W_o)$$

Now, we use O_t and Tan (h) of the updated cell state to calculate the current hidden state. The hidden state is a function of the Long term memory (C_t) and the current output.

To find the output of the current timestamp, SoftMax activation is applied on the hidden state H_t. The token with the maximum score is the output.

IV. SVR ALGORITHM

Support Vector Regression is a type of Support Vector Machine that supports both linear and non-linear regression, and is used to predict discrete values. The main objective of the algorithm is to find a best fit line, which is the hyperplane that has the maximum number of points on it.

Unlike standard regression models which aim to minimize error between the predicted and actual values, SVR tries to fit the best fit line within a threshold value. The threshold value is the distance between the hyperplane and the boundary line, which are parameters in SVR. These parameters are known as hyperparameters are as follows.

1) Hyperplane
   Hyperplanes are decision boundaries used to predict continuous output. The data points closest to the hyperplane are known as support vectors, and are used to influence the orientation of the hyperplane.

2) Kernel
   A kernel is a set of mathematical functions that takes input and transforms it into the required form. They are used to find hyperplane in higher dimensional spaces. By default, Radial Basis Function (RBF) is used as the kernel [24]. The mathematical expression of the RBF kernel is as follows:

$$K(X_1, X_2) = \exp\left(-\frac{||X_1 - X_2||^2}{2\sigma^2}\right)$$

   Where σ is the variance and $||X_1 - X_2||$ is the Euclidean distance between the two points X_1 and X_2.

3) Boundary Lines
   Two lines are around the hyperplane that are used to create a margin between the data points. They are located at a distance of ε (epsilon) from the hyperplane.

V. DATASET

The dataset contains hourly and daily level of various air pollutants across various stations in multiple cities of India. The data has been made publicly available by the Central

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 11 Issue 10, October-2022

Pollution Control Board: https://cpcb.nic.in/ which is the official portal of the Government of India.

For the purpose of this study, we will be using the readings of SO_2, NO_2 and PM2.5 [25]. These pollutants have been chosen because of they showcased the largest amount of recorded values, which helps increase the accuracy of the machine learning models.

In order to further increase the volume of data, we will be using the hourly readings of data. We will be focusing on data from stations in Delhi only, in order to provide more accurate predictions. Delhi is considered to be one of the most polluted cities in the world, and sees a sharp drop in air quality levels in winter. Delhi also had the least number of missing values among the cities, which makes it ideal for our models.

A. Data Preprocessing

Data quality and effective representation are important in ensuring good performance of the model. In order to successfully train our model, we need to fill out the missing values in our dataset [26]. These missing values are usually caused due to real world errors, such as those in the recording devices, and must be treated to make sure that the algorithms are trained properly.

The missing values have been filled using Linear Interpolation. Linear Interpolation is an imputation technique that assumes a linear relationship between data points and utilizes non-missing values from adjacent data points to compute a value for a missing data point.

B. Data Normalization

When input consists of multiple attributes with vastly different values, it is important to scale the attributes to the same range to give equal weightage to all the features. This is done through normalization. The normalization method used is MinMax scaling, which scales and translates each feature individually such that it is in the given range (0, 1) on the training set.

$$X_{sc} = \frac{X - X_{min}}{X_{max} - X_{min}}$$

Where X_{sc} is the normalized value, X is the original value and X_{min} and X_{max} are the minimum and maximum values respectively.

VI. RESULTS

After building and developing the models with both LSTM and SVR, it is important to evaluate the models using various error parameters. This is done to gain feedback on the performance of the model and to make suitable changes to achieve the desirable results. For this study, we will use 3 parameters – Mean Absolute Error (MAE), Root Mean Squared Error (RMSE) and R-Squared (R^2). The 3 pollutants, SO_2, NO_2 and PM2.5, are evaluated separately, and their performance is shown in Table 1, Table 2 and Table 3 respectively.

TABLE I. ERROR METRICS FOR PREDICTION OF SO_2

Parameter	RMSE	MAE	R^2
LSTM	1.579	1.009	0.878
SVR	1.540	1.230	-0.494

TABLE II. ERROR METRICS FOR PREDICTION OF NO_2

Parameter	RMSE	MAE	R^2
LSTM	4.148	2.845	0.950
SVR	4.737	3.093	0.369

TABLE III. ERROR METRICS FOR PREDICTION OF PM2.5

Parameter	RMSE	MAE	R^2
LSTM	3.495	2.626	0.973
SVR	4.508	4.516	0.662

The desired results are achieved, as can be seen from the error evaluations. In Table 1, LSTM fairs better than SVR in SO_2 prediction but both algorithms provide accurate results. In Table 2, higher error rates are observed for both the models. This can be attributed to a significant amount of missing input which has to be treated during preprocessing. In Table 3, the trend of higher error rates is observed as well, owing to lack of data. Considering these results, LSTM best serves our purpose, showcasing lower error in majority of the readings. Both models show higher accuracy when the prediction time period is shorter.

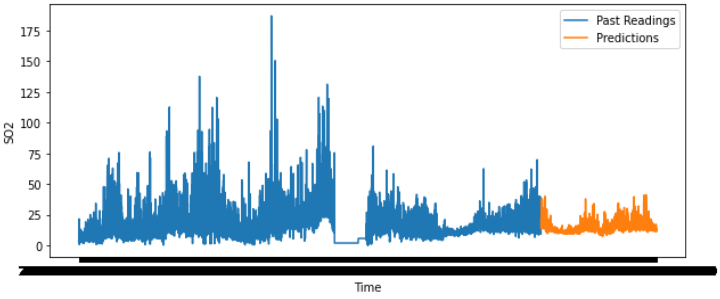

Fig. 1 Prediction of SO_2 levels in New Delhi.

The absence of values in a part of the graph can be attributed to measuring errors. Such values are treated during data preprocessing before the model is trained on the data.

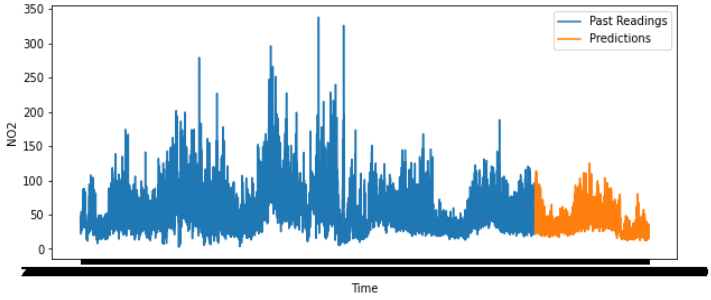

Fig. 2 Prediction of NO_2 levels in New Delhi

A downward trend is observed in both SO_2 and NO_2 levels. This can be attributed to the COVID-19 pandemic lockdowns, which caused reduction of emissions.

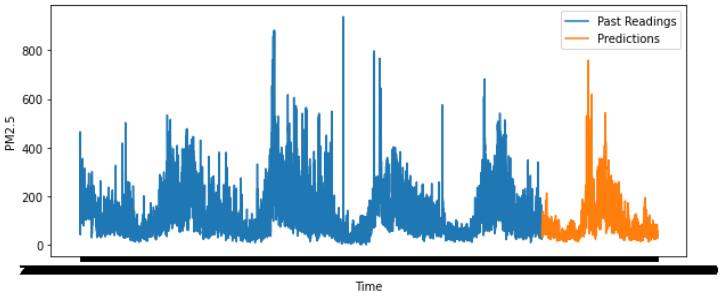

Fig. 3 Prediction of PM2.5 levels in New Delhi

PM2.5 levels are seen spiking each year in winter due to various agricultural and geographical reasons.

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 11 Issue 10, October-2022

VII. CONCLUSION

The task of forecasting pollutant levels is challenging and requires accurate long term data to provide specific results. This is inherently hard due to the volatility of the data. However, the use of suitable machine learning models can help in improving results significantly. In this study, we forecasted the values of SO_2, NO_2 and PM2.5 in New Delhi using publicly available data.

The findings above can help in accurate forecasting of air pollution in the future, and can help authorities in taking appropriate safety and precautionary measures and providing information to the general public about future pollution trends.

In the future, we would like to create models that use meteorological parameters, and to combine machine learning algorithms like neural networks and support vector machines to increase quality of the predictions. We would also like to use to larger datasets, which generally help in generalization of neural networks.

REFERENCES

[1] Zhou N, Cui Z, Yang S, Han X, Chen G, Zhou Z, et al. Air pollution and decreased semen quality: A comparative study of Chongqing urban and rural areas. *Environ Pollut.* 2014;187:145–52.

[2] Vermaelen K, Brusselle G. Exposing a deadly alliance: Novel insights into the biological links between COPD and lung cancer. *Pulm Pharmacol Ther.* 2013;26:544–54.

[3] Biggeri A, Bellini P, Terracini B. Meta-analysis of the Italian studies on short-term effects of air pollution – MISA 1996-2002. *Epidemiol Prev.* 2004;28(4-5 Suppl):4–100.

[4] Yamamoto SS, Phalkey R, Malik AA. A systematic review of air pollution as a risk factor for cardiovascular disease in South Asia: Limited evidence from India and Pakistan. *Int J Hyg Environ Health.* 2014;217:133–44.

[5] Rumana HS, Sharma RC, Beniwal V, Sharma AK. A retrospective approach to assess human health risks associated with growing air pollution in urbanized area of Thar Desert, Western Rajasthan, India. *J Environ Health Sci Eng.*

[6] Nagpure, B. Gurjar and J. Martel, "Human health risks in national capital territory of Delhi due to air pollution", *Atmos. Pollut. Res.*, vol. 5, no. 3, pp. 371-380, 2014

[7] P. Aggarwal and S. Jain, "Impact of air pollutants from surface transport sources on human health: A modeling and epidemiological approach", *Environ. Int.*, vol. 83, pp. 146-157, 2015

[8] Sopan T. INGLE, Bhushan G. PACHPANDE, Nilesh D. WAGH, Vijaybhai S. PATEL, Sanjay B. ATTARDE, Exposure to Vehicular Pollution and Respiratory Impairment of Traffic Policemen in Jalgaon City, India, Industrial Health, 2005, Volume 43, Issue 4, Pages 656-662, Released on J-STAGE March 17, 2006, Online ISSN 1880-8026, Print ISSN 0019-8366, https://doi.org/10.2486/indhealth.43.656

[9] L. Y. Siew, L. Y. Chin, and P. M. J. Wee, "Arima and integrated arfima models for forecasting air pollution index in shah alam, selangor," Malaysian Journal of Analytical Sciences, vol. 12, no. 1, pp. 257–263, 2008.

[10] Li, N. Hsu and S. Tsay, "A study on the potential applications of satellite data in air quality monitoring and forecasting", *Atmos. Environ.*, vol. 45, no. 22, pp. 3663-3675, 2011

[11] J. Zhu, R. Zhang, B. Fu, and R. Jin, "Comparison of arima model and exponential smoothing model on 2014 air quality index in yanqing county, beijing, china," Appl. Comput. Math, vol. 4, pp. 456–461, 2015.

[12] I. Alon, M. Qi, and R. J. Sadowski, "Forecasting aggregate retail sales: a comparison of artificial neural networks and traditional methods," Journal of retailing and consumer services, vol. 8, no. 3, pp. 147–156, 2001

[13] M. Baawain, "Systematic Approach for the Prediction of Ground-Level Air Pollution (around an Industrial Port) Using an Artificial Neural Network", Aerosol Air Qual. Res., 2014.

[14] M. Huang, T. Zhang, J. Wang and L. Zhu, "A new air quality forecasting model using data mining and artificial neural network", in 6th IEEE International Conference on Software A

[15] S. Saxena and A. Mathur, "Prediction of Respirable Particulate Matter (PM10) concentration using artificial neural network in Kota city", Asian Journal for Convergence in Technology, vol. 3, no. 3, 2018.

[16] Bai, Y. Li, X. Wang, J. Xie and C. Li, "Air pollutants concentrations forecasting using back propagation neural network based on wavelet decomposition with meteorological conditions", Atmos. Pollut. Res., vol. 7, no. 3, pp. 557-566, 2016

[17] Kumar and P. Goyal, "Forecasting of air quality index in Delhi using neural network based on principal component analysis", Pure Appl. Geophy., vol. 170, no. 4, pp. 711-722, 2012

[18] H. Drucker, C. J. Burges, L. Kaufman, A. Smola, V. Vapnik et al., "Support vector regression machines," Advances in neural information processing systems, vol. 9, pp. 155–161, 1997

[19] K. Hu, V. Sivaraman, H. Bhrugubanda, S. Kang and A. Rahman, "SVR based dense air pollution estimation model using static and wireless sensor network," IEEE SENS J , Orlando, FL, pp. 1-3, 2016

[20] S. Arampongsanuwat and P. Meesad, "Prediction of pm10 using support vector regression," in International Conference on Information and Electronics Engineering, IACSIT Press. Singapore, vol. 6, 2011.

[21] Dan wei: Predicting air pollution level in a specific city [2014]

[22] L. Cao, "Support vector machines experts for time series forecasting," Neurocomputing, vol. 51, pp. 321–339, 2003.

[23] Srivastava, Chavi & Singh, Shyamli & Singh, Amit. (2018). Estimation of Air Pollution in Delhi Using Machine Learning Techniques. 304-309. 10.1109/GUCON.2018.8675022.

[24] R. G. Brereton and G. R. Lloyd, "Support vector machines for classification and regression," Analyst, vol. 135, no. 2, pp. 230–267, 2010.

[25] Central Pollution Control Board, (Ministry of Environment, Forests & Climate Change), Govt of India, "National Air Quality Index", Central Pollution Control Board (CPCB), 2018

[26] S. B. Kotsiantis, D. Kanellopoulos, and P. E. Pintelas, "Data preprocessing for supervised leaning," International journal of computer science, vol. 1, no. 2, pp. 111–117, 2006.