International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 10 Issue 11, November-2021

To Predict Air Pollution using Machine Learning and Arima Model

Shakir Muhammad Abdullah
School of Computer Science
North China University of Water Resources and Electric Power, Zhengzhou, 450045, China

Abstract:
Background: Currently in many industrial and urban areas, air quality investigation and preservation has become one of the government's most essential responsibilities. The weather and traffic conditions, the combustion of fossil fuels, and the characteristics of industry all have a major influence in the air emissions. As a result of the rising levels of air pollution, Models that record information on air pollutant concentrations must be implemented. It is the testimony of these hazardous chemicals in the atmosphere that has an impact on the overall quality of people's lives, particularly in metropolitan areas. Due to the availability of environmental sensing networks and sensor data, many academics have begun to use Big Data Analytics techniques in recent years.

Materials and Methods: Predicting the concentration of SO2 in the environment is accomplished via the application of machine learning methods in this article. Moreover, Sulphur dioxide is unpleasantly influenced by skin and eyes, mucous membranes of the eyes, mouth, throat and lungs. For future years or months, time series models are used to predict quantities of SO2 in the atmosphere.

Results: The purpose of this study is to detect and forecast Zhengzhou's air pollution and quality, the province's capital in east-central China. It will be possible to forecast this using machine learning methods and the AIRMA model.
Keywords: Machine Learning, Air pollution, Prediction, Air Quality, Arima Model, Zhengzhou etc

1. INTRODUCTION

In developing nations such as China, the fast growth in population and economic boom in urban areas has resulted in environmental concerns such as air pollution and a lot of other issues. The health of people is directly impacted by air pollution. In Zhengzhou, there has been a rise in the general public's awareness of the issue. Air pollution has several long-term effects, some of which include global warming, acid rain, and a rise in the number of asthma sufferers. Air quality forecasting that uses predictive models has the potential to help people and the environment cope with periods of high pollution. The improvement of air quality projections is therefore one of society's most essential aims. Sulphur Dioxide is a gas that exists in the atmosphere [1]. It is one among the most significant pollutants found in the atmosphere. It is colorless and has a pungent, harsh odor about it. It readily interacts with other compounds to create dangerous molecules such as sulphuric acid and sulphureous acid, amongst others. When Sulphur dioxide is inhaled, it has a negative impact on human health. It will irritate the nose, throat, and airways producing a tight feel around the chest, coughing, wheezing, shortness of breath and. The quantity of Sulphur dioxide in the atmosphere may have an impact on the appropriateness of a habitat for plant communities as well as animal life, among other things.

The suggested system is capable of forecasting the concentration of Sulphur Dioxide in the atmosphere for the next month’s / years, according to the data.

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 10 Issue 11, November-2021

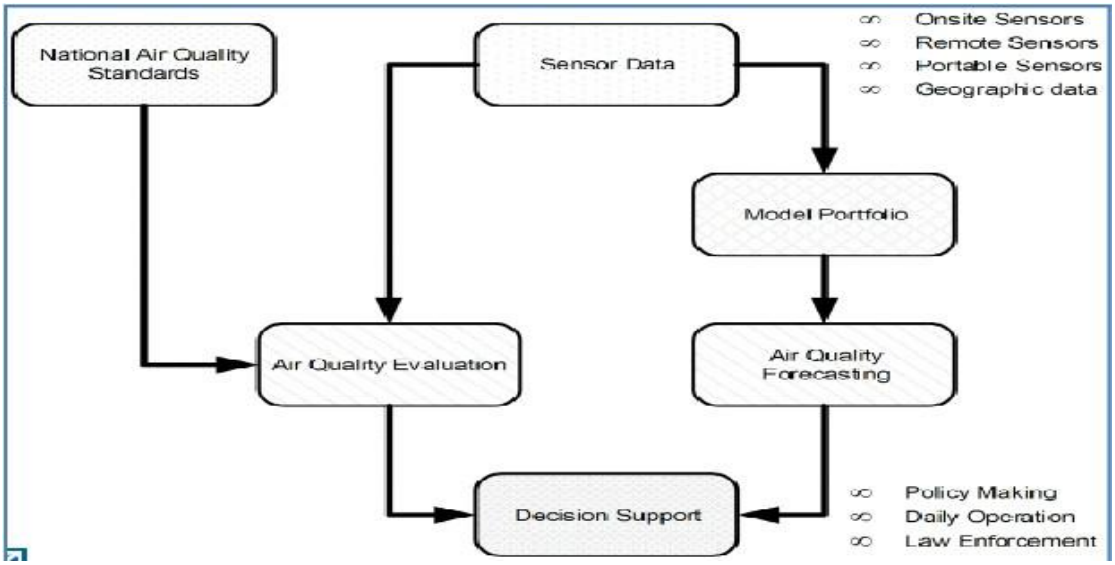

Figure 1 Big data-based decision support for air quality

1.1. Background of the research

Despite the country's improvements, air pollution levels are starting to drop relative to decades before. In 2015, the Chinese government presided over the government's decision to take immediate measures to reduce air pollution. Zhengzhou, China's east-central city has a mortality toll from air pollution of 29,000 people per year or more. NO2 was the main
contaminant in the air at the time. Diesel vehicle emissions are the most common source of NO2 emissions. Problems with the lungs are the outcome, and breathing becomes problematic as a result. The pollution level in Zhengzhou, China's east-central capital was more than 100 g/m3 in
2016, and a high warning was issued by a monitoring station. My motive for writing about this issue may be summed up in one sentence. It is my hope that the contribution made by this study will provide a reliable way to anticipate air pollution and assist people live healthier lives by decreasing their exposure to pollution and preventing health problems from arising.

The purpose of this study was to use the ARIMA model to forecast the amount of pollution in the air in Zhengzhou. Air pollution from fine particles from smoking, dost, and other sources is a big concern for Chinese cities today. Lung cancer has been linked to air pollution, according to WHO et al. (2018).

1.2. Objective of the study

The basic objectives of this research was to give an overview of the extensive research effort that has been carried out, and to take use of the state-of-the-art methods and techniques of applied Big Data for air quality assessment and forecasting. The map of air quality in Zhengzhou, China, was drawn and displayed using data from the city. In this study, artificial neural networks, Genetic Algorithm ANN Model, decision tree, and Deep belief network are the algorithms that were utilised, and the model's advantages and disadvantages were discussed in detail [11].

The proposed system performs two critical functions. I detect the presence of PM2.5 in the atmosphere depending on the values of the surrounding environment. (ii) Predicts the amount of PM2.5 for a certain day in the future For example, Logistic regression is used to evaluate whether a data sample has been corrupted or not. For the purpose of predicting future PM2.5 levels based on historical laboratory observations, auto regression is used [3]. The major goal is
to estimate the level of pollution in the city based on measurements taken on the ground.

1.3. Research Question

The air pollution in Zhengzhou has been caused by (Pollution, population and traffic) to what degree can we enhance it utilizing forecasting methods like ARIMA to extend people's lives?

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 10 Issue 11, November-2021

2. LITERATURE REVIEW

There are several air quality forecasting studies being conducted. To estimate the severity of air pollution, Ishan Verma and colleagues developed a Bi-directional LSTM model. Three
Bidirectional LSTM represent the short, long, and immediate impacts of the severity degree of

PM 2.5 in this system to increase prediction accuracy.

According to Temesegan Walelign Ayele and colleagues, an Internet of Things (IoT)-based air pollution system is being developed to monitor and evaluate air quality as well as anticipate pollution levels. The IoT and ML algorithms used in this system, notably the Recurrent Neural Network-LSTM, were created. The severity level of PM 10 pollution in Delhi is high. Using Multi-Layer Perception, which is an Artificial Neural Network, Nave Bayes, and Support Vector Machine, Aly Akhtar et al. built a system to forecast and assess Delhi's PM 10 pollution levels. For the purpose of determining which algorithm is more accurate, all of the aforementioned algorithms are compared.

The MLP method, which had a 98% accuracy rate, was used to build the model for predicting Delhi's pollution levels. Data mining was used by Shweta Taneja and colleagues to develop a method for forecasting air pollution trends [17]. Linear regression and Multi-Layer Perception are used in this system to understand patterns in many sorts of contaminants. This system analyzes and forecasts air pollution trends based on historical data. Dongping Qin's prediction model for PM 2.5 concentration in metropolitan areas used a combination of CNN and LSTM approaches. A Convolution Neural Network serves as the model's foundation, while a Long Short-Term Memory Neural Network serves as the model's output layer. The input layer is used to extract features, and the output layer is used to look at how pollutants change over time. Yue Shan Chang et al. created a cloud-based semantic ETL framework for air quality prediction and analysis. This system makes use of ontology to examine the connections between PM 2.5 in various data sets and to combine the results. After then, these datasets are evaluated to arrive at a prediction, which is then shown graphically. A comparative examination of air quality
forecasting systems based on Machine Learning approaches was provided by Saba Ameer et al. Chao Zhang et al. created a system for predicting air quality using an ensemble learning technique. The air pollution quality is predicted and analyzed using a multi-channel ensemble learning system with a supervised assignment mechanism [14].

3. RESEARCH METHODOLOGY

The system is divided into two distinct phases and these are given as:

- **Phase 1:** A model (line/curve) is fitted depending on the algorithm selected according to the data in the data set during the training phase.
- **Phase 2:** testing: the system is given inputs and tested to see whether it works. It's verified that everything is correct.

As a result, the data used to develop or validate the model must be accurate. The system's purpose is to identify and forecast PM2.5 levels, thus it has to utilize the right algorithms for the job. The accuracy of several algorithms was compared before they were picked for future usage. The person who was most qualified for the job was selected.

3.1. Data Set

This dataset was taken from UCI's repository and was used to train the air quality detection algorithm. The following characteristics were to be included in the data set:

1. Temperature
2. The speed of the wind
3. Dewpoint
4. Pressure
5. PM2.5 Concentration (in micrograms per cubic meter)

The data sample is either classed as contaminated or non-polluted as a result of the analysis.

3.2. Dataset Description

The collection has about 3000 entries representing all of Zhengzhou's states. We concentrated our efforts only on the Zhengzhou Dataset.

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 10 Issue 11, November-2021

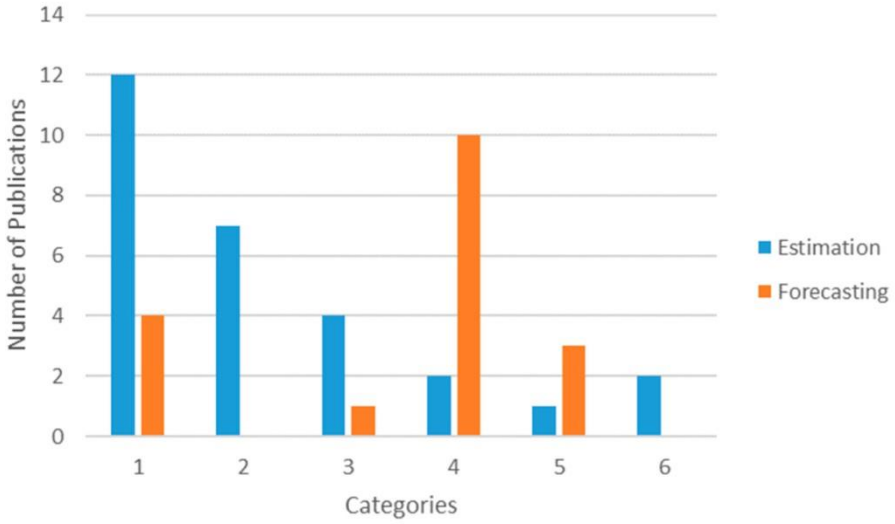

Figure 2 Machine learning approaches for outdoor air quality modeling.

3.3. Preprocessing and feature selection

We have solely researched and used Zhengzhou State data algorithms. Therefore, rows number was decreased to 3,200 and the state column is no longer automatically used. All pm2 5 values were null, therefore the column was deleted. The Agency's name has nothing to do with the level of contamination of the state. Also, stn code is also not useful. The date gives a better indication of the sample's date attribute and thus removes the redundancy [3]. The property location monitoring station is not needed again since it provides the monitoring station location which we do not need to take into account for analyzing To sum up, we have removed from our dataset the following features: State,pm2 5, agency, stn code, date of sampling, and place monitoring station[5].

3.4. Classification Algorithms

A variety of machine learning methods are used to train the dataset, including Logistic Regression, Naive Bayes, Support Vector Machines, and Decision Trees. Precision, recall, f1 score, specificity, sensitivity, and accuracy are the performance measurement parameters employed in the computation. Few of them are given below.

3.4.1. Logistic Regression

Logistic regression generates a binary output that can be used to forecast a categorical dependent variable's outcome. There should be a discrete/categorical result from the logistic regression, such as "one or zero," "yes or no," and "high or low". The sigmoid function used in this approach turns any continuous value into a discrete number.

3.4.2. Naïve Bayes

The nave Bayes algorithm is a statistical classification method based on Bayes' theorem of infinite variance. This theorem is based on the erroneous premise that the input variables are unrelated. Using the probabilities of each attribute belonging to each class, this approach for predictive analysis is simple and effective. Building a Nave Bayes classifier and using it on a large dataset are both simple and beneficial [8].

3.4.3. Decision Tree

Decision trees are widely used in machine learning because they are both effective and simple to implement in a new system. It uses a tree topology to create classification and regression models. Decide on which class of end result variable you want to forecast by using a decision tree method that learns from previous training data. When it comes to classification, the if-then rule is mutually exclusive and exhaustive [4].

3.4.4. Support Vector Machine

A separate hyperplane designed the SVM classification algorithm. This algorithm's goal is to separate the input data points in the best possible way as quickly as possible. In high-dimensional spaces, the SVM method works well. In the decision function, it uses only a small subset of training points, therefore it's memory-efficient as well [15].

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 10 Issue 11, November-2021

3.4.5. Random Forest

Random Forest is an effective machine learning ensemble strategy for both classification and regression. It's mostly used for putting together classifications. The random forest method mixes multiple different decision trees to produce a forest of trees. For each node in the decision tree, the Random Forest algorithm predicts the result and then picks out the best answer from among the guesses that it produces. This is an ensemble method that reduces over-fitting by averaging the outcomes [12].

4. DATA ANALYSIS

The graph below displays so2 concentration throughout the years. In 1997 and 2001 it was highest while in 1988 and 2003 it was lowest. It is nevertheless steady in recent years.

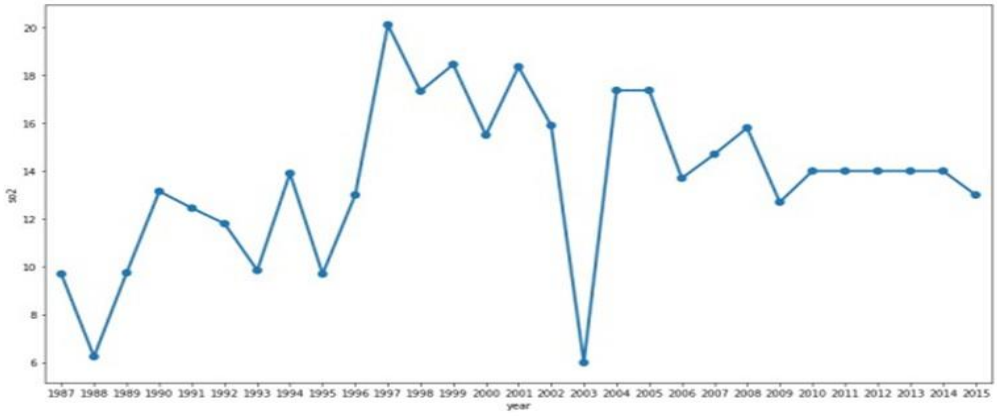

Figure 3 Analysis of Air Quality data set

This graph indicates that so2 in industrial regions is the highest. From this graph we may infer that Zhengzhou has the deadliest so2 in comparison to other cities.

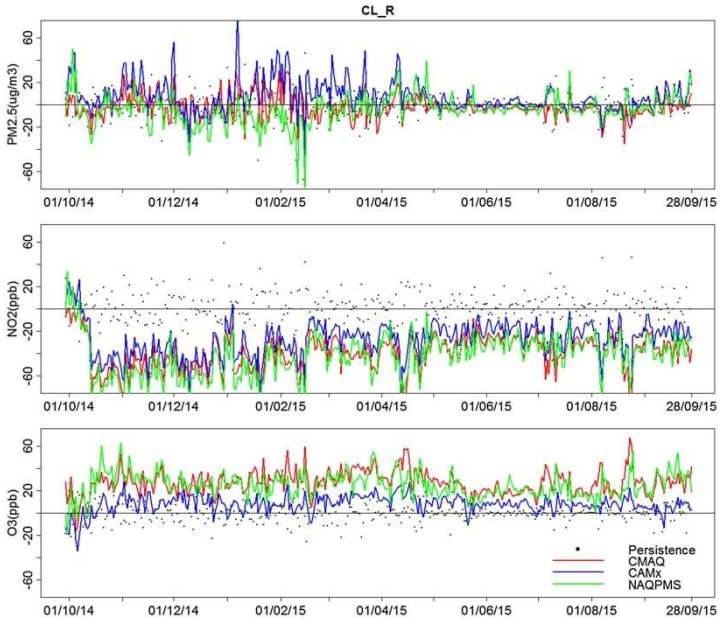

Figure 4 Series forecasting of Air Quality based on numerical model in Zhengzhou

4.1. Air pollution data set

There are 824 tuples and 9 attributes in this dataset. Country, state, city, location, last update, minimum, maximum, average, and pollutants are among the 9 attributes listed. They are grouped together. Furthermore, string and numeric are the names of these two types of data types.

4.2. Splitting data set

Training and testing datasets were created from the same dataset. When using a standard

80/20 split, this dataset is divided into two halves, with one half being used for training and the other half for testing.

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 10 Issue 11, November-2021

5. RESULTS AND DISCUSSIONS

The proposed method uses supervised machine learning algorithms to assess an air pollution dataset and predict with high accuracy the quality of the contaminants in the air. With time series analysis, we may detect future data points. The models utilised are the same:

5.1. ARIMA model

In the 1970s, Box and Jenkins developed the ARIMA model, which is a well-known time series prediction model. In an ARMA model, the current value of a time series is linearly represented by its prior value, current and previous residual sequences.

An ARIMA model for time series analysis and prediction is a class of statistical models. ARIMA is a further extension of the Autoregressive Move Average and includes the concept of integration.

AR: Auto regression. One observation is connected to several trailing observations through a dependent link in this paradigm.

Integrated. The use of raw data distinguishing (for example, the erasing of a fact from a prior period of time) for a stable time series.

Average moving. Using the moving average model's residual error dependence between an observation and the time delay.

5.2. Auto Regression (AR) Model

Auto regression is a time series model that uses previous stage data to predict the future value. It is a fairly basic concept that may lead to accurate predictions of a succession of issues:

$$Yhat = B0 + b1 * XI.$$

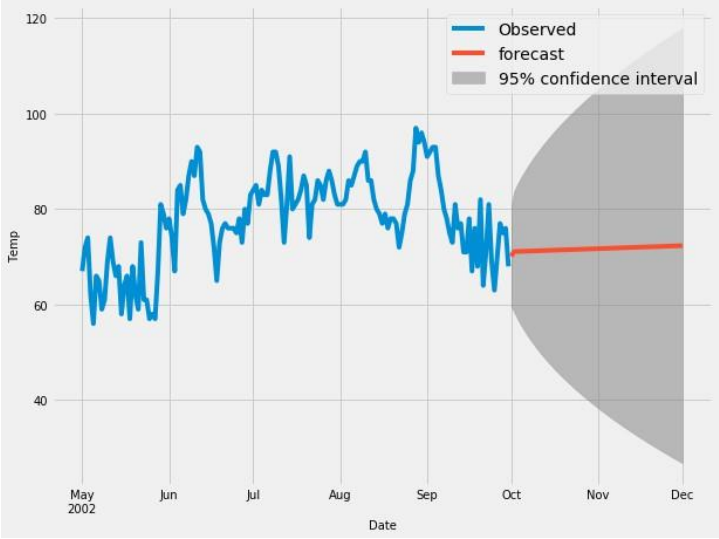

Figure 5 Applied forecasting in python, Air Quality data set (AIRMA MODEL)

B0 and b1 are model coefficients when trained on training data, while X is an input for the prediction. This method may be used to time series when lag variables (input variables) are used in previous observations [6]. It is possible to estimate the value of the following step (t+1) based on the results of the previous two steps (t-1 and t-2). It seems to be a regression model:

$$X (t+1) = b0 + b1 * X (t-1) * b2 + X (t-2)$$

Auto regressions are regression models in which the same input variable is used again throughout the course of the analysis (Self regression).

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 10 Issue 11, November-2021

5.3. Application of Algorithm

The ARIMA algorithm was employed to estimate the PM10 and PM2 parameters. Data from

80 percent of the collection was used to train the Machine Learning algorithm. The remaining

80% of the data was used to test the algorithm and predict the next value in accordance with the following procedure:

```
Series = read_csv ('file.csv', header=0,
parse_dates=[0], index_col=0, squeeze=True,
date_parser=parser)
X = series.values
size = int(len(X) * 0.80)
train, test = X[0:size], X[size:len(X)]
history = [x for x in train]
predictions = list()
for t in range(len(test)):
model = ARIMA(history, order=(5,1,0))
model_fit = model.fit(disp=0)
output = model_fit. forecast ()
yhat = output[0]
predictions.append(yhat)
obs = test[t]
history.append(obs)
print('predicted=%f, expected=%f' % (yhat, obs))
error = mean_squared_error(test, predictions)
print('Test MSE: %.3f' % error)
```

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 10 Issue 11, November-2021

After that, the true (measured) value will be added to the active data set and the absolute deviation determined. Every time you use this procedure, the set's average will be altered.

5.4. Predict Results

After completing our investigation and reaching the final step of prediction, we have divided our results into two categories, which are characterized into two classes and these are given as:

1. If the AQI value of pollution is > 100 then the cause of pollution is severe.

2. If the AQI value of pollution is < 100 then the cause of pollution is minimal.

This prediction system assists those who suffer from asthma to keep themselves away from polluted areas, moreover; it was developed to help the metrological department make accurate predictions about air quality. By showing the prediction result in either a web-based or a desktop application, this air quality forecasting method may one day be automated by using Artificial Intelligence. [14].

5.5. Experimental Setup

Tensorflow version 2.3.0 and Kera's version 2.4.3 were used to conduct our experiments and assessment. We choose a lower proportion for validation and testing datasets since we have a larger dataset and can exploit the vast quantity of data for training purposes using this strategy. For hyper-parameter tuning, the validation dataset was employed, and a grid search strategy was applied.

6. CONCLUSION

In the near future, controlling pollution levels in the air will be critical. People must be aware of the extent of pollution in their surroundings and take action to combat it. The findings suggest that machine learning algorithms (such as logistic regression and auto regression) may be effectively utilized to identify air quality and forecast future levels of PM2.5.

The suggested system will assist the general public, as well as meteorologists, in detecting and forecasting pollution levels, and then taking the appropriate action in response to that information. Additionally, this will serve as a data source for smaller towns and communities that are often overlooked in favor of major metropolitan areas.

Based on the plots of bars we conclude that certain cities are very dirty and urgently require care. We have researched from now on to avoid encounter issues in places like Kaifeng, Xinxiang, Xuchang, where the concentration of so2 is rising. For the prediction of so2 values, we utilised the AR model and the ARIMA model. Features like location monitoring station or station code are useless since they have nothing to do with so2 forecasts.

- The safe amounts are: So2
- Averaged over one hour, 0.20 ppm (parts per million).
- Averaged 0.08 ppm during a 24-hour period.
- Averaged 0.02 ppm throughout one year.

Pm2 5 is also an essential characteristic in order to forecast air quality. These particles must be documented in future since they have different health consequences including cardiovascular impacts such as heart rhythms and heart attacks as well as respiratory effects, such as asthma attacks and bronchitis. This model can't give the expected results, since the data doesn't follow the date column in sequence. The same applies for the municipalities. If we forecast the whole condition, it will not assist. We will now calculate AQI and apply additional categorization models. This model also brings us to understand future problems and research requirements like pm2.5, AQI, etc.

7.REFERENCES

[1] Arsov, M., Zdravevski, E., Lameski, P., Corizzo, R., Koteli, N., Mitreski, K. and Trajkovik, V., 2020, September. Short-term air pollution forecasting based on environmental factors and deep learning models. In *2020 15th Conference on Computer Science and Information Systems (FedCSIS)* (pp. 15-22). IEEE.

[2] Bhalgat, P., Bhoite, S. and Pitare, S., 2019. Air quality prediction using machine learning algorithms. International Journal of Computer Applications Technology and Research, 8(9), pp.367-390.

[3] Chang, Y.S., Abimannan, S., Chiao, H.T., Lin, C.Y. and Huang, Y.P., 2020. An ensemble learning based hybrid model and framework for air pollution forecasting. *Environmental Science and Pollution Research*, *27*(30), pp.38155-38168.

[4] Chen, S., Kan, G., Li, J., Liang, K. and Hong, Y., 2018. Investigating China's Urban Air Quality Using Big Data, Information Theory, and Machine Learning. Polish Journal of Environmental Studies, 27(2).

[5] Du, X., Lu, C., Wang, H. and Ma, J., 2012. Trends of urban air pollution in Zhengzhou City in 1996–2008. Chinese Geographical Science, 22(4), pp.402-413.

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 10 Issue 11, November-2021

[6] Gu, K., Zhou, Y., Sun, H., Zhao, L. and Liu, S., 2020. Prediction of air quality in Shenzhen based on neural network algorithm. Neural Computing and Applications, 32(7), pp.1879-1892.
[7] Hu, J., Li, X., Huang, L., Ying, Q., Zhang, Q., Zhao, B., Wang, S. and Zhang, H., 2017.Ensemble prediction of air quality using the WRF/CMAQ model system for health effect studies in China. Atmospheric Chemistry and Physics, 17(21), pp.13103-13118
[8] Kang, G.K., GAO, J.Z., Chiao, S., Lu, S. and Xie, G., 2018. Air quality prediction: Big data and machine learning approaches. *International Journal of Environmental Science and Development*, *9*(1), pp.8-16.
[9] Koo, J.W., Wong, S.W., Selvachandran, G. and Long, H.V., 2020. Prediction of Air Pollution Index in Kuala Lumpur using fuzzy time series and statistical models. *Air Quality, Atmosphere & Health*, *13*(1), pp.77-88.
[10] Kumar, T.S., Das, H.S., Choudhary, U., Dutta, P.E., Guha, D. and Laskar, Y., 2021. Analysis and Prediction of Air Pollution in Assam Using ARIMA/SARIMA and Machine Learning. In *Innovations in Sustainable Energy and Technology* (pp. 317-330). Springer, Singapore.
[11] LI, G., JIANG, W.J. and XU, J., 2013. Analysis and grey prediction for air major pollutants in Zhengzhou. Journal of Henan University of Urban Construction.
[12] Ma, J., Li, Z., Cheng, J.C., Ding, Y., Lin, C. and Xu, Z., 2020. Air quality prediction at new stations using spatially transferred bi-directional long short-term memory network. Science of The Total Environment, 705, p.135771.
[13] Pandey, G., Zhang, B. and Jian, L., 2013. Predicting submicron air pollution indicators: a machine learning approach. *Environmental Science: Processes & Impacts*, *15*(5), pp.996-1005.
[14] Wang, J., Zhang, X., Guo, Z. and Lu, H., 2017. Developing an early-warning system for air quality prediction and assessment of cities in China. Expert systems with applications, 84, pp.102-116.
[15] Wang, J. and Song, G., 2018. A deep spatial-temporal ensemble model for air quality prediction. Neurocomputing, 314, pp.198-206.
[16] WEI, L., ZHU, W.J. and CHEN, H.S., 2009. Statistical Forecasting Method of Air Quality in Zhengzhou City. Journal of Nanjing Institute of Meteorology, 32(2), pp.314-320.
[17] Zhao, Z., Qin, J., He, Z., Li, H., Yang, Y. and Zhang, R., 2020. Combining forward with recurrent neural networks for hourly air quality prediction in Northwest of China. Environmental Science and Pollution Research, 27, pp.28931