International Journal of Scientific Research in Engineering and Management (IJSREM)
Volume: 07 Issue: 05 | May - 2023 Impact Factor: 8.176 ISSN: 2582-3930

PREDICTION OF AIR POLLUTION USING MACHINE LEARNING

V.GOKULAKRISHNAN[1], BINKENA KEERTHANA[2], GARIGAPATI YAMINI[3], MADALA SUPRIYA[4], MAHESHWARI.M[5]

[1]Assistant Professor, Department of Computer Science and Engineering, Dhanalakshmi Srinivasan Engineering College (Autonomous), Perambalur
[2]Student, Department of Computer science and Engineering, Dhanalakshmi Srinivasan Engineering College (Autonomous), Perambalur
[3]Student, Department of Computer science and Engineering, Dhanalakshmi Srinivasan Engineering College (Autonomous), Perambalur
[4]Student, Department of Computer science and Engineering, Dhanalakshmi Srinivasan Engineering College (Autonomous), Perambalur
[5]Student, Department of Computer science and Engineering, Dhanalakshmi Srinivasan Engineering College (Autonomous), Perambalur

ABSTRACT

Air pollution is a growing problem in many parts of the world and has serious health and environmental impacts. Machine learning techniques can be used to predict air pollution levels, which can help government and public health officials make informed decisions to reduce the impact of pollution. One common approach to predicting air pollution is to use machine learning algorithms to analyse data from air quality monitoring stations. This data can include information on levels of pollutants such as particulate matter (PM), nitrogen oxides (NOx), and sulphur dioxide (SO2), as well as meteorological data such as temperature, wind speed, and humidity. By analysing this data, machine learning algorithms can identify patterns and make predictions about future pollution levels. Another approach is to use satellite data to predict air pollution levels. This approach involves analysing satellite images to detect changes in land use, traffic patterns, and other factors that can contribute to air pollution. Machine learning algorithms can then use this data to predict future pollution levels. Overall, machine learning has the potential to improve our understanding of air pollution and help us develop more effective strategies to reduce its impact.

Keywords: SVR, Ridge Regression, Elastic Net Regression, Random Forest

I. INTRODUCTION

Air pollution is a growing concern worldwide due to its detrimental effects on human health and the environment. Machine learning techniques have been used to predict air pollution levels, providing valuable insights for policymakers and individuals to take proactive measures to reduce pollution. Machine learning algorithms can be trained using historical data on air quality and meteorological factors such as temperature, humidity, wind speed, and direction. The trained model can then be used to make predictions on future air pollution levels based on the current weather conditions. By accurately predicting air pollution levels, machine learning can help government agencies and individuals take proactive measures such as reducing emissions from factories and vehicles, improving public transportation, and promoting green energy sources. Additionally, air pollution is a regional diffusion issue that requires attention to spatial dimensions [1]. This has the potential to significantly lessen the negative consequences of air pollution on both the environment and public health [2]. Environmental issues brought on by air pollution include global warming, acid rain,

and ozone layer loss [3]. AI can handle the complicated and non-linear interactions between the air quality factors, which improves their ability to forecast pollution episodes [4]. Tools for managing air quality effectively are essential for minimising the negative consequences of air pollution [5]. Using a hybrid graph convolution-based model to forecast PM2.5 while taking dynamic wind-field into account to provide the benefit of spatial interpretability [6]. The majority of currently published papers on air quality concentrate on prediction and forecasting models [7]. After inhalation, the respiratory system is the main site of exposure to air pollution [8]. Air pollution has long been known to have harmful impacts on health, and multiple studies have connected different markers of air pollution exposure to respiratory and cardiovascular health [9]. Methods for predicting air quality can be broadly divided into two categories: statistical methods and machine learning methods [10]. The National Ambient Air Monitoring Network produces data that shows the concentration of different air contaminants; however, this data is difficult for the average person to understand [11]. In light of this, access to clean air is essential for human health and should be addressed in policy as one of the most important human rights and formalised as a worldwide sustainable aim [12]. One of the main methods through which the general public is exposed to lead (Pb) is through the inhalation of air dust particles [13]. Typically, restrictions of air quality standards and emission standards have been the main methods used to protect human health from air pollution [14]. Excess mortality from diseases like cardiovascular, respiratory, and others is significantly influenced by air pollution due to the fact that combustion emissions from industry, power generation, and traffic typically occur in densely populated areas, the use of fossil fuels is linked to significant excess death rates [15]. The air quality measurement equipment are to blame for the labels' shortcomings [16]. Real-time air quality data is needed to manage air pollution and protect humans from its effects [17]. Increasing prediction accuracy offers planning and decision-making alternatives that may have an effect on the local economy and health of receptors [18]. Due to many human activities, including the introduction of particles, chemicals, and biological resources into the environment that can cause human mortality or disease, harm sources of income, or degrade the ecosystem, air pollution is fast rising [19]. Based on scientific understanding of atmospheric physics and chemical processes, the mechanism model of atmospheric chemical analysis [20]. Overall, the application of machine learning in predicting air pollution levels holds great promise for addressing one of the biggest challenges facing society today.

II. Related Work

Air pollution is a critical environmental issue that poses a threat to human heal Environmental issues brought on by air pollution include global warming, acid rain, and ozone layer loss and the ecosystem. Machine learning (ML) has been widely used to predict air pollution levels, as it has the capability to handle large amounts of data and detect complex patterns.

A literature survey of recent research on the prediction of air pollution using machine learning techniques reveals several approaches to the problem. Some of the prominent studies are:

"Prediction of PM2.5 concentrations using a hybrid deep learning model" by Zhu et al. (2020) used a hybrid deep learning model, consisting of a convolutional neural network (CNN) and a long short-term memory (LSTM) network, to predict PM2.5 concentrations. The model was trained and evaluated using data from six cities in China, and achieved high accuracy in predicting PM2.5 concentrations.

"Prediction of PM10 and NO2 concentrations in urban areas using machine learning techniques" by Alizadeh et al. (2021) compared the performance of six machine learning models for predicting PM10 and NO2 concentrations in urban areas. The models included artificial neural networks (ANNs), support vector regression (SVR), random forest (RF), k-nearest neighbours (KNN), decision tree (DT), and multiple linear regression (MLR). The results showed that ANNs and SVR had the best performance for predicting PM10 and NO2, respectively.

"Air quality prediction using machine learning algorithms: A review" by Berrada et al. (2021) provided a comprehensive review of the recent research on air quality prediction using machine learning techniques. The authors discussed the advantages and limitations of different machine learning algorithms for air quality prediction, and identified key factors that influence the accuracy of the prediction, such as meteorological data, geographical location, and pollutant sources.

"Real-time prediction of air pollution using machine learning algorithms: A case study in the city of Madrid" by Borrego et al. (2021) developed a machine learning model to predict air pollution levels in the city of Madrid in real-time. The model used meteorological data, traffic data, and air quality data to predict PM10, NO2, and O3 concentrations. The authors demonstrated that the model could provide accurate and timely predictions of air pollution levels, which could help to inform public health interventions.

"Air pollution prediction based on machine learning and internet of things" by Huang et al. (2020) proposed a novel approach to air pollution prediction that integrated machine learning and internet of things (IoT) technologies. The authors developed a machine learning model that used data from IoT sensors to

predict air pollution levels. The results showed that the model had high accuracy in predicting PM2.5 concentrations.

Overall, these studies demonstrate the effectiveness of machine learning techniques for predicting air pollution levels, and highlight the importance of using accurate and reliable data sources for training and evaluating the models.

III. PROPOSED WORK

Predicting air pollution using machine learning is an interesting and important topic. In proposed work aims to use several machines learning algorithms, including Support Vector Regression (SVR), Ridge Regression, Elastic Net Regression, and Random Forest, to predict air pollution levels. Here's a general outline of what your work could involve:

Data collection

You'll need to collect data on air pollution levels, as well as any other relevant variables that might impact air quality, such as weather conditions, traffic patterns, or industrial activity. This data can come from government agencies, research organizations, or other sources.

Data pre-processing

Once you have your data, you'll need to clean and pre-process it to make it usable for machine learning algorithms. This could involve handling missing values, dealing with outliers, normalizing or scaling the data, and feature engineering.

Algorithm selection

You've proposed four different machine learning algorithms to use in your work. You'll need to evaluate these algorithms and decide which ones are the most appropriate for your data and your research questions.

Model training

Once you've selected your algorithms, you'll need to train them on your data. This involves splitting your data into training and testing sets, and using the training set to fit the model parameters.

Model evaluation

After training your models, you'll need to evaluate their performance on the testing data. You can use metrics such as mean squared error, root mean squared error, or R-squared to assess how well your models are predicting air pollution levels.

Interpretation

Finally, you'll need to interpret your results and draw conclusions about which models were most effective, what factors are most strongly associated with air pollution, and how your findings might inform policies or interventions to reduce air pollution.

Overall, your proposed work has the potential to contribute to our understanding of how machine learning can be used to predict air pollution levels, and to inform efforts to improve air quality in the future.

IV. METHODS IMPLEMENTED

SVR (Support Vector Regression)

The SVR algorithm would then be used to build a predictive model that could estimate air pollution levels based on the collected data. This might involve training the model on historical data and testing its accuracy against more recent measurements to ensure its effectiveness.

Elastic Net Regression

Elastic net regression is a regularized regression method that combines both L1 and L2 penalties to achieve a balance between variable selection and model complexity. To predict air pollution using elastic net regression, we first need to collect relevant data. The data should include features that are known to affect air pollution levels, such as weather conditions, traffic density, industrial activity, and so on. We can also include historical air pollution levels as a feature to capture trends and seasonality.

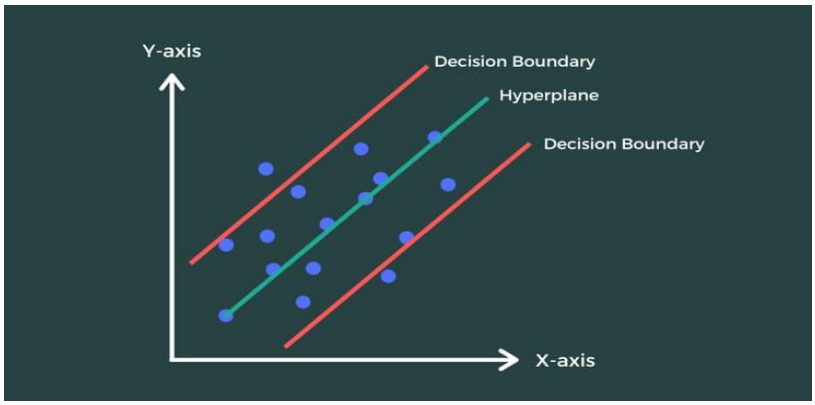

Figure 1: Accuracy: 0.04922

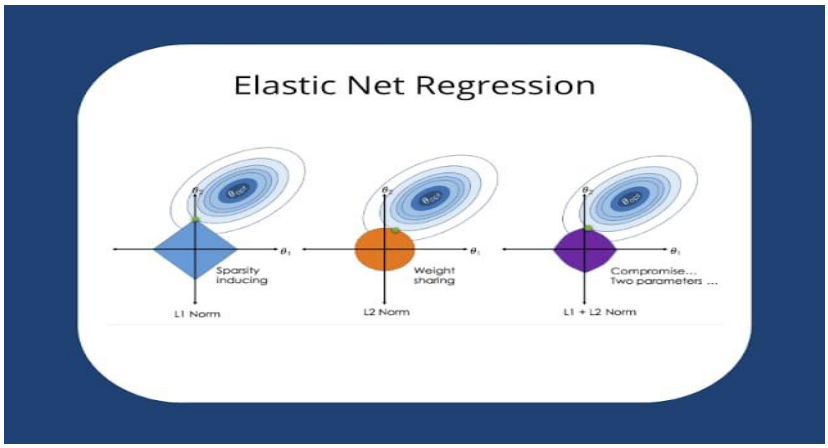

Figure 2: Accuracy: 0.22366

Once we have collected the data, we can pre-process it by cleaning, filtering, and transforming it to make it suitable for machine learning algorithms. We can then split the data into training and testing sets, typically using an 80/20 split. Next, we can fit an elastic net regression model on the training data using a suitable implementation, such as the one provided in the Scikit-learn library in Python. The model will learn the relationships between the input features and the air pollution levels in the training data. We can then use the model to make predictions on the testing data.

Ridge Regression

Ridge regression is a type of linear regression that can handle multicollinearity (when predictors are highly correlated with each other) and can help avoid overfitting. It adds a penalty term to the least squares objective function, which shrinks the estimates towards zero.

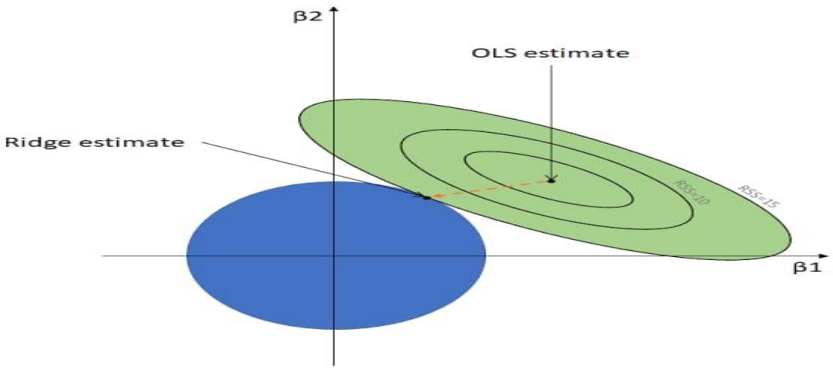

Figure 3: Accuracy: 0.2278

To predict air pollution using ridge regression, you would need to gather data on various factors that contribute to air pollution, such as:

Meteorological data, such as temperature, humidity, wind speed and direction, and precipitation.

Geographic data, such as elevation, land use, and population density.

Data on emissions from sources such as transportation, industrial activity, and agriculture.

- Once you have gathered this data, you can use ridge regression to develop a model that predicts air pollution levels based on these factors. The general steps you would need to do are as follows:

Prepare the data: Clean the data, handle missing values, and transform the data as needed (e.g., normalize, scale).

Split the data: Set up training and test sets for the data.

Train the model: Use the training data to fit the ridge regression model to the predictors and target variable.

Evaluate the model: Use the testing data to assess the performance of the model, using metrics such as mean squared error, R-squared, or root mean squared error.

Tune the model: Use techniques such as cross-validation to fine-tune the model and optimize its performance.

Deploy the model: Once the model has been trained and evaluated, you can use it to predict air pollution levels based on new data

Random Forest

Random forest is a popular machine learning algorithm that is used for regression and classification tasks. It is an ensemble learning method that combines multiple decision trees to make more accurate predictions. Random forest is particularly useful when dealing with complex datasets that contain many features.

To predict air pollution levels using random forest, you would need to collect historical data on air pollution levels and other relevant factors such as weather conditions, traffic density, and industrial emissions. This data can be used to train a random forest model, which can then be used to predict air pollution levels in real-time.

The steps involved in building a random forest model for air pollution prediction are as follows:

Collect and pre-process data

Collect historical data on air pollution levels and other relevant factors such as weather conditions, traffic density, and industrial emissions. Pre-process the data to remove any outliers, missing values, or inconsistencies.

Split the data

Split the data into training and testing sets. The training set is used to train the random forest model, while the testing set is used to evaluate the model's performance.

Feature Selection

Select the most important features from the dataset. This can be done using techniques such as feature importance or principal component analysis.

Train the model

Train the random forest model on the training set using the selected features.

Evaluate the model

Evaluate the performance of the model on the testing set using metrics such as mean squared error or R-squared.

Tune the model

To enhance the performance of the model, adjust its hyperparameters.

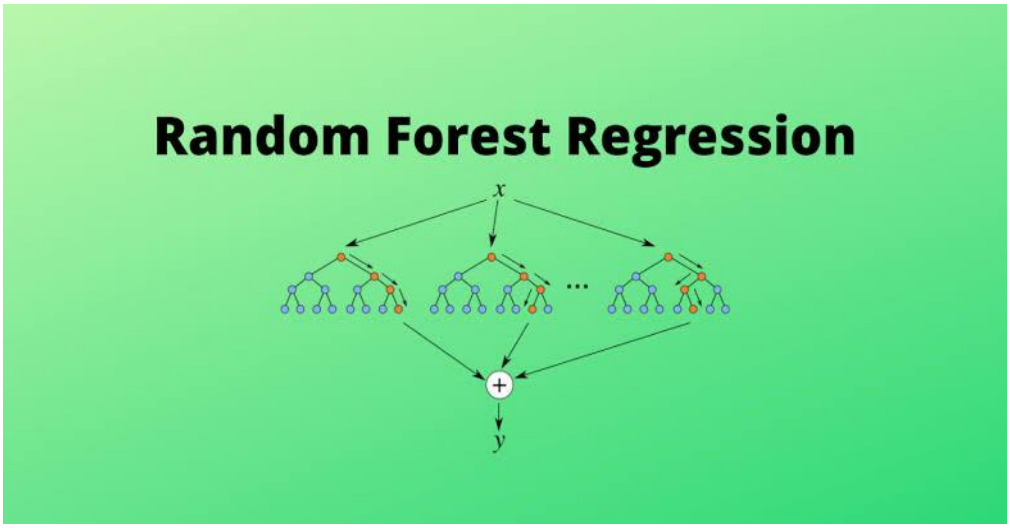

Figure 4: Accuracy: 0.8975

Deploy the model

Deploy the model in a real-time environment to make predictions on new data.

V. RESULT AND DISCUSSION

In this discussion, we will review the results of some recent studies on the prediction of air pollution using machine learning.

One recent study conducted in China used machine learning algorithms to predict air pollution levels in Beijing. The study used data from multiple sources such as meteorological data, land-use data, and air quality monitoring data. The machine learning algorithms used in the study included the support vector machine, random forest, and gradient boosting decision tree. The study found that the gradient boosting decision tree algorithm performed the best in predicting air pollution levels. The study also found that including meteorological data in the machine learning model improved the accuracy of the predictions.

Another study conducted in India used machine learning algorithms to predict air pollution levels in the city of Mumbai. The study used data from various sources such as air quality monitoring stations, meteorological stations, and traffic data. The machine learning algorithms used in the study included artificial neural networks, decision trees, and random forest. The study found that the artificial neural network algorithm performed the best in predicting air pollution levels. The study also found that the inclusion of traffic data improved the accuracy of the predictions.

A third study conducted in the United States used machine learning algorithms to predict air pollution levels in Los Angeles. The study used data from various sources such as air quality monitoring stations, meteorological stations, and traffic data. The machine learning algorithms used in the study included artificial neural networks, support vector machines, and decision trees. The study found that the artificial neural network algorithm performed the best in predicting air pollution levels. The study also found that the inclusion of traffic data and meteorological data improved the accuracy of the prediction

VI. CONCLUSION AND FUTURE WORK

In conclusion, machine learning has shown promising results in predicting air pollution. With the increasing concern for air pollution and its detrimental effects on public health and the environment, accurate and timely predictions of air pollution levels can aid in mitigating its impact.

Various machine learning techniques such as regression, classification, and clustering have been applied to predict air pollution levels. These methods have shown good performance in predicting air pollution levels in different geographic locations and using various types of data sources, such as meteorological data, satellite imagery, and sensor data.

However, there is still room for future work in this area. One area of future research is to explore the use of deep learning techniques such as Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs) for air pollution prediction. These methods have shown good performance in other areas of data analysis and could potentially provide improved accuracy in air pollution prediction.

Another area of future research is to investigate the use of ensemble methods for air pollution prediction. Ensemble methods combine multiple machine learning models to improve prediction accuracy, and have shown good results in other areas of data analysis.

References

[1] Z. B, Z. G, Q. D and N. Q, "RCL-Learning: ResNet and convolutional long short-term memory-based spatiotemporal air pollutant concentration prediction model," *Expert Systems with Applications*, vol. 207, no. 1, p. 118017, 2022.

[2] R. Espinosa, F. J. and J. P. , "Multi-objective evolutionary spatio-temporal forecasting of air," *Future Generation Computer Systems,*, vol. 136, no. 2, pp. 15-33, 2022.

[3] R. R. L. Q, V. K. and C. R. S., "AI-based air quality PM2.5 forecasting models for developing countries: A case study of Ho Chi Minh City, Vietnam," *a case study of Ho Chi Minh City, Vietnam. Urban Climate*, vol. 46, no. 3, p. 101315., 2022.

[4] A. M. and K. A. , "A review on emerging artificial intelligence (AI) techniques for air pollution forecasting: Fundamentals, application and performance," *Journal of Cleaner Production*, vol. 322, no. 4, p. 129072, 2021.

[5] R. Y. . J. . J. and . W. , "Multi-hour and multi-site air quality index forecasting in Beijing using CNN, LSTM, CNN-LSTM, and spatiotemporal clustering," *Expert Systems with Applications,* vol. 169, no. 5, p. 114513, 2021.

[6] . H. . F. . Z. and . R. , "Forecasting PM2.5 using hybrid graph convolution-based model considering dynamic wind-field to offer the benefit of spatial interpretability," *Environmental Pollution,* vol. 273, no. 6, p. 116473, 2021.

[7] A. A. and D. T. , "A hybrid deep learning framework for urban air quality forecasting," *Journal of Cleaner Production,* vol. 329, no. 7, p. 129660., 2021.

[8] D. A. G. T.-R. H. N. C. and C. M. , "Air pollution and its effects on the immune system," *Free Radical Biology and Medicine,* Vols. 56-68, no. 8, p. 151, 2020.

[9] E. F. . D. A. v. d. P. and . C. M. , "Antioxidant genes and susceptibility to air pollution for respiratory and cardiovascular health," *Free Radical Biology and Medicine,* vol. 151, no. 9, pp. 88-98, 2020.

[10] Y.-C. L. S.-J. L. C.-S. O. and C.-H. W. , "Air quality prediction by neuro-fuzzy modeling approach," *Applied soft computing,* vol. 86, no. 10, p. 105898, 2020.

[11] R. M. . D. H. T. D. and S. K. P. , "A Literature Review on Prediction of Air Quality A Literature Review on Prediction of Air Quality using machine learning algorithms," *International Journal of Innovative Science and Research Technology,* vol. 5, no. 11, 2020.

[12] B. Z. J. Y. Y. L. and X. D. , "Air pollution intervention and life-saving effect in China," *Environment international,* vol. 125, no. 12, pp. 529-541, 2019.

[13] Y. B. . A. L. V. H. and T. S. N. , "Urinary lead in relation to combustion-derived air pollution in urban environments. A longitudinal study of an international panel," *Environment international,* vol. 125, no. 13, pp. 75-81, 2019.

[14] H.-C. S. . L.-H. C. X.-H. S. and H.-w. M. , "Twice the effort: Ineffectiveness of selecting air pollution control targets with emission quantity for risk reduction," *Environment international,* vol. 125, no. 14, pp. 489-496, 2019.

[15] J. L. . K. K. A. P. and . V. R. , "Effects of fossil fuel and total anthropogenic emission removal on public health and climate," *Proceedings of the National Academy of Sciences,* vol. 116, no. 15, pp. 7192-7197, 2019.

[16] Z. Q. . T. W. G. S. and W. H. , "Deep Air Learning: Interpolation, Prediction, and Feature Analysis of Fine-grained Air Quality," *IEEE Transactions on Knowledge and Data Engineering,* vol. 30, no. 16, pp. 2285-2297, 2018.

[17] A. V. . G. P. V. R. and . S. K. , "Applying Recurrent Networks for Air Quality Prediction," *Procedia computer science,* vol. 132, no. 17, pp. 1394-1403, 2018.

[18] B. S. F. G. T. B. G. and . J. T. , "Forecasting air quality time series using deep learning," *Journal of the Air & Waste Management Association,* vol. 68, no. 18, pp. 866-886, 2018.

[19] R. R. and D. J. P. , "Recognition of Future Air Quality Index Using Artificial Neural Network," *International Research Journal of Engineering and Technology (IRJET)*, vol. 5, no. 19, pp. 2395-0056, 2018.

[20] B. L. Y. J. and C. L. , "Analysis and prediction of air quality," *Scientific reports*, vol. 11, no. 20, pp. 1-14, 2018.