Fuzzy Systems and Soft Computing
ISSN : 1819-4362

A FRAMEWORK OF PREDICTION OF AIR POLLUTANTS USING MACHINE LEARNING ALGORITHMS

Mrs.R.Sujatha Assistant Professor, Department of Computer Applications, J.H.A.Agarsen College, Madhavaram, (Affiliated to the University of Madras), Chennai, 600 060, Tamil Nadu India.
kamalihars2000@gmail.com
Dr.K.Shyamala Associate Professor PG & Research Department of Computer Science, Dr.Ambedkar Government Arts College (Autonomous), (Affiliated to the University of Madras), Chennai, 600 039, Tamil Nadu India. shyamalakannan2000@gmail.com

Abstract

"Air pollution" is the term used to describe when emissions in the air have a negative effect on the environment, materials or the health of humans or other living things. Carbon monoxide, nitrogen dioxide, particulate matter, and sulphur dioxide are the contaminants that are most harmful to human health. By creating respiratory disorders and other conditions, environmental and indoor air pollution greatly increases morbidity and mortality. This paper provides a framework which consists of four phases for predicting the air pollutants. The foremost aim of this study work is to predict the air pollutants RSPM, SO2 and NO2 during different seasons, in peak and non-peak hours in different cities of Tamilnadu, and Chennai and its suburbs by collecting hourly, daily, monthly and yearly data from pollutants dataset. The prediction work is carried out in four phases. Every phases of the implementation has proposed different algorithms. The algorithms are implemented and compared with bench mark algorithms to exhibit higher prediction accuracy and lower error rates.

Keywords: Air pollution, Pollutants, Hybrid technique, Prediction, Machine learning algorithms

1. Introduction

Air pollution is a combination of dangerous substances that can be both man-made and occur naturally. The primary human-caused pollution sources include motor vehicle emissions, fuels like petrol and natural gas that are used to heat homes, industrial and power plant waste, particularly from coal-fired power plants and chemical manufacturing odours. The urban environment is impacted by extreme pollutants in the air almost everywhere in the world. Poor urban pollution levels vary significantly between cities and suburbs and are primarily caused by local sources such as anthropogenic pollution. A detailed examination of meteorological circumstances is necessary to comprehend the variation in concentrations and distributions of airborne pollutants. The characteristics of aerosols are also impacted by meteorological driving parameters, which have an impact on dial-level air pollution concentrations [1].

Within artificial intelligence, machine learning enables systems to learn from experience and become more intelligent without explicit programming. Its numerous real-world uses across numerous industries have made it a focus of rising interest in recent years.

In the same way that no one size fits all caps, no one machine-learning approach works for every issue. A high-dimensional input space may be beyond the capabilities of some algorithms, which perform well with noisy data. While some people might have no trouble handling high-dimensional input space, others might find it difficult to deal with sparse data [2]. These conditions offer a strong basis for leveraging machine collaborative learning methods to enhance the potential approaches and use one to offset the inadequacies of the others.

The main aim of this work is to create framework to predict air pollutants.

- Air Pollutant Prediction for Chennai and suburbs
- Air Pollutant Prediction in different cities of Tamilnadu
- Air Pollutants Prediction in different seasons
- Air Pollutants Predictions in peak hours

2. Literature Review

This section summaries the current research being done by various researchers on forecasting air pollution.

Chi-Yeh Lin et al. [3] suggested a hybrid model for predicting air pollution. The study employed a combined model of Support Vector Machine (SVM), Long Short Term Memory (LSTM) and Gradient-Boosted Tree regression (GBT). The author evaluated the PM2.5 pollutant concentration and suggested a hybrid model. To show the efficiency of the hybrid model, GBT, LSTM, SVM and other models were compared. The results show that the hybrid model outperforms the other methods.

Samit Bhanja et al. [4] suggested a fusion deep learning algorithm to forecast air value over instance series. The author makes predictions regarding the quantity of air pollutants, such as the concentration of PM2.5. Analysis was done on the meteorology data sets from Delhi, India and Sydney, Australia. The author suggested a Hybrid Deep Neural Network (HDNN) structure to anticipate the pollutant PM2.5 by fusing a Convolutional Neural Network (CNN) and a Bidirectional Long Short Term Memory (BDLSTM). Better performance is generated using CNN and BDLSTM. In comparison to other models, the suggested model HDNN has minimal error.

The E-LSTM model, developed by Y. Bai et al. [5] and stimulated by ensemble learning develops various LSTMs in various modes and outperforms Feed-forward Neural Networks and single LSTMs in aspects of correlation metrics, root-mean-square (12.077 gm) and mean absolute inaccuracy (19.604% and 16.929% respectively) (0.994 and 0.991 respectively).

J.Zhao et al. [6] have proposed model, called an LSTM-FC (long short-term memory-fully connected) neural network, is based on data. It looks at past air quality data, meteorological data, weather prediction data, and the day of the week to guess how polluted a certain air quality monitoring station is with PM2.5 over the course of 48 hours. The dataset that the author used to evaluate our model contains data from 36 Beijing-area air quality monitoring stations between May 1, 2014, and April 30, 2015. The dataset is also used to compare our model with models from artificial neural networks (ANN) and long-short-term memory (LSTM). The outcomes demonstrate the superior prediction ability of the LSTM-FC Neural Network Model.

According to Zsolt Bodor et al. [7], the general quality of the air was at its best during summer while declining throughout the winter. Even though this was the case, the average amount of pollution in the air was below what was allowed. PM10 and other gaseous pollutants had a clear positive relationship, which was different from the expected negative relationship among temperature and wind speed. Although the 1-year maximum concentration (14.93 g/m3) is significantly lower than that of the legal maximum (40 g/m3) and is dropping compared to the previous 5 years, the daily pollutant concentrations were less than the legal limit.

According to Nabanita Ghosh et al. [8], just 25% of the days throughout the month of January are classified as moderate, with the rest being classified as poor. The proportion of "bad" and "moderate" days in December is roughly 43% and 57%, respectively. In February, just 17% of days with bad air quality were reported; as summer arrives and the air quality improves, this number rises to a moderate level in March. It is clear that PM10 is a severe contaminant in Kolkata location as a result. This study estimated the hazard of chronic obstructive pulmonary disease, respiratory disorders, cardiovascular diseases and elevated death rates and morbidity risks for a population exposed to PM10.

Hui Min, et al. [9] recommended the XGBoost algorithm to identify important miRNAs using CART (Regression and Classification Trees) on different type of sequence-based parameters. This technique is called XGEM (XGBoost for vital miRNAs). The success rate of XGEM predictions is capable. XGEM outperformed further high-tech methods highlighting its potential for finding important miRNAs.

To contrast XGBoost with Multi-Layer Perceptron models, M Poongodi et al. [10] offered higher precision and linkages between real-time data. Lastly, he found that XGBoost is significantly more suitable and efficient for taxi trip period estimates than Multi-Layer Perceptron after comparing the two aforementioned algorithms.

This selection gave a eye-opener of the research went on in the filed air pollution, various algorithms applied and different models developed. This has given rise to development of Frame work for prediction of Air pollutants using Machine learning algorithms.

3. Research Proposed work

The major aim of this study effort is to predict the air pollutants during different seasons, in peak and non-peak hours, in different cities of Tamilnadu, by collecting hourly, daily, monthly, yearly data from pollutants dataset. The prediction work is carried out in four phases. The four phases are explained in the following selections.

3.1 Harimax Model for Air Pollutant Prediction

ARIMA (Auto Regressive Integrated Moving Average Method) deals with statistical investigation that uses time – series information to identify with the dataset and to forecast upcoming trends. Auto ARIMA model use the pmd ARIMA package in the library which is used to fit the model. It will generate the optimal p, q and q values suitable for the dataset and provide better forecasting SARIMAX (Seasonal Regressive Integrated Moving Average through eXogenous factors) is a restructured edition of ARIMA representation. It uses past standards but also takes into description every seasonality prototype.

ARIMA model face the difficulty to predict turning point and it has poor performance for long term forecast and it cannot be used for seasonal time series. Auto ARIMA specifies the values for‘d’ parameter to make the model stationary. SARIMAX model make use of past values but also takes into account of seasonality pattern. By combining the advantages of ARIMA, Auto ARIMA and SARIMAX, a hybrid algorithm called HARIMAX is proposed in this work. [13]

The algorithm HARIMAX incorporates seasonal impacts, exogenous parameters with the auto regressive and moving average element. The algorithm HARIMAX predict the pollutants RSPM, SO2, and NO2 for the years 2020 to 2023 for the various areas of Chennai, AnnaNagar, Adyar, T.Nagar, Kilpauk, Nugambakkam. The predicted value of pollutants are compared with actual values of pollutants. The algorithm HARIMAX was compared with ARIMA and the HARIMAX algorithm exhibits better prediction accuracy and reduced Mean Absolute Percentage Error (MAPE).

3.2Modified LSTM for Air pollution Prediction

This work predict the daily concentrations of air pollutants. A network has been developed to track the environment air quality in Chennai and five cities of Tamilnadu – Salem, Thoothukudi, Madurai and Coimbatore. 648 datasets from public database in Tamilnadu Pollution Control Board [11] was collected for the pollutants PM2.5, PM10, NO2, SO2 between November 25, 2020 and September 13, 2022.

Regardless of its presence, a new layer will be formed for each iteration of the Long Short Term Method (LSTM). The approach will take longer to train if new layers are created throughout each iterations. This research suggests a novel procedure identified as Modified Long Short Term Memory (MLSTM) [14] in which layers are not formed with each cycle. Only if it does not already exist will a layer be created. The layer's reference will be the only thing created if the layer already exists. As a result, the training duration will be automatically decreased. The proposed model uses the H5 model. The LSTM model cannot be run in the cloud since it is not portable; however, the modified LSTM can function on a local workstation as well as in JCB, Amazon Assure, and other clouds.

The performance of the Modified LSTM was compared to LSTM method. The Modified LSTM method exhibits higher prediction accuracy and lesser execution time compared to LSTM method.

3.3 Prediction of Air pollutants in various seasons

Seasonal variations in air pollution are related to the diversity of the seasons and particularly of the months that makes up the so called summer and winter seasons. The incidence of greater air contamination levels in specific months of the year is related to the kind of atmosphere and therefore to the various weather conditions in those months, the varying climate conditions on any given activities.

The major purpose of this effort is to understand the difference among the different pollutants in various seasons. 720 records were gathered from Tamilnadu Pollution Control Board (TNPCB) website [11]. The Pollutants data set for four year duration 2017 - 2021 was taken from suburbs of Chennai – Adyar, AnnaNagar, Kilpauk, Nugambakkam and T.Nagar.

The Pollutant RSPM, SO2 and NO2 were taken for the study for the season's summer, Autumn, Spring and Winter. The Modified Adaptive Boosting Algorithm (MABA) was proposed [15] in this work for analyzing the variations among the different pollutants in various seasons and to forecast the seasonal variations of air pollutants for the year 2021. The algorithm Modified Adaptive Boosting algorithm (MABA) was compared with Adaptive Boosting Algorithm (ABA) and Modified Adaptive Boosting algorithm shows improved prediction accuracy and lesser execution times.

3.4 Prediction of Air Pollutants during Peak and Non-Peak hours

This work demonstrates how urban center's monitoring and evening rush hours are related to air pollution from transportation. The major aim of this effort is to comprehend how the pollutants PM2.5, PM10, SO2 and NO2, vary during peak and Non-Peak hours. The pollutants data set was collected from the suburbs of Chennai including Alandur, Arubakkam, Kodungayur, Manali and Velachery.

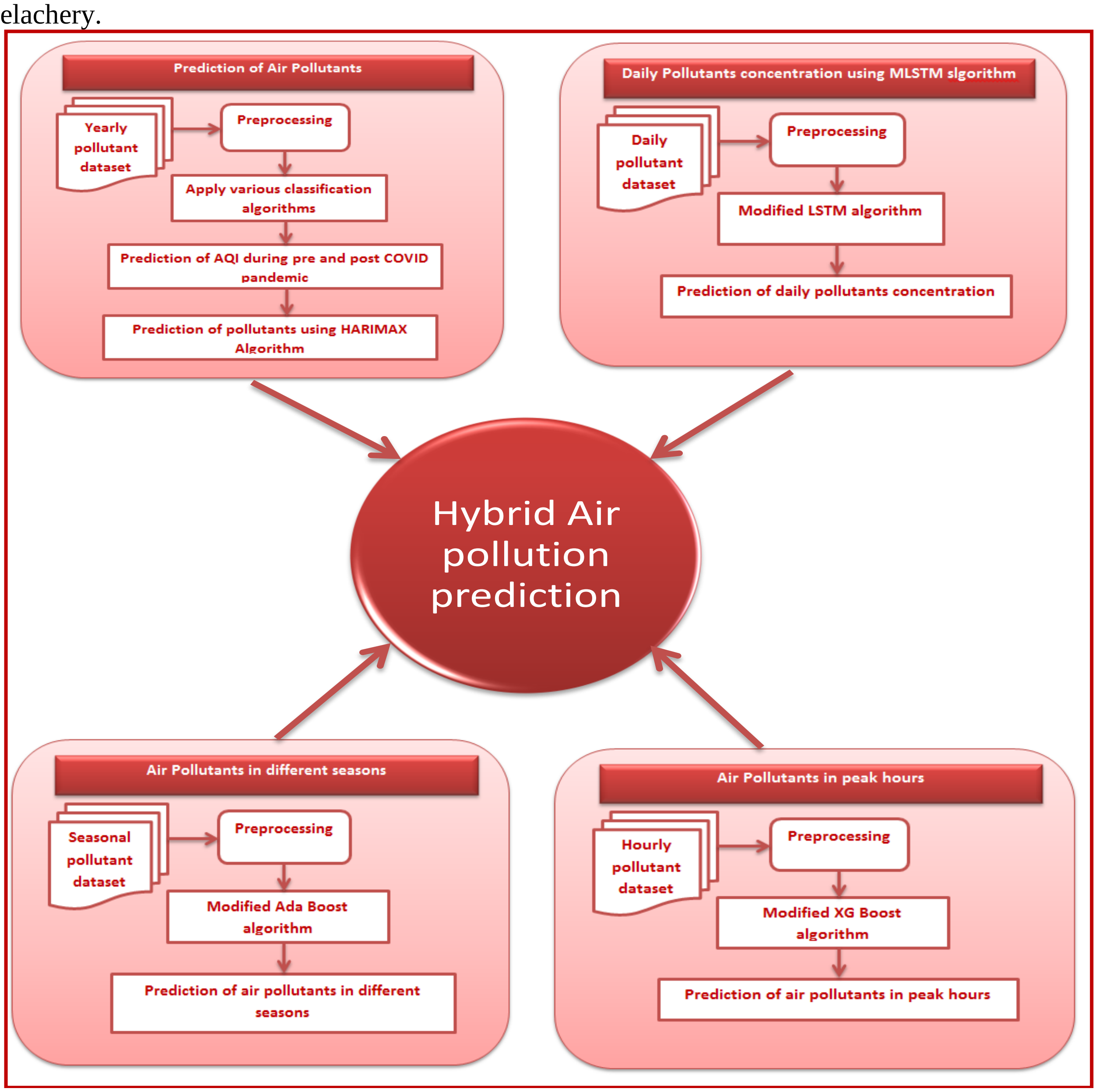
Prediction of Air Pollutants
Yearly pollutant dataset
Preprocessing
Apply various classification algorithms
Prediction of AQI during pre and post COVID pandemic
Prediction of pollutants using HARIMAX Algorithm
Daily Pollutants concentration using MLSTM slgorithm
Daily pollutant dataset
Preprocessing
Modified LSTM algorithm
Prediction of daily pollutants concentration
Hybrid Air pollution prediction
Air Pollutants in different seasons
Seasonal pollutant dataset
Preprocessing
Modified Ada Boost algorithm
Prediction of air pollutants in different seasons
Air Pollutants in peak hours
Hourly pollutant dataset
Preprocessing
Modified XG Boost algorithm
Prediction of air pollutants in peak hours

Figure 1.Frame work for Air Pollution Prediction system

The Modified XG Boosting algorithm (MXGBA) was proposed in this work [16] to forecast the fluctuations among the pollutants PM2.5, PM10, SO2 and NO2 during peak and non-peak hours. The study shows increase in level of pollutants PM2.5, PM10, SO2 and NO2 during peak hours of morning and evening. The concentrations of pollutants is more during peak hours than non- peak hours.

The Modified Extreme Gradient Boosting algorithm (MXGBA) was compared with the performance of Extreme Gradient Boosting Algorithm (XGBA) and Modified Extreme Boosting Algorithm (MXGBA) demonstrate higher accuracy and low error rates compared to Extreme Gradient Boosting Algorithm (XGBA)

4. Conclusion

The framework shows the four phases of work. In the first phase, yearly pollutant dataset was collected. Various classification algorithm were applied to dataset for predicting the Air quality levels, out of which Decision tree gives 98.1 accuracy. The Air quality level was calculated using pre and post COVID pandemic period and level of pollutants were less during lock down period. Hybrid algorithm HARIMAX was proposed incorporating the advantages of ARIMA, Auto ARIMA and SARIMAX.HARIMAX was compared with ARIMA and it exhibits better prediction accuracy and reduced Mean Absolute Percentage Error (MAPE).

The next phase of aim is to predict the daily concentration of air Pollutants. Modified Long Short Term (MLSTM) algorithm was proposed in this phase to forecast the daily concentration of air pollutant. The performance of the Modified LSTM was compared with LSTM method. The Modified shows higher prediction accuracy and lesser execution time compared to LSTM method.

In Third Phase, deals with analyzing the variation among the different pollutants in various seasons. The Modified Adaptive Boosting Algorithm (MABA) was proposed in this phase to forecast the seasonal variations of air pollutant. Its performance was compared with Adaptive Boosting Algorithm (ABA) and MABA exhibits higher prediction accuracy with lesser execution time.

The fourth phase relates between air pollution and peak hours. It also shows the variations in level of pollutants during peak and non-peak hours. The Modified Extreme Gradient Boosting algorithm (MXGBA) was proposed in this phase and it demonstrate higher accuracy and low error rates. When compared with Extreme Gradient Boosting Algorithm (XGBA).

This research work provide a framework that depicts prediction of air pollutants by taking hourly, daily, yearly and seasonal datasets by utilizing machine leaning algorithms.

References

1. Unique identification authority of India, "https://uidai.gov.in/images/state-wise-aadhaar-saturation.pdf", [Online; accessed 16 – July - 2020].
2. 6 of the world's 10 most polluted cities are in India" https://www.weforum.org/agenda/2020/ 03/6-of-the", [Online; accessed 22- August-2020].
3. CY Lin, YS Chag, HT Chiao and S Abimannan, "Design a Hybrid Frame work for Air Pollution Forecasting", *2019 IEEE International Conference on Systems, Man and Cybernetics (SMC)*. IEEE, pp 2472 – 2477, 2019.
4. S Bhanja and A Das, "A hybrid deep learning model for air quality time series prediction", *Indonesian Journal of Electrical Engineering and Computer Science*, Vol.22 (3), pp 1611-1618, 2021.
5. Y. Bai, B. Zeng, C. Li, and J. Zhang, "An ensemble long short term memory neural network for hourly PM2.5 concentration forecasting", *Chemosphere*, Vol. 222, pp. 286–294, 2019.
6. J. Zhao, F. Deng, Y. Cai, and J. Chen, "Long short-term memory—fully connected (LSTM- FC) neural network for PM2.5 concentration prediction", *Chemosphere*, Vol. 220, pp. 486–492, 2019.
7. Z Bodor and K Bodor, "Major air pollutants seasonal variation analysis and long-range transport of PM10 in an urban environment with specific climate condition in Transylvania (Romania)", *Environmental Science and Pollution Research*, Vol. 27 , pp 38181–38199, 2020.

8. N Ghosh, A Roy, R Mandal and A Dutta. “Degradation of Air Quality (PM10) with Seasonal Change and Health Risk Assessment in Metro City Kolkata”, *International Journal of Advancement in Life Sciences Research,* Vol. 3(1), pp 24-31, 2020.

9. M. Amjad, I. Ahmad, M. Ahmad, P. Wróblewski, P. Kamiński and U. Amjad, “*Prediction of Pile Bearing Capacity Using XGBoost Algorithm: Modeling and Performance Evaluation*”, Applied Sciences, doi: 10.3390/app12042126, Vol.12(4), pp 1- 24, 2022.

10. X. Deng, M. Li, S. Deng, and L. Wang, “Hybrid gene selection approach using XGBoost and multi-objective genetic algorithm for cancer classification”, *Medical & Biological Engineering & Computing*, doi: 10.1007/s11517-021-02476-x ,Vol. 60(3), pp 663–681, 2022.

11. Data for Air Pollutants: ”https://tnpcb.gov.in/air-quality.php”, [Online; Accessed March 2023].

12. Mrs. R. Sujatha and Dr. K. Shyamala,” Analysis of Air Pollution Data for Prediction of Air Quality Levels”, *Journal Telematique,* ISSN: 1856 - 4194, Vol. 21(1), pp 2598 - 2606, 2022.

13. Mrs.R.Sujatha and Dr.K.Shyamala,”Harimax Model for Prediction of Air Pollutants in Chennai”, *Neuroquantology,* DOI : 10.14704/Nq.2022.20.11.Nq66432, pp 4270 – 4279, 2022.

14. Dr. K. Shyamala and Mrs. R. Sujatha, "Modified LSTM for prediction of Pollutants Concentration", *In Human Machine Interaction in the Digital Era*, pp 171-181, CRC Press, 2024.

15. Mrs. R. Sujatha and Dr. K. Shyamala, "Modified Adaptive Boost Algorithm for Prediction of Air Pollutants in Various Seasons" *In proceedings of Emerging Research Trends in Computer Science and Information Technology*”, ISBN 978-81-948053-3-5, pp 25 -30, March 2023, Chennai.

16. Dr. K. Shyamala and Mrs. R. Sujatha ,"Modified Extreme Gradient Boosting Algorithm for Prediction of Air Pollutants in Various Peak Hours", *In International Conference on Advancements in Smart Computing and Information Security*, Vol. 2037, pp 125-141, 2023, Cham: Springer Nature Switzerland.