WWW.IJRPR.COM

International Journal of Research Publication and Reviews

Journal homepage: www.ijrpr.com ISSN 2582-7421

Air Quality Prediction Based on Deep Learning

G Gurupriya [1], Mr. J. Jayapandian [2]

[1]Master of Computer Application, Krishnasamy College of Engineering &Technology, Cuddalore,India
[2]MCA., M. Phil.,(Ph.D).,Associate Professor, Master of Computer Applications, Krishnasamy College of Engineering &Technology, Cuddalore, India.

ABSTRACT

Air pollution, a mixture of particulate matter and gases like NO2, CO, O3, and SO2, poses significant health risks, particularly for the young and elderly. As urbanization accelerates, forecasting air pollution becomes crucial for improving life quality in developing countries. Traditional machine learning models for air quality prediction often rely on limited data and standard regression techniques, which struggle with time and computational complexity. To address these limitations, we propose an XGBoost-based approach with a spatial transformation component and a deep distributed fusion network. This model leverages spatial correlations in air quality data to improve prediction accuracy, offering a more robust solution for air pollution forecasting.

Keywords: Air pollution, XGBoost-based , deep distributed fusion network, NO2, CO, O3, and SO2,

I. INTRODUCTION

With economic development and population growth in cities, environmental pollution issues such as air pollution, water pollution, noise, and land resource shortages have garnered increasing attention. Among these, air pollution's direct impact on human health has raised public awareness in both developing and developed countries. Air pollution, caused by power plants, industries, residential heating, vehicles, and natural disasters, poses significant health risks, especially in urban areas. Moreover, global warming from anthropogenic greenhouse gas emissions is a long-term consequence of air pollution. Accurate air quality forecasting can mitigate the effects of pollution peaks on populations and ecosystems, making it a critical goal for society. Air pollution involves the introduction of harmful substances, including particulates, biological molecules, and gases, into the Earth's atmosphere, leading to disease, death, and damage to living organisms and the environment. Pollutants can be solid particles, liquid droplets, or gases, classified as primary or secondary. Primary pollutants are directly produced from processes like volcanic eruptions or vehicle exhaust, while secondary pollutants form through chemical reactions in the atmosphere. The six criteria pollutants—ground-level ozone (O3), fine particulate matter (PM2.5), carbon monoxide (CO), nitrogen dioxide (NO2), sulfur dioxide (SO2), and lead—pose significant health threats, with O3, PM2.5, and NO2 being the most widespread.

Extensive experiments on real-time air pollution data demonstrate the effectiveness of machine learning in forecasting general patterns and sudden changes in air quality. Ensemble techniques, particularly gradient boosting with dropouts, improve the stability and accuracy of predictions. For sudden changes, recurrent neural networks with memory units yield the highest accuracy in classifying spikes. These machine learning results are compared with national air quality services to evaluate practical application, showing significant potential for improving air quality forecasts. The scope of air quality prediction using machine learning encompasses environmental monitoring, public health management, urban planning, and climate change mitigation. Machine learning algorithms can forecast pollutant levels like ozone, nitrogen dioxide, and carbon monoxide, aiding environmental agencies in timely interventions. These predictions can help urban planners optimize city infrastructure and transportation systems to reduce pollution. Integrating IoT devices and data fusion techniques enhances prediction accuracy, enabling personalized health recommendations and supporting research for improved air quality management strategies.

II. LITERATURE SURVEY

Machine learning has emerged as a powerful tool for crop yield forecasting, leveraging large-scale data and advanced algorithms to enhance prediction accuracy. Paudel et al. (2021) highlight the efficacy of machine learning in forecasting large-scale crop yields, underscoring its ability to process vast datasets and optimize predictive models. Similarly, Abbas et al. (2020) demonstrate the potential of proximal sensing combined with machine learning to predict crop yields, showcasing how these technologies can offer precise and timely insights into agricultural productivity. These studies emphasize the growing importance of machine learning in addressing the challenges of modern agriculture by providing reliable yield predictions that can inform management practices and policy decisions. Several researchers have explored the integration of various machine learning techniques with remote sensing and environmental data to improve crop yield predictions. Bian et al. (2019) utilized multispectral UAV data and machine learning methods to predict wheat yield at a field scale, achieving significant accuracy improvements.

Palanivel and Surianarayanan (2019) proposed a big data approach for crop yield prediction, demonstrating how combining machine learning with extensive datasets can enhance forecasting capabilities. Van Klompenburg et al. (2020) conducted a systematic literature review on crop yield prediction using machine learning, identifying key trends and methodologies that have proven effective across different studies. These works illustrate the diverse applications and potential of machine learning in agricultural forecasting. The impact of climatic factors on crop yield prediction has also been a focal point of research. Mishra et al. (2020) employed adaptive boosting techniques to account for climatic variability in crop production forecasting, finding that this approach significantly improves prediction accuracy. Ogutu et al. (2018) used dynamic ensemble seasonal climate forecasts for probabilistic maize yield prediction in East Africa, demonstrating the benefits of integrating climate models with machine learning. Dash et al. (2018) applied artificial intelligence to rainfall prediction in Kerala, India, highlighting the role of machine learning in anticipating environmental conditions that directly affect agricultural outputs. These studies underscore the necessity of considering climatic factors in crop yield models to enhance their reliability and applicability. Recent advancements in sensing technologies and data fusion have further propelled the capabilities of machine learning in agriculture.

Talaviya et al. (2020) explored the use of artificial intelligence for optimizing irrigation and pesticide application, emphasizing the operational efficiencies and environmental benefits that machine learning can bring to farm management. Holzman et al. (2018) assessed crop yield using remotely sensed water stress and solar radiation data, demonstrating early detection capabilities for agricultural monitoring. The survey by Bali and Singla (2022) on emerging trends in machine learning for crop yield prediction reveals a growing interest in integrating advanced sensing techniques and data analytics to improve forecasting precision. Collectively, these studies highlight the transformative impact of machine learning on agricultural practices, enabling more informed decision-making and sustainable farming strategies.

Drawback of Existing System:

Utilizing statistical models for air pollution prediction, the existing system employs support vector machines for numeric prediction, Random Forest for flexibility in both classification and regression, and Genetic algorithms to enhance performance through natural selection. Additionally, it incorporates an incremental K-means clustering methodology for weather forecasting based on limited datasets, complemented by Naive Bayes algorithm for classifying weather datasets with predefined labels. While statistical models have been widely used for air pollution prediction, many studies have overlooked the problem's nature and correlation between sub-models in different time slots

III. PROPOSED SYSTEM

Monitoring and preserving air quality has become one of the most essential activities in many industrial and urban areas today. The quality of air is adversely affected due to various forms of pollution caused by transportation, electricity, fuel uses etc. With increasing air pollution, we need to implement efficient air quality monitoring models which collect information about the concentration of air pollutants and provide assessment of air pollution in each area. In this project, we can input the air quality datasets and using the XGBoost model of the deep learning ensemble algorithm is performed on the six pollutant concentrations that currently mainly affect air quality, and the hourly prediction of AQI was achieved.

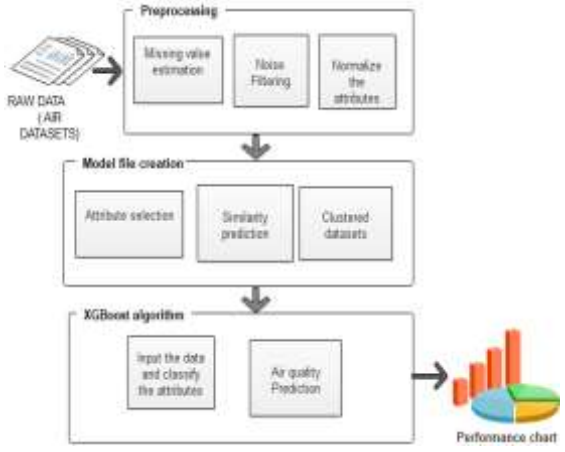

Figure 1: System Architecture of the proposed system

3.1 IMPLEMENTATION

Our project constituted of the below modules,

- Data Collection
- Datasets Acquisition
- Preprocessing
- Deep Learning Framework

- Performance Evaluation
- Accuracy on test set
- Saving the Trained Model

1. Data Collection

In the first module of Air Quality Prediction using machine learning, we developed the system to get the input dataset. Data collection process is the first real step towards the real development of a machine learning model, collecting data. This is a critical step that will cascade in how good the model will be, the more and better data that we get; the better our model will perform. There are several techniques to collect the data, like web scraping, manual interventions. Our dataset is placed in the project and it's located in the model folder. The dataset is referred from the popular standard dataset repository kaggle where all the researchers refer it. The dataset consists of patient numerical data. The following is the URL for the dataset referred from kaggle.

2. Datasets Acquisition

A data set (or dataset, although this spelling is not present in many contemporary dictionaries like Merriam-Webster) is a collection of data. Most commonly a data set corresponds to the contents of a single database table, or a single statistical data matrix, where every column of the table represents a particular variable, and each row corresponds to a given member of the data set in question. The data set lists values for each of the variables, such as height and weight of an object, for each member of the data set. Each value is known as a datum. The data set may comprise data for one or more members, corresponding to the number of rows. The term data set may also be used more loosely, to refer to the data in a collection of closely related tables, corresponding to a particular experiment or event. In this module, we can upload the air datasets which contains the attributes such as Carbon Monoxide (CO), Lead (Pb), Nitrogen Dioxide (NO2), Ozone (O3), Particulate matter (PM), Sulfur Dioxide (SO2). In this module we can upload the weather datasets which includes temperature, humidity, wind speed and lighting values. These datasets are collected from UCI repository.

3. Preprocessing

Data preprocessing is a data mining technique that involves transforming raw data into an understandable format. Real-world data is often incomplete, inconsistent, and/or lacking in certain behaviors or trends, and is likely to contain many errors. Data preprocessing is a proven method of resolving such issue. Missing Value treatment becomes important since the data insights or the performance of predictive model could be impacted if the missing values are not appropriately handled. Imputation of missing values from predictive techniques assumes that the nature of such missing observations is not observed completely at random and the variables chosen to impute such missing observations have some relationship with it, else it could yield imprecise estimates and Convert the unstructured data into structured datasets. Then calculate the missing value estimation and irrelevant data removal approaches.

4. Deep Learning Framework

Deep Learning algorithms can develop a layered, and hierarchical architecture of learning and representing data.In the feature extraction method, we extract the aspects from the processed dataset.And also implement XGBoost algorithm to classify the features. Decision trees create a model that predicts the label by evaluating a tree of if-then-else true/false feature, and estimating the minimum number of features needed to assess the probability of making a correct decision. Decision trees can be used for classification to predict a category, or regression to predict a continuous numeric value. Based on these features, predict the Air quality. The air quality prediction based on machine learning method, because of its flexible non-linear modeling ability, is often superior to the traditional statistical method in the prediction effect, but a single machine learning model often depends on expert knowledge and feature engineering to improve the prediction effect of the model. Boosting ensemble learning is a new rising machine learning mode. Its basic idea is to constantly use a "weak" classifier to make up for the shortcomings of the previous "weak" classifier, and finally form a "strong" classifier serially. XGBoost model is one of the boosting ensemble algorithms, which is based on the lifting tree model, so it ensembles many tree models together to form a strong classifier. At the same time, XGBoost model is improved on the basis of GradientBoostingDecisionTree (GBDT), making it more powerful and applicable to a wider range. Therefore, XGBoost model has the advantages of fast computing speed, strong model generalization ability, and significant model improvement effect. It is often used in some big data competitions, which can be used for both classification and regression problems

5.Performance Evaluation

We can evaluate the performance the system using accuracy parameters. The accuracy metric is evaluated as Accuracy = (TP+TN)/(TP+TN+FP+FN) The proposed algorithm provide improved accuracy rate than the machine learning algorithms.

6. Saving the Trained Model

Once you're confident enough to take your trained and tested model into the production-ready environment, the first step is to save it into a .h5 or .pkl file using a library like pickle. Make sure you have pickle installed in your environment. Next, let's import the module and dump the model into .pkl file

IV. RESULTS AND DISCUSSION

System testing aims to identify errors and ensure that software meets its requirements and user expectations through various testing types. Unit testing validates internal logic and outputs of individual software units, while integration testing assesses combined software components for consistency and correctness. Functional testing systematically checks that functions perform as specified by requirements, documentation, and user manuals. System testing ensures the entire integrated system functions correctly, focusing on process flows and integration points. White box testing examines the internal workings of the software, whereas black box testing assesses functionality without knowledge of the internal structure. A comprehensive test strategy includes field testing and detailed functional tests to ensure proper field entries, correct link activations, and timely responses, culminating in successful software integration and user acceptance testing to confirm the system meets all functional requirements.

Figure 1: Login Page

Figure 2: Enter Input Page

Figure 3: Prediction

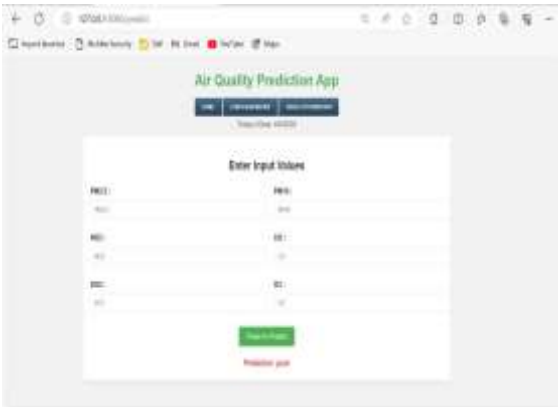

Figure 4: Prediction Result

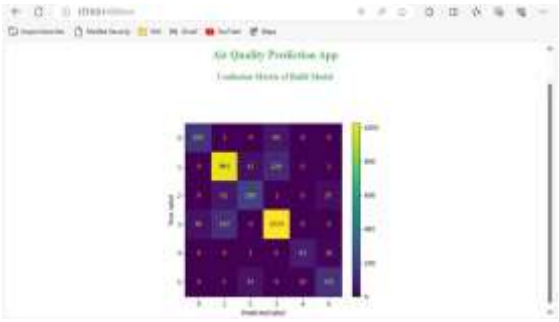

Figure 5: Analysis Screen

V. CONCLUSION

Air pollution play hazardous role in the health of the humans and plants. As there are many different sources of pollution, finding the effects of air pollution on health are very complex and their individual effects differ from one to the other. The data are preprocessed and Data can be further processed by data mining tool and proper decision support can be given to the policy makers. This project proves that data mining techniques are valuable tools that could be used for environmental monitoring and natural resource science field. It can be used to make reliable air quality index predictions. This model also facilitates decision making in day to day life. It can yield better results when applied to cleaner and larger datasets. Pre-processing of the datasets can be effective in the prediction as unprocessed data can also affect the efficiency of the model

REFERENCE

1 Paudel, Dilli, et al.: Machine learning for large-scalem crop yield forecasting. Agricultural Systems 187:103016. (2021)

2 Abbas, Farhat, et al.: Crop yield prediction through proximal sensing and machine learning algorithms.Agronomy 10.7:1046. (2020)

3 Anantha, K. H., et al.: Impact of best management practices on sustainable crop production and climate resilience in smallholder farming systems of South Asia.Agricultural Systems 194: 103276. (2021)

4 Palanivel, Kodimalar, and Chellammal Surianarayanan.: An approach for prediction of crop yield using machine learning and big data techniques.International Journal of Computer Engineering and Technology 10.3: 110-118. (2019)

5 Bian, Chaofa, et al.: Prediction of Field-Scale Wheat Yield Using Machine Learning Method and Multi-Spectral UAV Data.Remote Sensing 14.6 (2022): 1474. (2019)

6 Nishant, Potnuru Sai, et al.: Crop yield prediction based on indian agriculture using machine learning. 2020 International Conference for Emerging Technology (INCET),IEEE. (2020)

7 Mishra, Subhadra, Debahuti Mishra, and Gour Hari Santra.: Adaptive boosting of weak regressors for forecasting of crop production considering climatic variability: An empirical assessment.Journal of King Saud University-Computer and Information Sciences 32.8: 949-964. (2020)

8 Talaviya, Tanha, et al.: Implementation of artificial intelligence in agriculture for optimisation of irrigation and application of pesticides and herbicides.Artificial Intelligence in Agriculture 4: 58-73. (2020)

9 Van Klompenburg, Thomas, Ayalew Kassahun, and Cagatay Catal.: Crop yield prediction using machine learning: A systematic literature review.Computers and Electronics in Agriculture 177: 105709. (2020)

10 Pandith, Vaishali, et al.: Performance evaluation of machine learning techniques for mustard crop yield prediction from soil analysis.Journal of Scientific Research 64.2: 394-398. (2020)

11 Ogutu, Geoffrey EO, et al.: Probabilistic maize yield prediction over East Africa using dynamic ensemble seasonal climate forecasts.Agricultural and forest meteorology 250: 243-261. (2018)

12 Holzman, Mauro E., et al.: Early assessment of crop yield from remotely sensed water stress and solar radiation data.ISPRS journal of photogrammetry and remote sensing 145: 297-308. (2018)

13 Dash, Yajnaseni, et al.: Rainfall prediction for the Kerala state of India using artificial intelligence approaches.Computers & Electrical Engineering 70: 66-73. (2018)

14 Iu, Zhuo, et al.: Deep reinforcement learning with its application for lung cancer detection in medical Internet of Things.Future Generation Computer Systems 97:1-9. (2019)

15 Khodayar, Mahdi, Jianhui Wang, and Mohammad Manthouri.: Interval deep generative neural network for wind speed forecasting.IEEE Transactions on Smart Grid 10.4: 3974-3989. (2018)

16 R. Whetton, Y. Zhao, S. Shaddad, and A. M. Mouazen.: Nonlinear parametric modelling to study how soil properties affect crop yields and NDVI,Comput. Electron. Agricult., vol. 138, pp. 127–136, Jun. (2017)

17 Andrew Crane Droesch.: Machine learning methods for crop yield prediction and climate change impact assessment in agriculture, Published by IOP Publishing Ltd, vol. 05, OCT (2018)

18 W. Wieder, S. Shoop, L. Barna, T. Franz, and C. Finkenbiner.: Comparison of soil strength measurements of agricultural soils in Nebraska.J. Terramech., vol. 77, pp. 31–48, Jun. (2018)

19 Y. Cai, K. Gua.: A high-performance and in-season classification system of field-level crop types using time-series Landsat data and a machine learning approach,Remote Sens. Environ., vol. 210, pp. 35–47, Jun. (2018)

20 X. E. Pantazi.: Wheat yield prediction using machine learning and advanced sensing techniques,'' Comput. Electron. Agricult., vol. 121, pp. 57–65, Feb. (2016)

21 Bali, Nishu, and Anshu Singla.: Emerging trends in machine learning to predict crop yield and study its influential factors: a survey.Archives of computational methods in engineering 29.1: 95-112. (2022)