Air Quality Prediction Using Machine Learning Approaches.

[1]Dyana Priya, [2]Edara Uday Venkata Shanmuk, [3]Rudranadh Chinnam,
[4]Anthati Uday Goud, [5]Ankipalli Siddardha

[1]Asst. Professor, [2,3,4,5]Students
department of Computer Science and Engineering
Bharath Institute of Higher Education and Research
Chennai, India

***Abstract*—By utilizing machine learning to estimate the air quality index of a certain place, we forecast India's air quality. The air quality index of India is a commonly used indicator of pollution levels (so2, no2, rspm, spm, etc.) through time. Based on historical data from prior years and utilizing ML techniques, we created a model to forecast the air quality index for a certain forthcoming year. For our prediction problem, we use cost estimation to increase the model's efficacy. When given historical data on pollutant concentration, our model will be able to accurately estimate the air quality index for an entire county, any state, or any contiguous region. We improved performance over the baseline regression models in our model by applying the suggested parameter-reducing formulations. Our model has a 91% accuracy rate when used to estimate the air quality index for the entirety of India. We also utilise the AHP MCDM approach to determine the order of preference based on how closely it resembles the ideal solution. This research can assist in building MDS by utilising deep learning techniques like XGBoost, Random Forest (RF), and Convolution Neural Network (CNN), which can monitor the information entering mobile devices and separate out unlawful occurrences. With an accuracy rate of roughly 90%, CNN outperforms the other algorithms among the two techniques. Air quality prediction is carried out in our project.**

Keywords—pollutants, machine learning, random forest, AQI, CNN

I. INTRODUCTION (HEADING 1)

India, the world's fastest-growing industrial country, is creating record levels of pollutants, including Co2, PM2.5, and other dangerous airborne contaminants. According to the Indian air quality standard, contaminants are indexed according to their scale, and these air quality indexes show the amounts of significant pollutants in the atmosphere. Air quality of a specific state or country is a measure of the impact of pollutants on the respected regions. Many atmospheric gases harm our ecosystem by causing pollution. Every type of pollution has a unique index and scales at various degrees. The principal pollutants The data may be classified depending on the restrictions using individual AQIs like (no2, so2, rspm, and spm indices, for example). We gathered the information from the Indian government's database, which includes information on pollution concentrations that occur across the country. To establish the appropriate AQI for the area, we first calculate the individual pollutant index for each accessible data point. In order to estimate the air quality of India in any given place, we have developed a model that can predict the air quality index for each accessible data point in the dataset. We can identify the main pollutants that cause pollution and the areas of India that are most severely impacted by those pollutants by forecasting the air quality index. Using this forecasting model, different information about the data is collected using different methodologies to determine the regions that are most severely affected in a certain location (cluster). This provides more knowledge and information about the origin and age of the contaminants.

II. LITERATUE SURVEY

A. Air Quality Indices Prediction using Sugeno's Fuzzy Logic

Publisher: IEEE 2021

Rajan Kumar; Surendra Kumar; Eeshan Amiy.

In the built environment, air quality plays a significant role in maintaining residents' well-being, comfort, prosperity, and productivity. One of the most urgent issues facing India's major leading cities is the indoor air quality. 13 out of the 20 cities with the worst levels of air pollution are located in India, making air pollution a severe threat to the nation's health. Due to hazardous chemicals and other toxic compounds, indoor air pollution has a ten times larger influence on the deterioration of human life quality than outside air pollution. India's Central Pollution Control Board, which monitors air pollution, has issued several individual ratings. The current investigation is centred on Ranchi, Jharkhand's air pollution concentration.

B. Prediction of Air Quality Index Based on Improved Neural Network.

Publisher: IEEE 2021

Wang Zhenghua; Tian Zhihui

This research employs the upgraded BP neural network to create a prediction model of air quality index in order to achieve the prediction of city air quality status. The model addresses the issue that air quality has several influencing elements, is nonlinear, and difficult to predict by using the features of nonlinear fitting approximation of BP neural network. A genetic algorithm is employed to optimise, with the goal of addressing the issue of the BP neural network's sluggish convergence and ease of falling into local optimum solutions. This essay uses Xuchang City as an illustration. Findings indicate that the air quality index has an average relative inaccuracy of 22%. The degree of accuracy was 80.44%. The air quality accuracy rate is 82.5%.

C. Air Quality Prediction Of Data Log By Machine Learning

Publisher: IEEE 2021

Venkat Rao Pasupuleti; Uhasri; Pavan Kalyan; Srikanth; Hari Kiran Reddy

To ensure optimum air quality, the air quality monitoring system collects data from several places on different air contaminants. In the current situation, it is the pressing issue. The introduction of hazardous gases into the atmosphere from industrial sources, vehicle emissions, etc., pollute the air. In many major cities today, the air pollution level has exceeded the government-set air quality index value, and it has reached critical levels. It has a significant effect on a person's health. Machine learning technology has advanced to the point that it is now able to anticipate contaminants based on historical data. In this research, we provide a machine learning-based system that uses previous pollution data to predict future pollution data using a device that can accept current pollution as input. The detected data is stored for further analysis inside an Excel sheet. The data on the pollutants is gathered using these sensors on the Arduino Uno platform.

D. Prediction of Air Quality Index Using Supervised Machine Learning Algorithms

Publisher: IEEE 2021

Karlapudi Saikiran; Gottapu Lithesh; Birru Srinivas; S Ashok

In order to estimate the Air Quality Index, which is used to limit pollution and reduce serious health risks, this article applies a variety of machine learning techniques. The Air Quality Index displays the level of air pollution. Particulate matter, nitrous oxide (NO 2), sulphur dioxide (SO 2), and carbon monoxide are the main pollutants (CO). The air quality is forecasted using older approaches like probability and statistics, however these techniques are exceedingly difficult to predict. To get around problems with earlier methods, machine learning algorithms provide a superior strategy for forecasting air pollution levels. Random forest regression, support vector regression, and linear regression are a few examples of machine learning techniques. The root mean square error (RMSE) approach is used to assess the accuracy of various models.

III. Existing System

Few Existing ststems are based on Random Forest and XGB methodologies:

As an ensemble learning technique for classifying individual trees, random forests or random decision forests are used. We will partition our dataset into two samples, and each sample will be organized into a tree-shaped architecture. The tendency of decision trees to overfit their training set is corrected by random decision forests. The best voting method is used, producing the best-fitting tree. We can divide the dataset during training based on the best-fitting tree.

For regression and classification problems, gradient boosting is a machine learning approach that creates a prediction model in the form of a group of weak prediction models, often decision trees. The resultant technique, known as gradient boosted trees, typically beats random forest when a decision tree is the weak learner. Similar to previous boosting techniques, it constructs the model in stages, but it generalises them by enabling the optimisation of any differentiable loss function.

Disadvantages of existing system:

- The accuracy level is not very high.
- not applicable to all datasets.
- Precision and recall do not meet expectations.
- complex model.

IV. PROPOSED SYSTEM

Deep learning convolutional neural networks, often known as ConvNets, are a kind of deep neural networks that are most frequently used to analyse visual data. As a result of its shared-weights design and translation invariance properties, they are often referred to as shift invariant or space invariant artificial neural networks (SIANN). They may be used in a variety of fields, including image and video recognition, recommender systems, image classification, image segmentation, and medical image analysis. They can also be used in brain-computer interfaces, natural language processing, and financial time series. Multilayer perceptrons are regularised variants of CNNs. Fully linked networks, or multilayer perceptrons, are those in which every neuron in one layer is connected to every neuron in the following layer. These networks are vulnerable to overfitting data because of their "fully-connectedness." Adding some kind of magnitude measurement of weights to the loss function is a typical method of regularisation. CNNs tackle regularisation differently; they make use of the data's hierarchical structure to piece together more complicated patterns out of smaller, simpler ones. CNNs are therefore at the lower end of the connectivity and complexity spectrum.

Because of how closely the connection pattern between neurons mirrors the structure of the animal visual cortex, convolutional networks were inspired by biological processes. Only in the constrained area of the visual field known as the receptive field do individual cortical neurons respond to inputs. Different neurons' receptive areas partially overlap one another to fill the whole visual field.

Comparatively speaking to other image classification algorithms, CNNs employ a minimal amount of pre-processing. This implies that the filters, which were manually designed for traditional techniques, are learned by the network. This feature design's independence from past knowledge and human effort is a significant benefit.

Proposed System Advantages:

- Algorithm limits our search area.
- Reduced time / Estimation time is Low.

- Feature Selection with defined area.
- Improved accuracy.
- Can be implemented in all datasets.

V. System Implementation

Modules:
1. Input dataset
2. Analysis of size of data set.
3. Oversampling.
4. Training and Testing.
5. Apply algorithms.
6. Predict results.

Modules Description:

1) Dataset input: Datasets can be obtained from online data sources on the internet. In order to estimate the accuracy effectively, we need to gather a sizable amount of data.

2) Analysis of the data set: This section contains an examination of the dataset. While processing data, the size of the data is taken into account.

3) Oversampling (Using SMOTE): We have compiled a thorough history of the air pollution that has been produced in a certain location over an extended period of time.

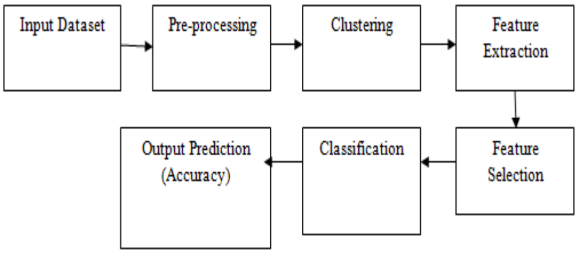
4) Training and Testing Subset: Many classifiers exhibit bias towards majority classes because the dataset is unbalanced. Minority-class characteristics are dismissed as noise and ignored. Thus, choosing a sample dataset is suggested.

5) Using the algorithm: The classification methods that were tested on the dataset for the sub-sample are listed below.

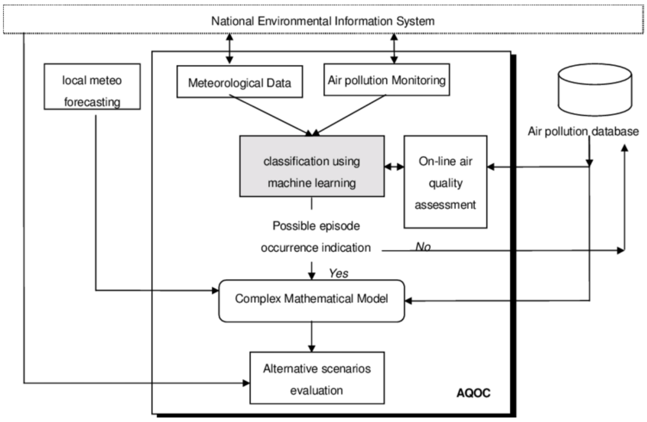
A. Random Forest (RF), B. Convolution Neural Network, and C. X Gradient Boost (CNN)

6) Making predictions about outcomes: The training model is used with the test subset. Accuracy is the metric that is utilized. The desired outcomes are obtained once the ROC Curve is displayed.

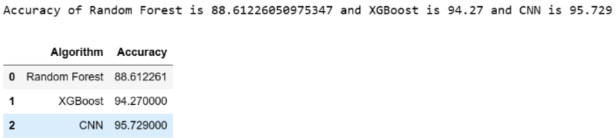
VI. System Architecture



VII. DATAFLOW DIAGRAM

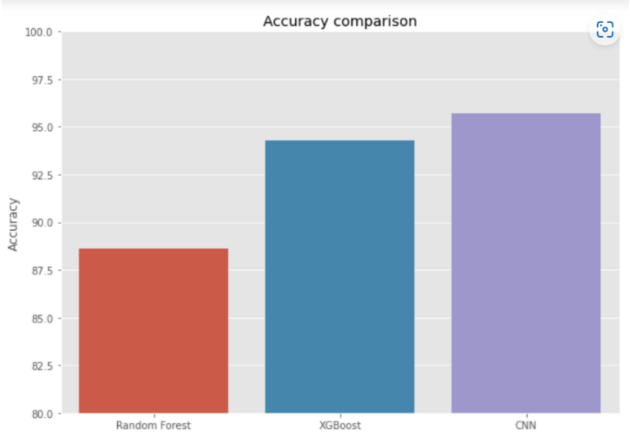

National Environmental Information System
local meteo forecasting
Meteorological Data
Air pollution Monitoring
Air pollution database
classification using machine learning
On-line air quality assessment
Possible episode occurrence indication
No
Yes
Complex Mathematical Model
Alternative scenarios evaluation
AQOC

VIII.Results and Discussion



```
Accuracy of Random Forest is 88.61226050975347 and XGBoost is 94.27 and CNN is 95.729
```

	Algorithm	Accuracy
0	Random Forest	88.612261
1	XGBoost	94.270000
2	CNN	95.729000



Our model performed better than Random Forest and XGBoost, our CNN model performed well, its accuracy stands at 95.7%.

IX. CONCLUSION

Our algorithm can effectively forecast the forthcoming air quality index of any specific data inside a specified location since it can anticipate the present data with 91% accuracy. With the help of this model, we can predict the AQI and warn the relevant regions of the nation. Additionally, because it uses a progressive learning algorithm, it is able to go back and pinpoint the exact location that needs attention, assuming that we have time series data for every potential affected region. The air quality data used in this project comes from the China air quality checking and investigation stage and includes the daily average fine particle problem (PM2.5), the inhalable particulate problem (PM10), the ozone problem (O3), CO, SO2, NO2 fixation, and the air quality record (AQI). The many sources of the poison focus as well as the factors that influence its fixation should be taken into consideration while assessing it. In terms of accuracy and precision, it was shown that the CNN approach used for classification in this study is more effective than the ones currently in use.

X. FUTURE SCOPE

In terms of accuracy and precision, it was shown that the CNN approach used for classification in this study is more effective than the ones currently in use. We can do for deep learning methodologies in the future. India's meteorological service aims to automatically determine if the eligibility procedure results in excellent or bad air quality (real time).
To automate this procedure by displaying the prediction outcome in a desktop or online application. to efficiently carry out the task in an environment using artificial intelligence.

REFERENCES:

[1] Verma, Ishan, Rahul Ahuja, HardikMeisheri, andLipikaDey. ” Air pollutant severity prediction using Bi-directional LSTM Network.” In 2018 IEEE/WIC/ACM International Conference on Web Intelligence (WI), pp. 651-654. IEEE, 2018.

[2] Figures Zhang, Chao, Baoxian Liu, Junchi Yan, Jinghai Yan, Lingjun Li, Dawei Zhang, XiaoguangRui, and RongfangBie. ”Hybrid Measurement of Air Quality as a 5 Fig. 8. RH w.r.t tin oxide Fig. 9. RH w.r.t C6H6 Mobile Service: An Image Based Approach.” In 2017 IEEE International Conference on Web Services (ICWS), pp. 853- 856. IEEE,2017.

[3] Yang, Ruijun, Feng Yan, and Nan Zhao. ” Urban air quality based on Bayesian network.” In 2017 IEEE 9th Fig. 10. RH w.r.t NO Fig. 11. RH w.r.t NO2 International Conference on Communication Softwareand Networks (ICCSN), pp. 1003-1006. IEEE,2017.

[4] Ayele, TemeseganWalelign, and RutvikMehta.” Air pollution monitoring and prediction using IoT.” In 2018 Second International Conference on Inventive Communication 6 Fig. 12. RH w.r.t Temperature Fig. 13. RH w.r.t CO and Computational Technologies (ICICCT), pp. 1741-1745. IEEE,2018.

[5] Djebbri, Nadjet, and MouniraRouainia. ” Artificial neural networksbased air pollution monitoring inindustrial sites.” In 2017 International Conference on Engineering and Technology (ICET), pp. 1-5. IEEE,2017.

[6] Kumar, Dinesh. ” Evolving Differential evolution method with random forest for prediction of Air Pollution.” Procedia computer science 132 (2018): 824-833.

[7] Jiang, Ningbo, and Matthew L. Riley. ” Exploring the utility of the random forest method for forecasting ozone pollution in SYDNEY.” Journal of Environment Protection and Sustainable Development 1.5 (2015): 245-254.

[8] Svetnik, Vladimir, et al.” Random forest: a classification and regression tool for compound classification and QSAR modeling.” Journal of chemical information and computer sciences 43.6 (2003): 1947-1958.

[9] Biau, GA˜ Srard. ” Analysis of a random forest model.” ˇJournal of Machine Learning Research 13. Apr (2012): 1063- 1095.

[10] Biau, Gerard, and ErwanScornet. ” A random forest ´ guided tour.” Test 25.2 (2016): 197-227.

[11] Grimm, Rosina, et al.” Soil organic carbon concentrations and stocks on Barro Colorado Island— Digital soil mapping using Random Forests analysis.” Geoderma 146.1- 2 (2008): 102-113.

[12] Strobl, Carolin, et al.” Conditional variable importance for random forests.” BMC bioinformatics 9.1 (2008): 307.

[13] Svetnik, Vladimir, et al.” Random Forest: a classification and regression tool for compound classification and QSAR modeling.” Journal of chemical information and computer sciences 43.6 (2003): 1947-1958.

[14] Verikas, Antanas, AdasGelzinis, and MarijaBacauskiene. ” Mining data with random forests: A survey and results of new tests.” Pattern recognition 44.2 (2011): 330-349.

[15] Ramasamy Jayamurugan,1 B. Kumaravel,1 S. Palanivelraja,1 and M.P. Chockalingam2 International Journal of Atmospheric Sciences Volume 2013, Article ID 264046, 7 pages http://dx.doi.org/10.1155/2013/264046

[16] V. M. Niharika and P. S. Rao, "A survey on air quality forecasting techniques," International Journal of Computer Science and Information Technologies, vol. 5, no. 1, pp.103-107, 2014. 2.

[17] NAAQS Table. (2015). [Online]. Available: https://www.epa.gov/criteria-air-pollutants/naaqs-table 3.

[18] E. Kalapanidas and N. Avouris, "Applying machine learning techniques in air quality prediction," in Proc. ACAI, vol. 99, September 2017.

[19] Questioning smart urbanism: Is data-driven governance a panacea? (November 2, 2015). [Online]. Available: http://chicagopolicyreview.org/2015/11/02/questioningsmart-urbanism-is-data-driven-governance-a-panacea/.

[20] D. J. Nowak, D. E. Crane, and J. C. Stevens, "Air pollution removal by urban trees and shrubs in the United States,"Urban Forestry &Urban Greening, vol. 4, no. 3, pp. 115-123, 2014`.

[21] T. Chiwewe and J. Ditsela, "Machine learning based estimation of Ozone using spatio-temporal data from air quality monitoring stations," presented at 2016 IEEE 14th International Conference on Industrial Informatics (INDIN), IEEE, 2016.

[22] Y. Zheng, X. Yi, M. Li, R. Li, Z. Shan, E. Chang, and T. Li, "Forecasting fine-grained air quality based on big data,"in Proc. the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 2267-2276, August 10, 2015