2023 International Conference on Recent Advances in Electrical, Electronics, Ubiquitous Communication, and Computational Intelligence (RAEEUCCI) | 979-8-3503-3742-6/23/$31.00 ©2023 IEEE | DOI: 10.1109/RAEEUCCI57140.2023.10134166

# Air Quality Index (AQI) Prediction using Automated Machine Learning with TPOT-ANN*

1st Meghna Sharma
*Department of Computing Technologies*
*SRM Institute of Science And Technology*
Chennai, India

2nd Eishita Gupta
*Department of Computing Technologies*
*SRM Institute of Science And Technology*
Chennai, India

3rd D.Viji
*Department of Computing Technologies*
*SRM Institute of Science And Technology*
Chennai, India

***Abstract*—Pollution is a critical and disturbing problem that people encounter daily in today's world and also has an impact on the quality of air. The issue is so crucial that it cannot be overlooked and its effects are felt everywhere. The climatic variables that affect the AQI, such as NO2, NH3, SO2, CO, O3, fog ,temperature, smoke ,dew ,mist, benzene, toluene , xylene, etc.The AQI measures the severity of the pollutants present in the air. It classifies the severity of air quality into six categories, each with its own range of values. The categories are as follows: Good, which ranges from 0-50 on the AQI scale, indicating that the air quality is generally safe and healthy for everyone to breathe. Moderate, which ranges from 51-100 on the AQI scale, indicating that the air quality is acceptable but may pose a moderate risk for certain individuals, such as those with respiratory issues. Unhealthy for Sensitive Groups, which ranges from 101-150 on the AQI scale, indicating that the air quality is dangerous for certain individuals, for example the youth or the younger and older ones or people having respiratory problems. Unhealthy, which ranges from 151-200 on the AQI scale, indicating that the air quality is hazardous and can cause serious health problems for everyone. Very Unhealthy, which ranges from 201-300 on the AQI scale, indicating that the air quality is extremely dangerous and can cause severe respiratory and cardiovascular problems. Hazardous, which ranges from 301 and higher on the AQI scale, indicating that the air quality is life-threatening and can cause serious health problems even for those who are otherwise healthy. Overall, the AQI is an essential tool for assessing the severity of air pollution levels and determining the appropriate measures that need to be taken to protect public health.The suggested model aims to evaluate the air quality. The proposed model suggests a strategy for measuring future AQI data from the present and historical AQI data by using automated machine learning techniques. Threshold value might be specified as a similar parameter since TPOT increases the iterations in number, which increases the depth of the node. The data on air pollutants is obtained from the sensors, processed according to a single schema, and then saved as a dataset. This dataset has undergone many preprocessing operations, including normalization, discretization and attribute selection.The machine learning system would learn from the data (pertaining to point in the time) and database to offer the user with comparable statistics to minimize processing time and increase platform efficiency.**

***Index Terms*—air quality index, automated machine learning, Tree-based pipeline optimisation tool (TPOT), normalization, deep learning, local binary pattern**

## I. INTRODUCTION

Monitoring air pollution, identifying the zone which is hazardous, and predicting future air quality have lately grown to be major concerns for many academics. Air pollution has reached a deadly level as a result of increased mobility, accelerating global warming, and abrupt climatic shifts. The negative impact of poor air quality on people's wellbeing has made accurate and timely air quality index (AQI) forecasting important.In numerous places where air quality levels are well above WHO standards, almost 91% of the world's population has recently resided[12]. Surprisingly, around 7 million people die every year due to polluted causes globally, including lung cancer, chronic respiratory disorders, heart disease, and stroke.Therefore, it is very important that all the parameters should be considered before calculating the AQI prediction.

In order to calculate the Quality of air, there are some atmospheric factors which need to be taken into consideration such as the direction of the wind or the humidity and sometimes the temperature. Additionally, environmental factors such as smoke, fog, mist, and dew can also impact the atmospheric conditions. One of the most popular metrics used in many nations to gauge how polluted a particular location in a city or the countryside is the air quality index (AQI). In general, the public health becomes riskier the higher the AQI number. As a result, people frequently use the AQI level to determine the local air quality so that they can take the necessary precautions to protect their families.As the AQI value increases, the level of pollution in the air and the associated health risks also increase. For instance, a score of 50 or below on the AQI scale signifies good air quality, while a score above 300 indicates the air quality as hazardous. The AQI consists of six categories, each with its own level of health risk and associated color code. By observing the color code, people can quickly determine if the air quality in their area has reached harmful levels.

Pollutants can be dispersed with the aid of wind. Pollutants that linger over a region may be spread out by the wind, which lowers the concentration of more strong pollutants in any one area. As happened when smoke from wildfires in the west of the United States carried particle contaminants as far as western Europe, this may likewise carry pollutants far from their source[3]. Areas with stronger winds usually exhibit lower levels of air pollution because the higher wind speeds disperse the pollutants more efficiently. As the earth heats up during the day, the air becomes more turbulent, leading to the scattering of air contaminants. Because circumstances are more steady at night because the air is colder, contaminants have a tendency to spread less.The amount of hazardous or dangerous substances in the air rises with high humidity. Additionally, it lowers the air quality in our houses by bringing in dust mites. Small, airborne particles that are 2.5 microns in dimension and invisible to the human eye make up particulate matter[17]. Dust, pollen, ash, germs, smoke, soot, aerosol droplets, and other materials can be used to create these. These tiny particles can all readily penetrate our natural defenses and enter your lower respiratory system through deep inhalation. Because of high and low humidity, viral and bacterial organisms which cause respiratory illnesses flourish. Germs in the air are also caused by low humidity.By initially sorting the nation, state, and city, the apps allow users to locate the Air Quality Index (AQI). In order to reduce processing time and improve platform effectiveness, the machine learning system would learn from the data (relative to a certain period) and database. It would then provide the user with similar statistics.

The following are the primary goals of the suggested model

- Prediction of the AQI around the globe.
- To determine if air is Good ranging from 0 to 50, Medium ranging from 51 100, Sensitive for unhealthy group (101 to 150), Unhealthy ranging from 151 200, Very Unsafe or unhealthy ranging 201 to 300, and perilous ranging from 301 and above.

## II. STATE OF THE ART (LITERATURE SURVEY)

UAVs levitate at high altitudes and have great mobility, according to Y. Liu, J. Nie, W. Y. B. Lim, X. Li, S. H. Ahmed, and C. Miao; as a result, communications between UAV and ground base stations may have certain problems. Therefore, it is difficult to carry out learning-related activities utilizing centralized systems, especially when transferring a significant volume of data wirelessly [7]. Public and private organizations collaborate to develop AQI through crowd-sourcing, monitor the models by sharing the data which is gathered.. However, due to privacy concerns, the global data protection law (GDPR) forbids direct exchange of user data between agencies, creating a situation known as data islands. Consequently, it's important to ensure precise AQI monitoring while maintaining privacy[20].

It is proposed that LSTM and GRU be used to solve the "exploding gradient" and "vanishing gradient" difficulties[4]. Although the gradient problem of RNN has been partially resolved by LSTM and GRU, it is nonetheless insufficient. When working with length sequences, they are still unable to understand the relationships between the data more thoroughly. These gated units do, however, have certain processing flaws with time series data. The significance of meteorological data characteristics varies according to the temporal dimension, and the features concealed in aberrant data are more useful than those in data of meteorological[16] which is normal. Standard LSTM and GRU struggle to capture this type of variance in temporal aspects.

To undertake the prediction of wind speed which is short-term, an UKF technique with SVR was used[16]. The hybrid UKF-SVR model, however, outperformed the other four models in terms of predicting performance. To combine the pseudo-range, ranging information, for localization the information of location, two popular Kalman filtering techniques—EKF and UKF—have been used. The experiment findings revealed that when Kalman Filter framework was used, the placement performance of the nodes significantly improved. In order to estimate traffic flow with less input data, Kumar developed the Kalman filtering method (KFT)[3]. Due to the absence of sufficient data, the accuracy of the outcome is demonstrated.

In order to explore the early stages of the variance rule of concentration of pollutant, individuals employed classic statistical methods like multiple regression and the AQI to measure the concentration of different contaminants in the air [15]. To fit the model of the change in $O3$ concentration, several strategies employed air contaminants and numerous climatic conditions as variables using multiple regression analysis techniques. The results of the experiment demonstrated how well this approach predicted $O3$ concentration. To address the issue of multicollinearity across dependent variables, the study used PCA and MLR[21].One of the most popular ways to explain how dependent variables rely on a number of independent variables is the MLR model. However, this model's limitation to addressing linear problems with modest data fluctuation ranges or those with less severe nonlinearity was a drawback[6].The issues which were in the majority were nonlinear and needed to be addressed immediately, but the MLR model was unable to construct nonlinear problems with a high degree of fitting.

[10] proposes an approach for big data-based air simulation; the authors compare MapReduce Hadoop and Spark for air quality modeling. They used the Texas 179 sensor dataset and discovered performance gains of 20–25% for Spark over MapReduce. Although real-time decision-making has been stated, the forecast accuracy was left out. In [21], another Random Forest-based AQI prediction system utilizing Spark dispersed across many clusters is presented. Although Random Forests may be used to categorize data, the technique is not employed to analyze time series data in real-time. Currently, many systems employ supervised learning techniques to train the dataset, which can be a drawback as the data is continually changing. Previous methods were either laborious or ineffective. However, the suggested model will leverage automated learning by creating and evaluating a pipeline, randomly adjusting certain components of the pipeline, and continuously searching for better algorithms. By considering all the significant parameters, such as Nitrogen Oxide, Nitrogen Dioxide, Carbon monoxide, Ammonia, Sulphur Dioxide, Ozone, Benzene, Toluene, and Xylene, the model will close the gap between data and accurately predict the air quality index.

## III. PROPOSED WORK

The proposed model aims at determining the AQI of the particular city. It is built using JavaScript. The user first searches for the city for which AQI has to be determined by typing the city name in the search bar. Then it displays the AQI and with date and time, also, the parameters like CO, NO, Nitrogen dioxide, Ozone, SO2, Fine particles matter, Ammonia. At the backend, the model is merged, where the data analysis is performed which is data preprocessing which is done in four stages that are Data Cleaning, Data Integration, Data Transformation, Data Reduction. Log Transformation and Linear Discriminant Analysis is used in Data Transformation and Data reduction respectively. Now, the processed data is now taken by TPOT-ANN model to extract the Feature that is done in parallel, feature selector, feature preprocessing, feature transformation. Now it is passed to MLP and later on the features are combined and hence the prediction is made. The response is given back to the user.

For data preprocessing, log transformation and linear discriminant analysis are employed for data transformation and data reduction, respectively.For data analysis, the dataset is worked upon by cleaning of the data using Kalman Filtering Algorithm. Kalman filtering, alternatively referred to as linear quadratic estimation (LQE), is a smart technique that makes use of a range of measurement methods, including statistical noise and other types of inaccuracies, to make predictions about the probable distribution of variables across different time frames. By integrating information from multiple sources, this method is capable of generating highly accurate estimates of unknown variables, surpassing what can be achieved through a single measurement alone. In other words, Kalman filtering allows us to extract meaningful insights from noisy data by extracting useful signals from the noise. This approach is highly effective in a wide range of fields, including engineering, finance, and science, and has a proven track record of improving predictions and decision-making processes. In summary, Kalman filtering is a powerful tool that enables us to extract valuable insights from complex and noisy data sets, ultimately leading to more accurate predictions and improved outcomes.

The fig.1 shows the architecture diagram of the prediction model for the front-end website where the user interacts with the website and enters the city name, the results are displayed.The user enters the name of the place for which they need to obtain the AQI for, once city is entered the database searches for the past data inorder to predict the AQI and the result is displayed on the screen to the user.

The fig 2 shows the backend model of tpot-ann where MLP is used. Here, the data is scanned and preprocessed wherein the data is normalized and processes like feature selection, feature preprocessing and feature transformation is performed and the features based on the similarities are grouped together using Multi-layer Perceptron and therefore class prediction is done.

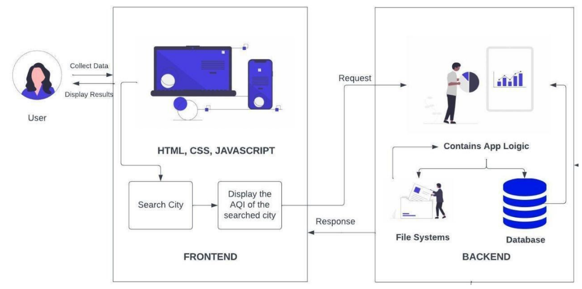


Fig. 1. Architecture diagram for AQI prediction-frontend

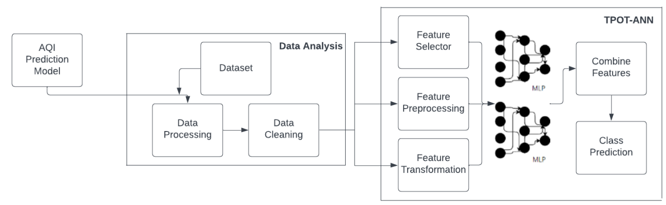


Fig. 2. Architecture diagram for AQI prediction-backend

## IV. IMPLEMENTATION

In order to implement the model, it is categorized into three modules: Data Analysis, Feature Engineering and Model Building and Prediction.These modules involve gathering various datasets in accordance with the parameters to obtain the desired results, choosing the most suitable algorithm—in this case, Tree-Based Pipeline Optimization Tool: Regressor—and then creating the model and predicting the accuracy with the assistance of the aforementioned algorithm and the neural networking.Artificial neural networks (ANNs) and automated machine learning (AutoML) have transformed the study of artificial intelligence by producing models that are very effective at solving a variety of inductive learning problems.

### A. Data Analysis

The proposed model aims at the globalization of the data. In this model, the user can find the Air Index Prediction of any place around the globe. One can find the AQI of Paris while sitting in Chennai. The following parameters influence the amount of pollutants in the ambient air : atmospheric wind speed, wind direction, relative humidity, temperature, smoke, fog, mist, and dew. The existing model has localized search pertaining to specific areas, say Delhi, for finding the AQI. The more elaborate dataset is constructed where data from various dataset are compiled to give a broader spectrum. The process of compiling data involves more than merely adding up the survey items. To raise acquired data to the level of the desired statistical output, statistical offices execute a number of checks, validations, and statistical processes. Large random mistakes made by respondents are often detectable through plausibility tests on the data, such as by comparing the reported statistics with prior values or the reported ratios with appropriate boundaries for the various sorts of operations.

The fig 3. shows the density of the data with respect to PM 2.5. The graph shows how density varies with the fine particulate matter.As the particle in air increases, the density of the air and the pressure increases thus increasing the temperature of the surrounding environment.

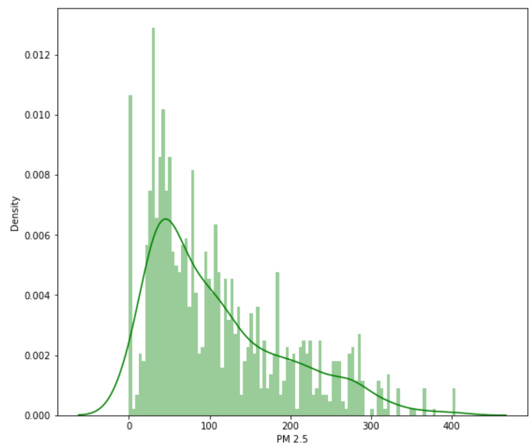


Fig. 3. Shows the density of data with respect to PM2.5

In order to compile the various dataset into a custom made, following datasets are used: AQI - Air Quality Index by Azmine Toushik Wasi, Global Air Pollution Dataset by Hasib Al Muzadid, Air Quality Data in India (2017 - 2022) by Fedesonario etc.

The Fig 4 shows the insight of the data preprocessing such as Data Cleaning, Data Integration, Data Transformation and Data Rotation.In this process, the data is searched upon to find the most appropriate results so that the desired outcome is achieved based on the requirements entered by the user.

The Fig 5 shows the Data cleaning process which has missing values, noisy data, inconsistent data. It aims at transforming the data based on the similarities so that the performance can be enhanced.

The Fig 6 shows the process involved in Data Transformation such as generalization and Normalization.Here the data is categorised into Generalisation and Normalisation.In this process of data summarization, higher level ideas are used to replace relatively low level values.

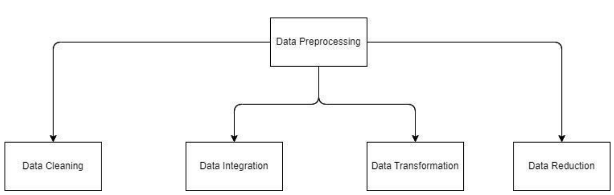


Fig. 4. Data analysis of dataset

Using CSV files to store large data sets is a simple and effective method.

- Use pandas to load the CSV into a DataFrame.
- The dataset's maximum number of incomplete data is then returned after missing values are dropped.

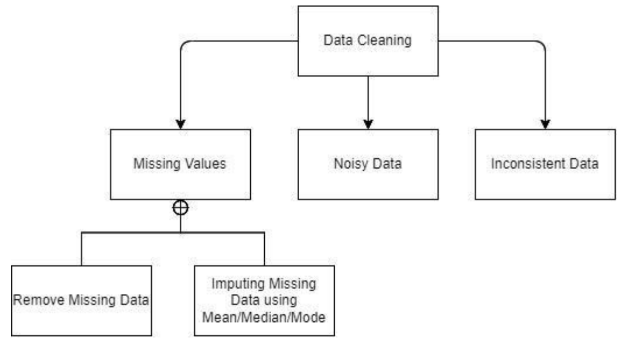


Fig. 5. Data cleaning of dataset

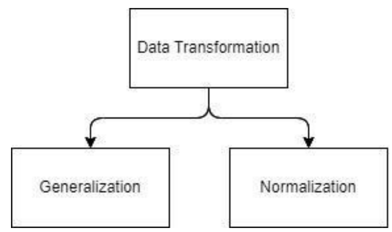


Fig. 6. Data transformation for dataset

- Determine the number of unique items now.
- A Seaborn library called Distplot illustrates the diversity in the data distribution. The distplot displays the data's univariate distribution (Distribution of a variable against the density distribution).
- To determine if the distributions of the two variables, i.e., theoretical amount and ordered value, are comparable or not with regard to the locations, a quantile-quantile plot is now created.
- The data is typically distributed if it falls close to or on the line; otherwise, pre-processing, reciprocal transformation, or log transformation is required.

There are two types of datasets used in this model, Air Quality Data in India and Air Quality Index from various locations of India. One data set contains, city, date, pm2.5, NO, NO2, NOx, NH3, CO, SO2, O3, Benzene, Toluene, Xylene and the other dataset contains the following data, id, country, state, city, pollutant id, pollutant min, pollutant max, pollutant avg, pollutant unit.

The fig 7 shows the visual representation of all the features listed in the dataset used for the model.The major pollutants such as NO, NO2, NOx, NH3, CO, SO2, O3, Benzene, Toluene, Xylene etc are plotted against the density in order to see the variation on AQI and factors affecting it.

### B. Feature Engineering

The automated machine learning tool in Python called Tree-based Pipeline Optimisation Tool optimizes the machine learning pipeline using the principles of genetic programming. By intelligently investigating thousandths of the conceivable to discover the finest potential parameter that matches the data, it automates the most laborious element of machine learning. It

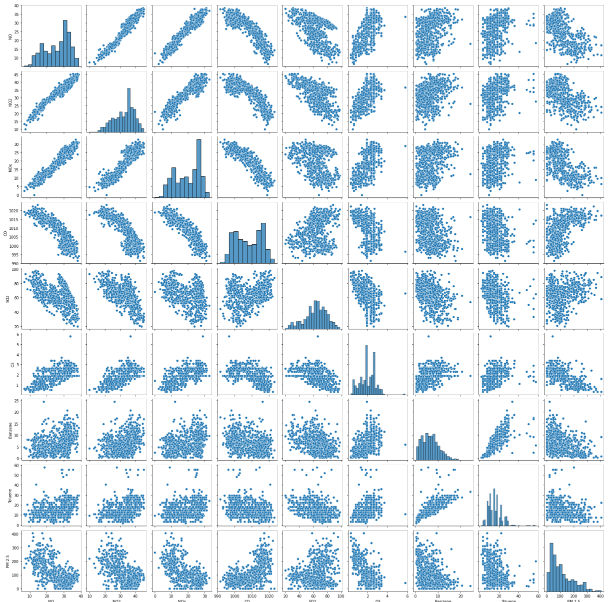

Fig. 7. Visual representation of all the features listed in the database

then defers to more restricted parameter tuning strategies like grid search for the final fine-tuning. TPOT aids in the discovery of effective algorithms.TPOT uses tree concepts during genetic programming.

The Fig 8 shows the graph plotted between the ordered values and the theoretical values for the features PM 2.5 present in the dataset used.

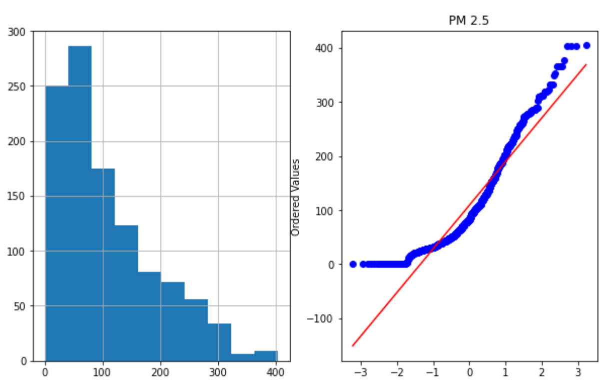


Fig. 8. Graph plotted between the ordered values and the theoretical values for the feature PM 2.5 present in the dataset used.

The TPOT process involves three stages. Firstly, an algorithm is selected based on its potential to provide the best results. Then, the selected algorithms are cross-bred to produce a hybrid solution. As time passes, these algorithms undergo mutations and improvements, resulting in the most optimal solution.

The fig 9 shows the feature engineering process such as preprocessing of features or the construction of features or the selection of features.Using statistical or machine learning techniques, feature engineering is the process of transforming unprocessed observations into desired features.

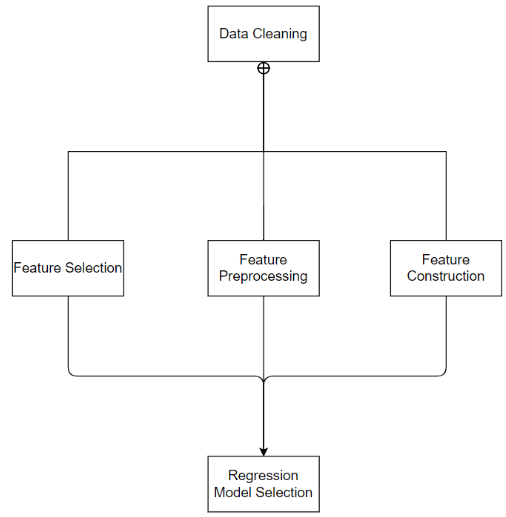


Fig. 9. Feature engineering for AQI prediction

In order to enhance the model, a new version of TPOT, called TPOT-regressor, has been developed. This new version utilizes an intelligent search method across various machine learning pipelines that include supervised regression models, preprocessors, feature selection strategies, and other transformers that conform to the scikit-learn API. AutoML algorithms, like TPOT-regressor, incorporate several machine learning algorithms, such as random forests, linear models, and SVMs, along with multiple preprocessing techniques, such as missing value imputation, scaling, PCA, and feature selection, to build accurate predictive models. These algorithms also optimize the hyperparameters of the models and preprocessing steps, and offer various options for ensembling or stacking the algorithms to create more robust models. Additionally, AutoML algorithms generate informative visualizations, such as heat maps, to illustrate the correlation between different variables, helping users to better understand the relationships between different data points. Overall, AutoML algorithms are a powerful tool for automating the process of building predictive models, enabling researchers and practitioners to quickly and easily gain insights from complex and noisy datasets.

The fig 10 displays a map of heat of the coefficients which allows for the visualization of the strength of correlation among variables.

### *C. Model Building and Prediction*

The TPOT program now includes a new feature that utilizes artificial neural network (ANN) technology. Operators in TPOT can either be natively implemented or imported from other developer tools, such as XGBoost and Scikit Learn, and each operator has controllable hyperparameters that can be adjusted as TPOT learns. TPOT will consider a total of

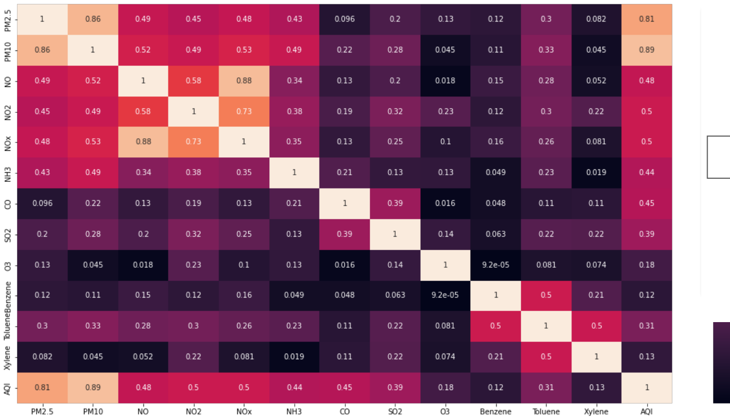


Fig. 10. Heat map to visualize the strength

100 pipeline configurations using default parameters, which is equivalent to a grid search with 100 different hyperparameter combinations for a machine learning system. TPOT-NN adds two additional classification estimators, a LR and a MLP, that are accessible to TPOT during the learning process. The computational complexity of transitional network layers, the convex optimization technique used in learning rate, and other variables of the LR and MLP estimators are optimized by TPOT using genetic programming (GP), which is a systematic method for enabling computers to solve problems and create new generations of programs. The dimensionality of the hidden layer in TPOT-NN estimators is optimized by GP, but the possibilities are determined by the count of features, which limits the prediction ability of individual layers and restricts TPOT from creating large networks that would be for the learning with respect to the TPOT's GP algorithm. In automated Machine learning, the TPOT works in the following process, firstly the raw data is send to the data cleaning and then the feature selection, feature preprocessing, feature construction takes place in parallel and the data is then send to to the process selection and then the parameter optimization takes place and at the end the model is validated. The visual representation of the process of TPOT is shown below.

The 11 fig shows the process for automated machine learning using Tree pipeline based optimization technique.By discovering unique pipeline operators, this technique may locate artificial feature builders that can improve classification accuracy in extremely demanding ways. To create a sequence of operations acting on the provided dataset, the TPOT operators are chained together.

The hybrid model is the combination of Artificial Neural Networks and TPOT, firstly below is the diagram of the Artificial Neural Network working.

The fig 12 shows the existing model of Artificial neural network.The data is trained using the Multi-Layer Perceptron and the class is predicted using the training data.

Here the diagram is the visual representation of the Neural Network process and procedure for working. Firstly, the data is trained and sent to the MLP and the prediction is done, it

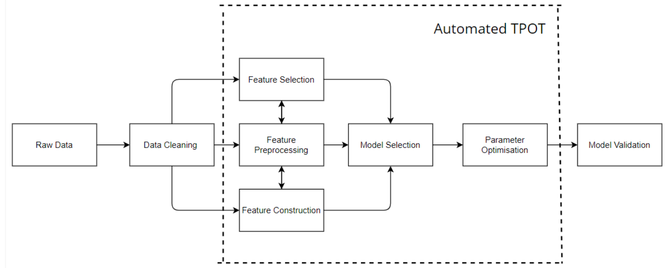


Fig. 11. Automated machine learning using TPOT

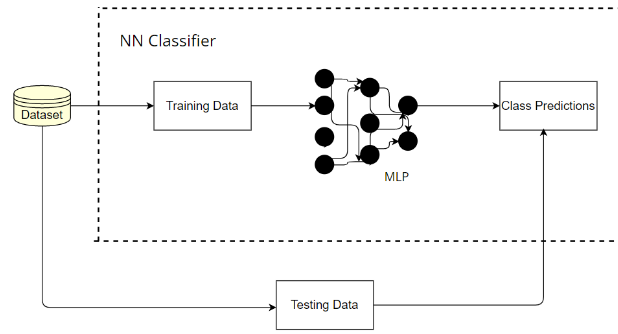


Fig. 12. Existing method of NN classifier.

is a feed-forward process. Now, the TPOT model is shown below.

The fig 13 shows the existing model of TPOT classifier. The data is trained then feature is extracted using a Regression model where the features based on similarities and dissimilarities are combined and the class is predicted by training the data simultaneously.

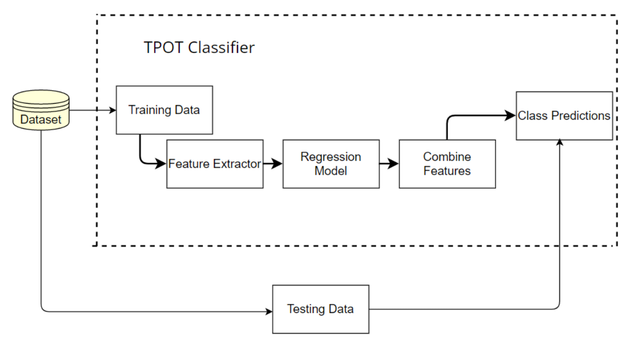


Fig. 13. Existing methods of TPOT models

Here, the working of automated machine learning using TPOT is shown.The data preprocessing is done after the data is trained and features are extracted and later the model is predicted and features are combined after which the accuracy is determined by the model.

The fig 14 shows the proposed hybrid model of TPOT and Artificial Neural Network. The data is obtained from the dataset and basic feature engineering processes are performed and is processed through various MLP model designed on different parameters and the features are combined and the class is predicted and the dataset is trained for better pipe lining of the vectors.

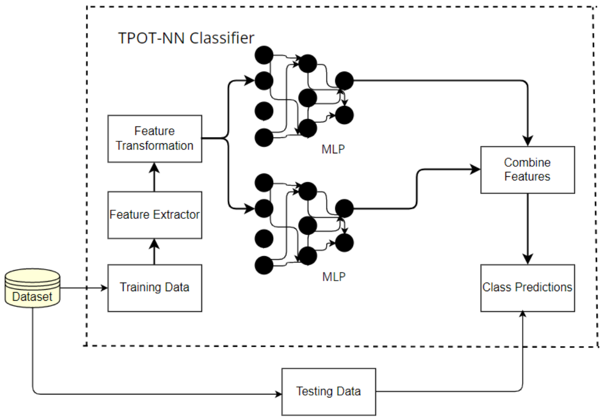


Fig. 14. Proposed TPOT-ANN model

It is the hybrid classifier of TPOT and ANN, here the MLP is taken into account after the features are transformed and after it is processed by MLP, the features are combined and the prediction is done. The table 1 shows the count, mean, std, min and max for the parameters present in the dataset used.

TABLE I
PARAMETERS PRESENT IN THE DATASET

| Stats | Parameters | | | | |
|---|---|---|---|---|---|
| | ***PM2.5*** | ***NO*** | ***NH3*** | ***CO*** | ***SO2*** |
| Count | 176502 | 176502 | 176502 | 176502 | 176502 |
| Mean | 0.1052 | 0.0672 | 0.0831 | 0.0127 | 0.0716 |
| Std | 0.1015 | 0.0948 | 0.0560 | 0.0354 | 0.0744 |
| min | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 |
| 25% | 0.0303 | 0.0122 | 0.0549 | 0.0034 | 0.0312 |
| 50% | 0.0764 | 0.0288 | 0.0783 | 0.0061 | 0.0583 |
| 75% | 0.1414 | 0.0748 | 0.0921 | 0.0099 | 0.0824 |
| max | 1.0000 | 1.0000 | 1.0000 | 1.0000 | 1.0000 |

## V. RESULTS AND DISCUSSION

In this research paper, the existing models like Linear Regression, XGBoost Regression model and Random Forest accuracy are compared below. The scatter plot is shown below, the scatter plot is to show how much one variable is affected by another.

### A. Linear Regression

The accuracy achieved by using Linear Regression is 48.52%. The fig 15 shows the accuracy of the LR model.

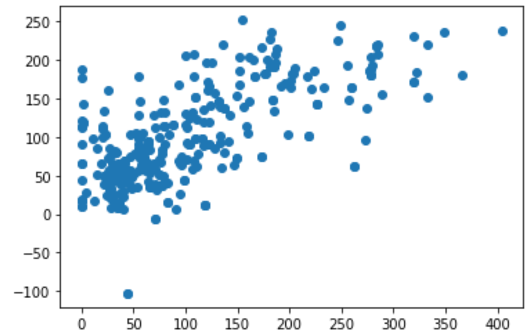


Fig. 15. Accuracy of LR model

### B. Random Forest

The accuracy achieved by using Random Forest is 76.75%. The fig 16 shows the accuracy of the RF model.

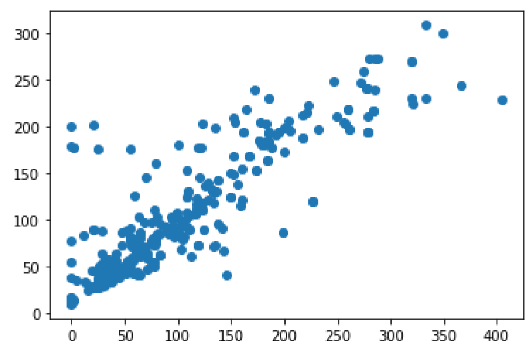


Fig. 16. Accuracy of RF model

### C. XG Boost Regression Model

The accuracy achieved by using XG Boost Regression is 78.59%. The fig 17 display the accuracy of the XG Boost R model.

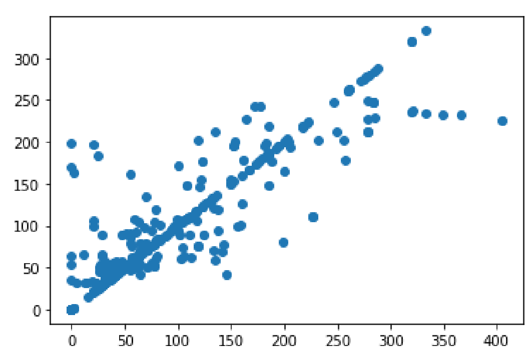


Fig. 17. Accuracy of XGB model

### D. TPOT Model

The accuracy of Automated machine learning using TPOT is 82.28%. The fig 18 shows the accuracy of the TPOT model.

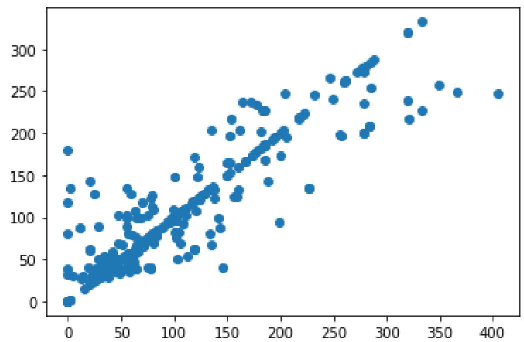


Fig. 18. Accuracy of TPOT model

This section applies a combination of the Tree-Based Pipeline Optimization Tool (TPOT) and Artificial Neural Network as a hybrid model to improve the model's accuracy and efficiency. After the data analysis, using TPOT the feature selector, preprocessing and transformation takes place in parallel and then data is passed through MLP Multi-layer Perceptron classifier and later features are combined and the class is predicted. The accuracy record with the hybrid model of TPOT and ANN is 94.22 %. The fig 19 shows the accuracy of the TPOT-ANN model.

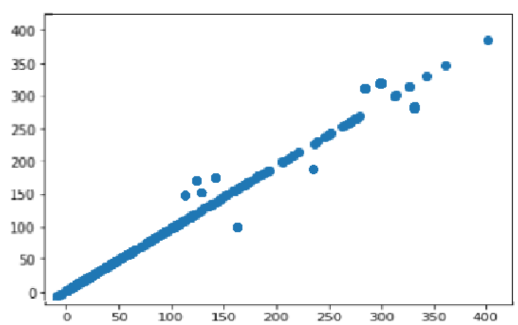


Fig. 19. Accuracy of TPOT-ANN model

The table 2 shows the comparison model for Linear regression, Random forest, XG Boost Regression model, TPOT Model to the TPOT-ANN model.

## VI. CONCLUSION

The hybrid model of the Tree-Based Pipeline Optimization Tool (TPOT) and Artificial Neural Network is used in this section to improve the model's efficacy and accuracy. The input is then fed through the MLP Multi-layer Perceptron classifier utilizing TPOT as the feature selector, where preprocessing and transformation are carried out concurrently. Finally, features are aggregated, and the class is predicted. The TPOT and ANN hybrid model's accuracy rate is 94.22

TABLE II
ACCURACY OF DIFFERENT MODELS

| Algorithm | Accuracy |
|---|---|
| Linear Regression | 48.52% |
| Random Forest | 76.75% |
| XG Boost Regression Model | 78.59% |
| TPOT Model | 82.28% |
| TPOT-ANN Model | 94.22% |

Air quality Prediction is quite a challenging task due to sudden changes in the environment i.e, the dynamic environment, variability, pollutants. In developing countries, quality analysis and monitoring is difficult because of the consequences of air pollution on climate, humans, plants, etc. Automated ML and ANNs are very useful for solving a wide variety of inductive learning tasks. In the TPOT, the use of the estimators ANNs increases the training time in few pipelines, which is useful in some situations. The accuracy was compared with all the other models and the result was the best in the case for the hybrid model of TPOT and ANN.

## REFERENCES


[1] R. Purbakawaca, A. S. Yuwono, I. D. M. Subrata, Supandi and H. Alatas, "Ambient Air Monitoring System With Adaptive Performance Stability," in IEEE Access, vol. 10, pp. 120086-120105, 2022, doi: 10.1109/ACCESS.2022.3222329.

[2] J. Wang, L. Jin, X. Li, S. He, M. Huang and H. Wang, "A Hybrid Air Quality Index Prediction Model Based on CNN and Attention Gate Unit," in IEEE Access, vol. 10, pp. 113343-113354, 2022, doi: 10.1109/ACCESS.2022.3217242.

[3] R. Kumar, S. Kumar and E. Amiy, "Air Quality Indices Prediction using Sugeno's Fuzzy Logic," 2022 International Conference on IoT and Blockchain Technology (ICIBT), Ranchi, India, 2022, pp. 1-5, doi: 10.1109/ICIBT52874.2022.9807786.

[4] H. Chen, M. Guan and H. Li, "Air Quality Prediction Based on Integrated Dual LSTM Model," in IEEE Access, vol. 9, pp. 93285-93297, 2021, doi: 10.1109/ACCESS.2021.3093430.

[5] M. Benhaddi and J. Ouarzazi, "Multivariate time series forecasting with dilated residual convolutional neural networks for urban air quality prediction", Arabian J. Sci. Eng., vol. 46, no. 4, pp. 3423-3442, Apr. 2021.

[6] A. Tripathy, D. Vaidya, A. Mishra, S. Bilolikar and V. Thoday, "Analysing and Predicting Air Quality In Delhi: Comparison of Industrial and Residential Area," 2021 International Conference on Smart Generation Computing, Communication and Networking (SMART GENCON), Pune, India, 2021, pp. 1-6, doi: 10.1109/SMARTGENCON51891.2021.9645787.

[7] Y. Liu, J. Nie, X. Li, S. H. Ahmed, W. Y. B. Lim and C. Miao, "Federated Learning in the Sky: Aerial-Ground Air Quality Sensing Framework With UAV Swarms," in IEEE Internet of Things Journal, vol. 8, no. 12, pp. 9827-9837, 15 June15, 2021, doi: 10.1109/JIOT.2020.3021006.

[8] K. Saikiran, G. Lithesh, B. Srinivas and S. Ashok, "Prediction of Air Quality Index Using Supervised Machine Learning Algorithms," 2021 2nd International Conference on Advances in Computing, Communication, Embedded and Secure Systems (ACCESS), 2021, pp. 1-4, doi: 10.1109/ACCESS51619.2021.9563323.

[9] K. Saikiran, G. Lithesh, B. Srinivas and S. Ashok, "Prediction of Air Quality Index Using Supervised Machine Learning Algorithms," 2021 2nd International Conference on Advances in Computing, Communication, Embedded and Secure Systems (ACCESS), 2021, pp. 1-4, doi: 10.1109/ACCESS51619.2021.9563323.

[10] Q. P. Ha, S. Metia and M. D. Phung, "Sensing Data Fusion for Enhanced Indoor Air Quality Monitoring," in IEEE Sensors Journal, vol. 20, no. 8, pp. 4430-4441, 15 April15, 2020, doi: 10.1109/JSEN.2020.2964396.

[11] R. Jain, I. Agarwal, S. Dwivedi, S. K. Singh, A. Purwar and D. Gopinathan, ”Smart Navigation System Using Air Quality Index,” 2020 6th International Conference on Signal Processing and Communication (ICSC), Noida, India, 2020, pp. 272-277, doi: 10.1109/ICSC48311.2020.9182777.
[12] Sur Soumyadeep, Rohit Ghosal and Rittik Mondal, ”Air Pollution Hotspot Identification and Pollution Level Prediction in the City of Delhi”, 2020 IEEE 1st International Conference for Convergence in Engineering (ICCE), pp. 290-294, 2020.
[13] D. A. Padilla, G. V. Magwili, L. B. Z. Mercado and J. T. L. Reyes, ”Air Quality Prediction using Recurrent Air Quality Predictor with Ensemble Learning,” 2020 IEEE 12th International Conference on Humanoid, Nanotechnology, Information Technology, Communication and Control, Environment, and Management (HNICEM), 2020, pp. 1-6, doi: 10.1109/HNICEM51456.2020.9400051.
[14] B. Liu et al., ”A Sequence-to-Sequence Air Quality Predictor Based on the n-Step Recurrent Prediction,” in IEEE Access, vol. 7, pp. 43331-43345, 2019, doi: 10.1109/ACCESS.2019.2908081.
[15] S. S. Anjum et al., ”Modeling Traffic Congestion Based on Air Quality for Greener Environment: An Empirical Study,” in IEEE Access, vol. 7, pp. 57100-57119, 2019, doi: 10.1109/ACCESS.2019.2914672.
[16] X. Song, J. Huang and D. Song, ”Air quality prediction based on LSTM-Kalman model”, Proc. IEEE 8th Joint Int. Inf. Technol. Artif. Intell. Conf. (ITAIC), pp. 695-699, May 2019.
[17] Y. Yang, Z. Zheng, K. Bian, L. Song and Z. Han, ”Real-Time Profiling of Fine-Grained Air Quality Index Distribution Using UAV Sensing,” in IEEE Internet of Things Journal, vol. 5, no. 1, pp. 186-198, Feb. 2018, doi: 10.1109/JIOT.2017.2777820.
[18] K. Gu, J. Qiao and W. Lin, ”Recurrent Air Quality Predictor Based on Meteorology- and Pollution-Related Factors,” in IEEE Transactions on Industrial Informatics, vol. 14, no. 9, pp. 3946-3955, Sept. 2018, doi: 10.1109/TII.2018.2793950.
[19] S. Mahajan, H.-M. Liu, T.-C. Tsai and L.-J. Chen, ”Improving the accuracy and efficiency of PM2.5 forecast service using cluster-based hybrid neural network model”, IEEE Access, vol. 6, pp. 19193-19204, 2018.
[20] Y. Yang, Z. Zheng, K. Bian, L. Song and Z. Han, ”Real-Time Profiling of Fine-Grained Air Quality Index Distribution Using UAV Sensing,” in IEEE Internet of Things Journal, vol. 5, no. 1, pp. 186-198, Feb. 2018, doi: 10.1109/JIOT.2017.2777820.
[21] K. Gu, J. Qiao and W. Lin, ”Recurrent Air Quality Predictor Based on Meteorology- and Pollution-Related Factors,” in IEEE Transactions on Industrial Informatics, vol. 14, no. 9, pp. 3946-3955, Sept. 2018, doi: 10.1109/TII.2018.2793950.
[22] A. El Fazziki, D. Benslimane, A. Sadiq, J. Ouarzazi and M. Sadgal, ”An Agent Based Traffic Regulation System for the Roadside Air Quality Control,” in IEEE Access, vol. 5, pp. 13192-13201, 2017, doi: 10.1109/ACCESS.2017.2725984.
[23] S. S. Ganesh, S. H. Modali, S. R. Palreddy and P. Arulmozhivarman, ”Forecasting air quality index using regression models: A case study on Delhi and Houston,” 2017 International Conference on Trends in Electronics and Informatics (ICEI), Tirunelveli, India, 2017, pp. 248-254, doi: 10.1109/ICOEI.2017.8300926.
[24] Z. Kang and Z. Qu, ”Application of BP neural network optimized by genetic simulated annealing algorithm to prediction of air quality index in Lanzhou,” 2017 2nd IEEE International Conference on Computational Intelligence and Applications (ICCIA), Beijing, China, 2017, pp. 155-160, doi: 10.1109/CIAPP.2017.8167199.
[25] J. Jin, J. Gubbi, S. Marusic and M. Palaniswami, ”An Information Framework for Creating a Smart City Through Internet of Things,” in IEEE Internet of Things Journal, vol. 1, no. 2, pp. 112-121, April 2014, doi: 10.1109/JIOT.2013.2296516