Revolutionizing Air Quality Forecasting in Bengaluru with Advanced Machine Learning Techniques

A. Neeraja[a], Ch. Mounika Begum[a], K. Vigneswara Reddy[a], T. V. Smitha[b*], and N. Neelima[c]

[a]Department of Computer Science, Amrita School of Computing, Amrita Vishwa Vidyapeetham,Bengaluru, India
[b]Computational Science Lab, Amrita School of Engineering, Amrita Vishwa Vidyapeetham, Bengaluru, India
[c]Department of Electronics and Communication, Amrita School of Engineering, Amrita Vishwa Vidyapeetham, Bengaluru, India
[*]tv_smitha@blr.amrita.edu

***Abstract*—Exposure to air pollution can significantly affect one's health, which can lead to itchy eyes, breathing issues, and hospitalization. There are a variety of pollutants such as indoor, biological, and outdoor pollutants. This paper emphasizes estimating the quality of air concerning outdoor pollutants. This model aims to effectively predict Air Quality Index (AQI) in major populated stations in Bengaluru city. Using data from five monitoring stations - Peenya, Bapuji Nagar, Silk Board, Hombegowda, and BTM Layout- collected between 2021 and 2023, machine learning models are developed to forecast AQI for 2024. The research involved data pre-processing, AQI calculation from key pollutants, and training of models for AQI prediction. The findings indicate that the Random Forest algorithm outperformed the Decision Tree in both regression and classification tasks, providing more accurate predictions. Additionally, the study included a visual mapping of AQI categories using color codes, revealing trends in air quality across different regions. This research serves as a foundation for future work, including expanding the model to predict AQI across India and integrating it with real-time data collection hardware to enhance prediction accuracy.**

***Index Terms*—Air Pollutants, Machine Learning, Bengaluru, Prediction, Air Quality Index.**

I. Introduction

Mother Earth, our home, is composed of various gases and pollutants collectively known as air, which is essential for all living organisms on Earth. This vital quality of air is diminishing like a ticking clock. The main reason behind this is the emission of harmful gases released by industries, vehicles and many more. The release of these harmful combinations into atmosphere is causing significant health problem for all living beings.

Due to such health risks of air pollution mainly in the urban regions, time-to-time calculations of pollution levels are much needed such that, forecasting the quality of air has become a major part of air pollution research. To provide timely pollution levels, government agencies introduced the air quality index (AQI) to demonstrate the air quality that everyone is breathing. With this, the government bodies were mainly trying to explain the risk of performing actions like deforestation, plastic, and chemical usage.

Air consists of PM2.5 particulates, causing approximately 6.4 million deaths worldwide, with over 2 million deaths recorded in India alone. With this information, everyone is learning about the severity of air pollution in India [1]. Certain areas of the country, such as Delhi, Haryana, Punjab, Bihar, Bengaluru, Hyderabad, and Visakhapatnam, are the worst affected. When looking at the northern regions of the country [2], Delhi stands at the top, where in November 2022, the state's air quality reached hazardous levels, leading to the closure of schools and restrictions on deforestation. In most parts of the city, the AQI values crossed over 500 indicating very poor air quality. This incident has impacted a lot of people in many ways [3]. In November 2022, Bengaluru, located in the southern part of the country, was reported as the most polluted city. It experienced a significant decline in air quality, with several areas of the city registering poor and very poor Air Quality Index (AQI) ranges, with values exceeding 300, which is considered harmful to human health. This poor air quality was attributed to construction dust and smoke from firecrackers during the Diwali festive season [4].

Apart from this, Bengaluru, also known as the Silicon Valley of India, has swiftly emerged as an urbanized city that has transformed into a technological and industrial hub. With this, the rise in IT companies and professionals has also increased the traffic, which is contributing tremendously to air pollution in the city [5]. The industrial areas on the outskirts of the city produce and emit pollutants through their manufacturing processes and combustion of fossil fuel.

Overall these past few years, the city's AQI has always crossed safe limits, particularly in heavy-population and industrially concentrated areas [6]. So by analyzing the city's historical AQI data which shows the frequent instances of high pollution levels and fluctuations, Bengaluru city is choosen for predicting AQI values by machine learning [7]. One major advantage of using machine learning in AQI prediction is that it holds the power of efficiency when handling vast data, extending in terms of history and real-time data [8].

In machine learning itself there are many algorithms but when looking into previous research, Linear Regression

presents a straightforward model that assumes a linear relationship between the input variables and AQI; simple to understand, but its performance is poor with respect to complex patterns [9]. Decision Trees creates a tree-like structure by splitting data on feature values, hence offering interpretability, but it overfits easily. Being the combination of several decision trees, Random Forest provides better accuracy and robustness to complex data sets. It is shown that Random Forest and Decision Trees are quite effective against high values and complex patterns [10]. These models can thus be evaluated by some performance metrics, which include the following: RMSE is a measure of the root mean square error - this puts much weight on large mistakes - and R-squared, a measure for the proportion of variance explained by the model, to evaluate and compare models for accuracy and effectiveness [11].

When looking into AQI it came to know that it is worked out by the CPCB [12] by considering only major health-impacting pollutants common in urban areas. Particulate matter comes under PM2.5 and PM10 because it can penetrate deep inside the respiratory system and cause serious health problems. NO2 forms ground-level ozone and fine particulate matter, while it is also emitted in great quantities from combustion processes. The other gas to be measured is SO2, which causes respiratory problems and contributes to the formation of fine particulate matter, formed mostly due to industrial activity. Other key parameters are Ozone and Carbon Monoxide. Ozone is formed from chemical reactions involving NO2; this may lead to respiratory problems, while CO reduces oxygen delivery within the body and causes symptoms like headaches. Ammonia (NH3) which is mainly emitted from the agricultural sector cause significant amount of harm to health because of their negative effects. CPCB chooses these pollutants as the base for AQI calculations under an all-inclusive parameter for air quality and associated health risks [13] - [16].

In this work the emphasize is on estimating the quality of air concerning outdoor pollutants by effectively predicting AQI using advanced machine learning techniques for Bengaluru city by analyzing data from five stations: Silkboard, Peenya, Bapuji Nagar, Hombegowda, and BTM Layout. The concentration values for each pollutant are obtained from the CPCB official website, covering the period from 2021 to 2023, and calculated the AQI based on this data. The machine learning model is designed using Random Forest Algorithm, Decision Tree Algorithm and trying to compare which is best, to predict AQI for the year 2024. Additionally, The mapping of AQI values for each station was done using color codes to represent the different conditions like Good, Satisfactory, Moderate, Poor, Very Poor or Severe, according to the ranges provided by the CPCB website.

II. WORKFLOW

The implementation involves following steps: Data Collection, Data Preprocessing, AQI Calculation, Model Building, Model Training, Model Testing, Prediction, and Mapping. The model is developed and explored using Python programming language. A Complete AQI Design Flow of the process is depicted in Fig. 1.

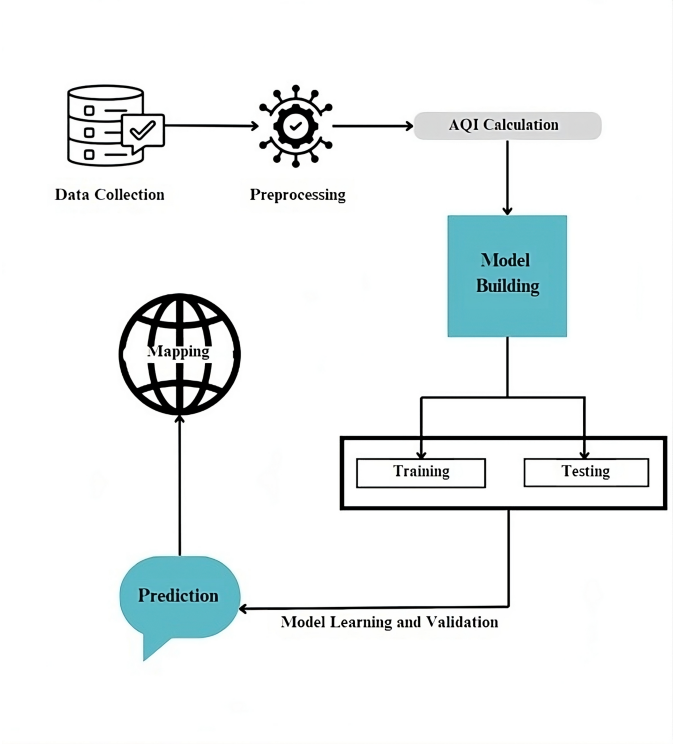

Fig. 1: A Complete AQI Design Flow

A. Data Collection

The air quality data of Bengaluru has been collected from the central pollution control board of India (Data source link 2023). The database covers from the time January 1, 2021 to 31st December 2023. It encompasses different type of pollutants and other meteorological data which are recorded frequently. The values of air quality index was obtained from the five different Air Quality monitoring stations in Bengaluru, managed by CPCB. Stations are efficient in measuring different air pollutants such as PM2.5, PM10, NO2, SO2, ozone, CO, and NH3. The target variable denoted as the Air Quality Index (AQI).

B. Data Preprocessing

During data preprocessing, concentration values of pollutants (PM2.5, PM10, NO2, SO2, ozone, CO, NH3) are gathered for five stations: Silkboard, Peenya, Bapuji Nagar, Hombegowda, and BTM Layout. The data used spans from the years 2021 to 2023. The missing values, records with blank or ”NAN” entries, are handled by replacing them with average concentration of that pollutant for that year. This process resulted in a clean and complete dataset.

C. AQI Calculation

The AQI is calculated using different pollutants with specific formula

Initially each pollutant’s concentration value should be collected (PM2.5, PM10, NO2, SO2, ozone, CO, NH3) For each pollutant, convert it’s concentration into a

sub-index value using the below formula

$$Iq = [(ACG - ACS)/(BCG - BCS)*(Cp-BCS)]+ACS \quad (1)$$

where,

BCG = Breakpoint concentration greater or equal to given concentration

BCS = Breakpoint concentration smaller or equal to given concentration

ACG = AQI value corresponding to BCG

ACS = AQI value corresponding to BCS; subtract one from ACS, if ACS is greater than 50

AQI = Max (Iq) (where; q= 1, 2,..., m); denotes m pollutants

After calculating sub-index values of each pollutant, the highest sub-index value among all pollutants determines the overall AQI value.

D. Model Building

1) Training: The machine learning model has been trained using the Random Forest Algorithm and the Decision Tree Algorithm.

2) Testing: The machine learning model has been tested using the Random Forest Algorithm and the Decision Tree Algorithm.

Random Forest:

This algorithm is under the category of ensemble learning methods, whereby many decision trees are constructed using bootstrap sampling and random feature selection. Each tree is fitted on a different subset of data, and the best split for each node comes from a random subset of the features, which offers diversity and robustness. For classification, it makes the final prediction by voting among the trees. For regression makes an average of tree outputs. The method increases the accuracy level while lowering the overfitting risk compared to a single decision tree [17] - [21]. The main advantage of this algorithm is it's potential for handling large datasets.

Decision Tree:

Decision Tree is an algorithm that involves supervised learning for splitting data into subsets in line with the input feature values, forming a tree-like model for decisions. Every node represents an attribute test; branches are possible test outcomes, and every leaf node classifies a label or value. Decision trees are easy to understand and interpret, but when not properly pruned, they can overfit the training data.

For both Random Forest and Decision Tree models, 80% of data is used for training. In the regression tasks, Decision Trees and Random Forest use pollutant concentrations in attempting to minimize the prediction error. Model accuracy will be assessed using Mean Squared Error and R-Squared metrics. The Random Forest model has several trees in which the outputs are averaged for the final prediction. Decision Trees uses a single tree to make predictions. For classification tasks, the two models classify AQI levels using pollutant concentrations. Random Forest aggregate predictions by taking a major vote of the different trees prediction. The remaining data of 20% is used for testing. In the case of regression testing, the quality of predictions by models regarding AQI is assessed with MSE and R-Squared. A lower value of MSE and a higher value of R-Squared both mean better performance. On the other hand, in classification testing, accuracy and classification reports are accuracy metrics against which to test the ability to predict AQI categories accurately.

E. AQI Categories and Color Codes

AQI is categorised within six different ranges, each associated with a certain remark and color code as shown in Fig. 2. This categorization helps to know the pollution levels and the expected health affects. The AQI categorization are given below:

Good : (0-50)- This indicates little or no risk resulting from air pollution; denoted color, Green.
Satisfactory : (51-100)- This is acceptable; however, because of some pollutants there is moderate health concern which is valid for very few persons; denoted color, Light Green.
Moderate : (101-200)- This is acceptable for most people; however, there may be health concern for those which show sensitivity towards air pollution; denoted color, Yellow.
Poor : (201-300)- Those who are sensitive might be affected by pollution, but the general public is unlikely to notice any health effects; denoted color, Orange.
Very Poor : (301-400)- Everyone start experiencing health effects but those who are more sensitive will get more affected; denoted color, Red.
Severe : (401-500)- Everyone may begin to experience serious health effects; denoted color, Dark Red.

AQI	Remark	Color Code
0-50	Good	
51-100	Satisfactory	
101-200	Moderate	
201-300	Poor	
301-400	Very Poor	
401-500	Severe	

Fig. 2: AQI Ranges with Remarks and Colors

III. Results

When tested with both algorithms by giving input of the concentrations of pollutants into the models, then the models generate an AQI value and a remark as shown in Fig. 3.

```
Enter pollutant values to predict AQI and AQI Value:
PM2.5 value: 16.46
PM10 value: 35.19
NO2 value: 25.72
NH3 value: 30.81
SO2 value: 10.35
CO value: 0.63
Ozone value: 22.39

Predicted AQI Category: Good
Predicted AQI Value: 35.771863141616315
```

Fig. 3: Prediction of Real-Time Data

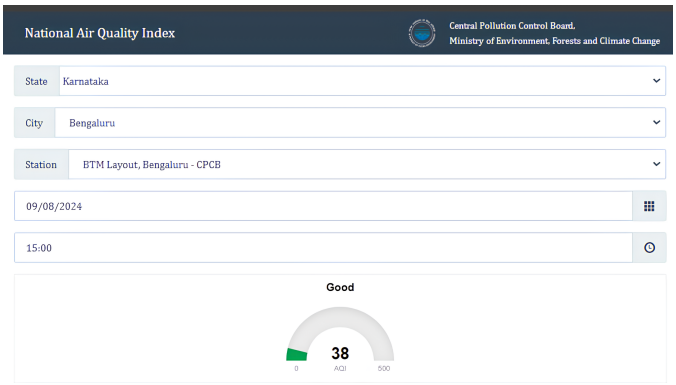

Fig. 4: Validation of Real-Time Data

The predictions are compared with data from CPCB official website where this confirms that the predictions are correct and precise as shown in Fig. 4.

TABLE I: Performance Metrics of Random Forest

Metric	Value
Mean Squared Error	3.643503389916514
R-squared	0.9980845274562784
Accuracy	0.9917808219178083

TABLE II: Classification Report of Random Forest

Class	Precision	Recall	F1-Score
Good	1.00	1.00	1.00
Moderate	0.99	1.00	0.99
Poor	0.73	0.79	0.76
Satisfactory	1.00	1.00	1.00
Severe	0.00	0.00	0.00
Very Poor	1.00	0.25	0.40

TABLE III: Performance Metrics of Decision Tree

Metric	Value
Mean Squared Error	49.103004306
R-squared	0.97418543459
Accuracy	0.99817351598

The regression and classification metrics of both algorithms are presented in TABLE I to TABLE IV.

A. *Comparison*

In this subsection, The performance of both models in predicting AQI values is compared.

TABLE IV: Classification Report of Decision Tree

Class	Precision	Recall	F1-Score
Good	1.00	1.00	1.00
Moderate	1.00	1.00	1.00
Poor	1.00	1.00	1.00
Satisfactory	1.00	1.00	1.00
Severe	0.00	0.00	0.00
Very Poor	0.80	1.00	0.89

TABLE V: Comparing Performance Metrics

Metric	Random Forest	Decision Tree
Regression Model Mean Squared Error(MSE)	3.64	49.10
Regression Model R-squared(R^2)	0.99	0.97
Classification Model Accuracy	0.99	0.99

From TABLE V It is observed that the Random Forest demonstrates better performance than the Decision Tree model, which is evident because of the substantially lower Mean Squared Error and higher R-Squared value. Therefore, this means that the random forest model makes a more accurate prediction and accounts for more of the variances in the AQI values. While both models give the same level of accuracy for classification, the random forest model is better for regression tasks which makes it better for AQI prediction.

B. *Mapping and Categorization*

This subsection presents the plots of AQI levels and categories for Bengaluru across the years 2021, 2022, 2023.

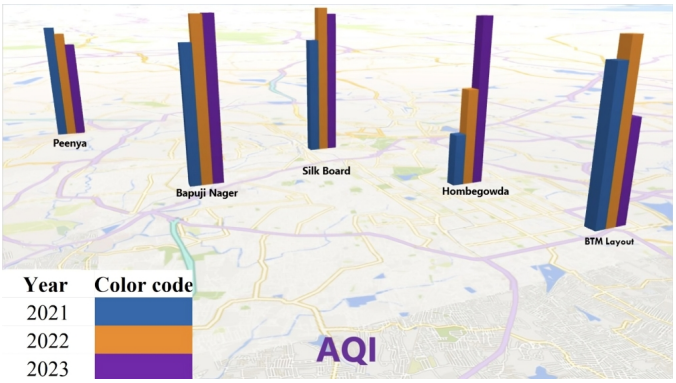

Fig. 5: AQI Trends across monitoring stations from 2021-2023

Fig. 5 depicts the comparison of AQI levels across five monitoring stations. The interpretations from Fig. 5 as follows:
1) Peenya : The AQI levels are on a down trend from 2021-23, hence an improving trend in air quality is probably due to effective control measures against pollution.
2) Bapuji Nagar : The AQI levels have raisen in 2022 from 2021 and got decreased in 2023. It may indicate that sources of pollution or weather conditions are different.
3) Silk Board : With a linear rise from 2021-23, the AQI levels here depicts degrading air quality, perhaps due to increased traffic or industrial activities.
4) Hombegowda : AQI levels increased from 2021-22, with a slight fall in 2023- indicating a persistent problem of air quality but slight improvements recently.

5) BTM Layout : There is a continuous decrease in AQI levels over three years, indicating that the air quality has been improving probably due to local policies or initiatives for pollution reduction.

From Fig. 5, it is observed that Peenya and BTM Layout show an increase in trend, while silk Board and Hombegowda need more efforts to be put in. where, Bapuji Nagar requires long-term sustained measures for improvement.

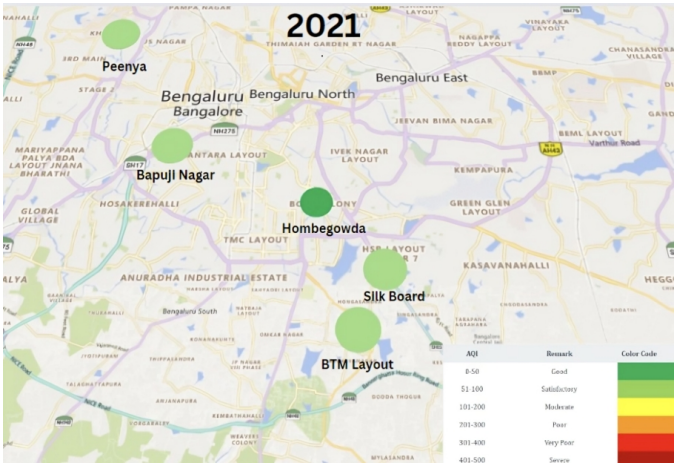

Fig. 6: AQI Categories and Color Codes for Bengaluru in 2021

The Fig.6. shows the AQI category visualisation with color codes across different locations in Bengaluru, 2021. Peenya, Bapuji Nagar, Silk Board, and BTM Layout are light green in color, indicating that they have satisfactory air quality. Meanwhile Hombegowda is colored green, which is even better in terms of air quality.

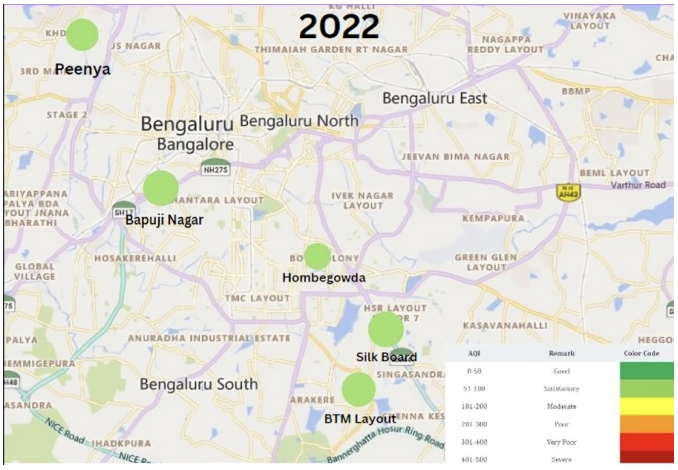

Fig. 7: AQI Categories and Color Codes for Bengaluru in 2022

From the Fig. 7 and Fig. 8 we observe that the AQI categories and color codes for Bengaluru in 2022 and 2023 are similar, with all the stations having light green in color, thus indicating satisfactory air quality. this consistency will point to the fact that the quality of air has to be stable and good across these years, likely to have been driven by some effective control measures on pollution or favorability of the environment.

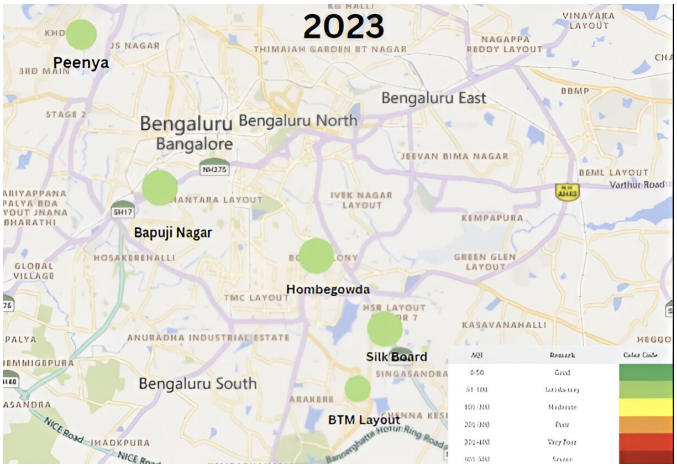

Fig. 8: AQI Categories and Color Codes for Bengaluru in 2023

IV. Conclusion and Future Scope

Quality of air has become a major concern with increasing population, and vehicle usage. Knowing the quality of air in the area where everyone lives can enhance the significant measures to take necessary action. In conclusion, this research effort has successfully exploited machine learning models to predict and analyze AQI levels over Bengaluru, where Random Forest has been more accurate than Decision Tree. AQI mapping for all these years showed some interesting trends in air quality in the city. In the future, the model is intended to be enriched with additional features for better visualization and real-time prediction. This further can be extended to a hardware model that supports real-time AQI forecast. Future work will therefore be geared to word the generalization of this model in order to predict AQI levels for the entire country, unbound by geographical limitations since this study only involved five stations in Bengaluru.

Acknowledgment

Sincere gratitude to all the agencies that contributed data to the Central Pollution Control Board (CPCB).

References

[1] Sharma, Gaurav, et al. ”Comparative Analysis of Machine Learning Techniques in Air Quality Index (AQI) prediction in smart cities.” International Journal of System Assurance Engineering and Management (2024): 1-16.

[2] Natarajan, Suresh Kumar, et al. ”Optimized machine learning model for air quality index prediction in major cities in India.” Scientific Reports 14.1 (2024): 6795.

[3] Kumar, Raj, et al. ”Air pollution and its effects on emergency room visits in tertiary respiratory care centers in Delhi, India.” Monaldi Archives for Chest Disease 94.1 (2024).

[4] Dasgupta, Anindita, and Uttam Kumar. ”Interactive influence of urban heat island and urban pollution island in two major cities of India (Bangalore and Delhi).” 2024 International Conference on Machine Intelligence for GeoAnalytics and Remote Sensing (MIGARS). IEEE, 2024.

[5] Kavyashree, N. V., and H. B. Rekha. ”Data Analysis & Prediction Of Air Quality Parameters In Bangalore City.” (2021).

[6] Atmakuri, Krishna Chaitanya, and K. V. Prasad. ”Urban air quality analysis and aqi prediction using improved knn classifier.” Journal of Pharmaceutical Negative Results (2023): 7673-7681.

[7] Mallappa, Sugurappa. ”A Study On Air Quality Management And Control Measures To Reduce Air Pollution In Bangalore.” Journal of Research Administration 5.2 (2023): 11668-11677.

[8] Saxena, Riya, et al. ”Air-Quality Index Prediction Using Auto Ml Library, TPOT.” 2023 14th International Conference on Computing Communication and Networking Technologies (ICCCNT). IEEE, 2023.

[9] Nair, Aadarsh Sathinarayan, Sangita Khare, and Amrita Thakur. ”Air Quality Index Prediction of Bangalore City Using Various Machine Learning Methods.” Information and Communication Technology for Competitive Strategies (ICTCS 2022) Intelligent Strategies for ICT. Singapore: Springer Nature Singapore, 2023. 391-406.

[10] Aram, S. A., et al. ”Machine learning-based prediction of air quality index and air quality grade: a comparative analysis.” International Journal of Environmental Science and Technology 21.2 (2024): 1345-1360.

[11] Ketu, Shwet. ”Spatial air quality index and air pollutant concentration prediction using linear regression based recursive feature elimination with random forest regression (RFERF): a case study in India.” Natural Hazards 114.2 (2022): 2109-2138.

[12] Website of Central Pollution Control Board, ”https://cpcb.nic.in/”, 2024.

[13] SK, Aruna, and Gokulan Ravindiran. ”Integrating machine learning techniques for Air Quality Index forecasting and insights from pollutant-meteorological dynamics in sustainable urban environments.” Earth Science Informatics (2024): 1-16.

[14] Gupta, N. Srinivasa, et al. ”Prediction of air quality index using machine learning techniques: a comparative analysis.” Journal of Environmental and Public Health 2023.1 (2023): 4916267.

[15] Sarkar, Pritisha, et al. ”Analyzing the Severity of Air Pollution in an Industrialized Suburb.” 2023 14th International Conference on Computing Communication and Networking Technologies (ICCCNT). IEEE, 2023.

[16] Kumar, Arun, and Anupam Jamatia. ”Prediction of air quality using machine learning.” International Conference on Frontiers of Intelligent Computing: Theory and Applications. Singapore: Springer Nature Singapore, 2022.

[17] Chandan, K., Nagaraja, K.V., Gamaoun, F., Smitha, T.V., Neelima, N., Khan, U. and Hassan, A.M., 2024. Improving flow efficiency in micro and mini-channels with offset strip fins: A stacking ensemble technique for Accurate friction factor prediction in steady periodically developed flow. Case Studies in Thermal Engineering, 56, p.104232.

[18] Panimathi, B., Chandan, K., Nimmy, P. and Smitha, T.V., 2024, January. An Efficient Prediction of battery capacity based on temperature and state of charge using impedance. In 2024 Fourth International Conference on Advances in Electrical, Computing, Communication and Sustainable Technologies (ICAECT) (pp. 1-5). IEEE.

[19] Bhutra, H., Chandan, K. and Smitha, T.V., 2024, January. Effective Prediction of Coefficients and Performance of Airfoil Using ANN. In 2024 Fourth International Conference on Advances in Electrical, Computing, Communication and Sustainable Technologies (ICAECT) (pp. 1-5). IEEE.

[20] Oberoi, K.G., Deepa, K., Sangeetha, S.T. and Neelima, N., 2024, March. Predicting Vegetables And Fruits Through Supply Chain Insights. In 2024 IEEE International Conference on Interdisciplinary Approaches in Technology and Management for Social Innovation (IATMSI) (Vol. 2, pp. 1-5). IEEE.

[21] Vinutha, K., Srilatha, P., Chandan, K., Sriram, D., Madhukesh, J.K., Nagaraja, K.V. and Varshney, G., 2024. Stacking regression model approach to mixed convection flow of ternary-nanofluid over slanted surface with magnetic field, waste discharge concentration, and joule heating effects. International Journal of Thermofluids, p.100731.