2025 International Conference on Advanced Computing Technologies (ICoACT)

Air Quality Analysis and Prediction in Tamil Nadu: AN ANN and CNN approach

Dr. S. Palanivel Rajan
Department of Electronics and Communication Engineering, Velammal College of Engineering and Technology,
Madurai, India.
drspalanivelrajan@gmail.com

R. Rahul
Department of Electronics and Communication Engineering, Velammal College of Engineering and Technology,
Madurai, India.
rahulkanna170504@gmail.com

T. Jegan
Department of Electronics and Communication Engineering, Velammal College of Engineering and Technology,
Madurai, India.
jknn007@gmail.com

S. Yasar Arafath
Department of Electronics and Communication Engineering, Velammal College of Engineering and Technology,
Madurai, India.
ya0387288@gmail.com

***Abstract*—Air quality poses a significant environmental and health challenge, particularly in developing nations like India. Tamil Nadu, situated in southern India, grapples with severe air pollution issues stemming from diverse sources like discharges from automobiles, production tasks, biomass burning, and dust storms. Traditional monitoring and forecasting methods like ground-based stations and satellite observations have limitations concerning spatial and temporal coverage, cost, and upkeep. Consequently, there arises a demand for alternative approaches harnessing advancements in machine learning and big data analytics to furnish high-resolution, real-time air quality insights. In this study, we introduce an innovative methodology for analyzing and predicting air quality in Tamil Nadu employing artificial neural networks (ANNs) and convolutional neural networks (CNNs). These sophisticated machine learning techniques adeptly capture intricate nonlinear relationships from extensive and diverse datasets while adeptly handling missing and noisy data. Leveraging historical datasets encompassing air pollutant concentrations, meteorological variables, and land use features sourced from multiple channels, we train and validate our models. We conduct comparative assessments with established methodologies, assessing our models' robustness and generalizability. Furthermore, we offer an understanding of the spatiotemporal dynamics and trends of air quality in Tamil Nadu, identifying pivotal factors and drivers of air pollution. Our proposed approach holds promise for policymakers, researchers, and the general populace to comprehend and ameliorate the air quality scenario not only in Tamil Nadu but also in analogous regions worldwide. Additionally, our model yields the following performance metrics: Mean Absolute Error (MAE): 0.037, R-squared (R^2): 0.9998, Root Mean Squared Error (RMSE): 0.4522.**

Keywords*—Air Quality, Atmosphere, Environmental monitoring, Machine Learning, Neural networks, Time Series Analysis.

I. INTRODUCTION

Air quality is a crucial factor in physical conditions, ecosystem/surroundings, and commerce. From this system, summit an approach for quality of air analysis and prediction in Tamil Nadu using ANN and CNN. ANN and CNN are powerful machine learning techniques that can understand complicated interactions from huge and diverse datasets and can handle missing and noisy data. In the face of escalating global concerns about air quality and its profound implies for physical health, our research aims to contribute significantly by integrating advanced technologies, specifically CNNs and ANNs, into analysis and prediction for quality of air. The body of literature encompassing our study reflects a rich tapestry of innovative techniques hired to solve the complexity of air pollution dynamics. Studies such as [1] advocate for the deployment of wireless sensor networks, addressing limitations associated with globally trained models. Simultaneously, the binding of Convolutional Neural Networks – Long Short Term Memory (CNN-LSTM) models, demonstrated in [2], has proven effective in predicting particulate matter (PM10) concentrations. Regional air quality analysis, as exemplified in [3], integrates decision trees and clustering techniques to identify pollution hotspots, while [4] underscores the enhanced predictive capabilities of Long Short Term Memory (LSTM) over traditional methods. Regression techniques confer the precise prediction of air quality in smart cities, and models like Attention Temporal Graph Convolutional Network (A3T-GCN) showcase excellence in predicting Nitrogen dioxide (NO_2) concentrations. Unconventional approaches, such as leveraging CNNs for particulate matter (PM2.5) estimation from natural images [7], provide noteworthy accuracy. Moreover, localized linear regression studies offer valuable insights into predicting Air Quality Index (AQI) classifications. The exploration extends to studies such as Airborne and Satellite Investigation of Asian Air Quality's (ASIA-AQ) plans for Asian air quality study using satellites [9] and urban air quality data mining and visualization using Seasonal Autoregressive Integrated Moving Average with Exogenous Factors (SARIMAX) and Prophet [10]. The acuteness of air pollution, as discussed in [11], emphasizes the health impacts of fine particles PM2.5 and proposes a predictive model using multiple neural networks [12]. In unconventional domains, studies address tool wear prediction for Computer Numerical Control (CNC) machine tools [13], spacecraft debris removal using Faster Region-based Convolutional Neural Network (R-CNN) [14], and comparison of Maximum power point tracking (MPPT) technologies [15].(Palanivel Rajan, 2020; Palanivel Rajan & Vivek, 2019) Our research synthesizes insights from these diverse studies, proposing a novel integrated framework leveraging CNNs and ANNs for air quality analysis and prediction. Through this endeavor, we provide meaningful information to the field of pollutant analysis, fostering advancements in environmental monitoring and public health.

This paper follows the methodology and design detailed in **Section II**, while **Section III** presents results, including model performance comparisons using RMSE, MAE, and R^2. The

979-8-3315-0985-9/25/$31.00 ©2025 IEEE

analysis involves a correlation heatmap and residual plots highlighting performance and variable relationships. **Section IV** summarizes key insights, and **Section V** provides references supporting this research.

II. Methodology

Air quality analysis and prediction involve measuring, monitoring, and forecasting atmospheric pollutants like PM, Ozone(O_3), NO_2, sulfur dioxide (SO_2), Carbon Monoxide (CO), and Volatile Organic Compounds (VOC). Data come from enormous sources, including sensors, satellites, and historical records. Challenges arise due to the difficult environment and variability of atmospheric processes affected by weather, topography, and human activities. Machine learning and deep learning methods, including ANN and CNN, help address these challenges by identifying patterns information for more precise predictions.

A. Data description

The dataset provided contains data associated with air quality in the state of Tamil Nadu, India. The key components of the dataset. Figure 1 to Figure 4 depict the relationship between air pollutants and sampling dates. Figure 1 and Figure 2 represents SO_2 and NO_2 concentrations, respectively, originating from industrial processes and combustion. Respirable suspended particulate matter (RSPM/PM10) and PM 2.5 indicate particulate matter concentrations, with PM 2.5 associated with more severe health impacts due to its smaller size. These figures illustrate the temporal variations in air pollutant concentrations, crucial for understanding environmental trends and potential health implications.

B. AQI (Air Quality Index)

The AQI is an analytical measurement helps to convey area. It is estimated based on several air pollutants. The AQI gives the potential health risks associated with the observed pollutant levels.

AQI = Sum of the considered pollutants

In general, AQI is split into multiple Ranges from excellent to dangerous with corresponding health It acts as an important tool for common society policymakers, researchers, and the general public to acknowledge the consequence of air pollution on human health. Table 1 explains calculation or classification classification of AQI. The dataset is utilized to analyze trends, assess the impact of different pollutants, and provide to knowledge of air quality in Tamil Nadu. Additionally, by incorporating AQI values, you can deliver a coherent evaluation of overall air quality conditions and their potential health implication potential health implication. Figure 5 shows the relationship between AQI and Date.

C. CNN (Convolution Neural Network)

Data Representation: Represent air quality data as images or grids, with each pixel representing a specific feature or parameter.

Convolutional Layers: Use these layers to automatically procure spatial uniqueness from the input data. This involves a sliding over the input grid, computing a weighted sum at each position, and applying an activation function. The convolutional operation is represented as equation (1).

$$Output(i,j) = Activation(\sum_{k=1}^{K}\sum_{l=1}^{L}(Weight(k,l) \times Input(i+k,j+l)) + Bias \qquad \ldots(1)$$

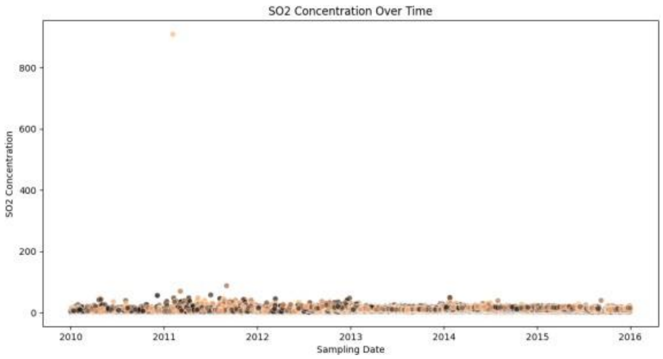

Figure 1: Scatter Plot between SO_2 and Sampling Date

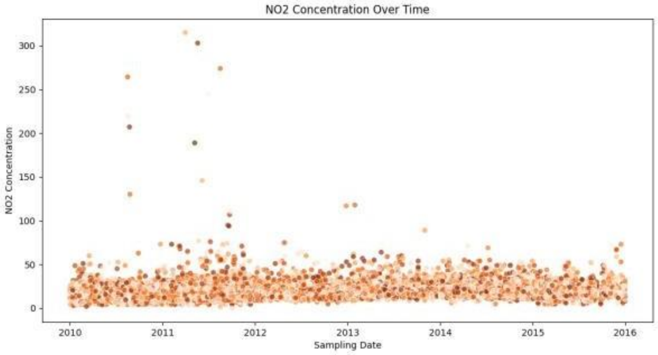

Figure 2: Scatter Plot between NO_2 and Sampling Date

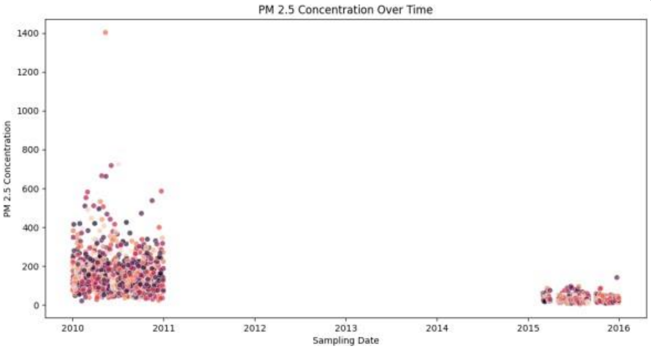

Figure 3: Scatter Plot between RSPM/PM10& Sampling Date

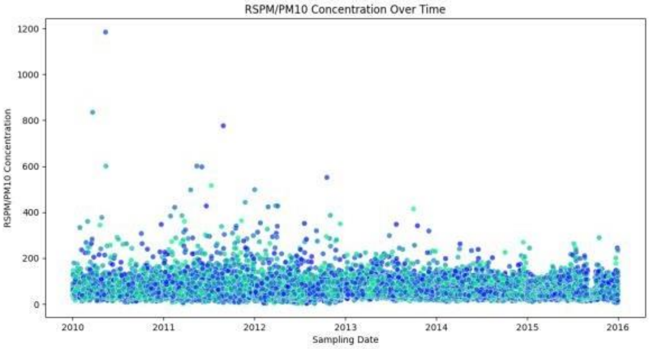

Figure 4: Scatter Plot between PM 2.5 and Sampling Date

TABLE 1. AQI Calculation

Level of Concern	Value of Index	Description
Good	AQI<50	Health Impact: Air quality is considered excellent and air quality comes with no risk. Environmental Impact: Air quality is excellent, and the air is generally clean. This level is beneficial for health and the environment.
Moderate	51<AQI<100	Health Impact: Air quality is moderate. however, some pollutants may be vulnerable to a few individuals reactive to air pollution. Environmental Impact: Air quality is moderate, so there shall be health concerns for a few of people for those who with respiratory or heart conditions.
Satisfactory	101<AQI<200	Health Impact: People of reactive groups may experience health effects,but the common society is hardly affected. Environmental Impact: members of reactive groups may indulged with physical health effects. The public is less likely to be negative impact .
Poor	201<AQI<300	Health Impact: Everyone may experience health effects; sensitive group members may experience health effects. Environmental Impact: Everyone may begin to experience more health effects. Sensitive groups may experience severe health effects.
Very Poor	301<AQI<400	Health Impact: The entire population is likely to experience hassle health effects Environmental Impact: Health warnings of emergency conditions. The entire population is more likely to be affected.
Severe	401<AQI<500	Health Impact: Health alert: everyone may experience more serious health effects. Environmental Impact: Health warnings of emergency conditions. The entire population is more likely to be affected.

Pooling Layers: Apply pooling layers to down reduced dimensionality while preserving essential features. Pooling layers typically use max or average pooling to reduce dimensionality. The equation (2) explains the Max Pooling and equation (3) explains Average pooling.

$$Output(i,j) = max\left(input\big((2i,2j),(2i+1,2j),(2i,2j+1),(2i+1,2j+1)\big)\right) \quad ...(2)$$

$$Output(i,j) = \frac{1}{4}[\sum_{m=0}^{1}\sum_{n=0}^{1} Input(2i+m, 2j+n)] \quad ...(3)$$

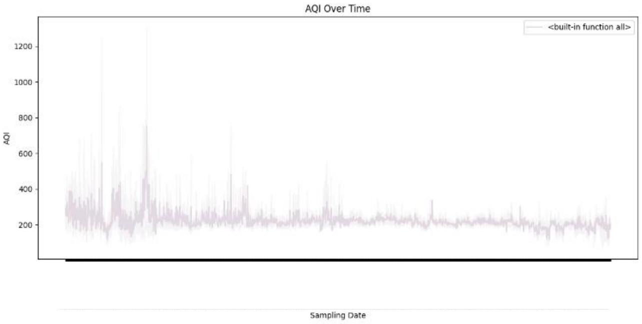

Figure 5: AQI vs Sampling Date

Fully Connected Layers: Connect these convolution layers to fully connected layers for prediction. The equation (4) conveys transformation occurs in layers.

$$Output(i) = Activation(\sum_{j=1}^{M}(weight(i,j) \times Input(j) + Bias(i))) \quad \dots(4)$$

Connect the convolutional layer's output to entirely connected layers for further processing and prediction. Each neuron node is connected to all neuron nodes in the previous layer.The architecture of CNN:

D. Training and Prediction of CNN

Train the CNN using historical air quality data and deploy it for predicting future air quality. The Equation (5) explains the Weight Update in Backpropagation and equation (6) gives Prediction in CNN

$$New\ Weight = Old\ Weight - (Learning\ rate \times (\frac{\partial loss}{\partial old\ weight})) \quad \dots(5)$$

$$Prediction\ Output = Activation(\sum_{j=1}^{M}\big(Weight(i) \times Input(j)\big) + Bias) \quad \dots(6)$$

E. ANN (Artificial neural Network)

ANN is a computational model where brain structure is inspired for forming this model. This model has 3 layers forming layers of interconnected nodes consist of an input layer, one or more hidden layers, and an output layer.

Data Preparation: Collect and preprocess air quality data, including features like pollutant quantity climate/weather conditions, and a few more factors.

$$Normalized\ Value = \frac{Original\ Value - Min\ Value}{Max\ Value - Min\ Value} \quad (7)$$

Normalize features to a common scale provided in the equation (7), which helps in training the neural network.

Network Architecture: Design the ANN architecture, specifying the number of layers, neurons per layer, and activation functions.

$$No\ of\ Parameters\ in\ a\ Layer = \begin{pmatrix}(No\ of\ Neurons\ in\ Previous\ Layer\ +1) \times \\ No\ of\ Neurons\ in\ Current\ Layer\end{pmatrix} \quad \dots(8)$$

The equation (8) and equation (5) calculates the count of parameters and elements (weights and biases) in a neural network's connected layer respectively.

Training: Historical air quality data is helped to train the ANN. The map input features are understood by the network to the corresponding air quality output.

Adjust the weights during training using gradient descent to reduce the loss function.

Validation and Testing: Evaluate the trained model on validation data to ensure generalization. Test the model on unseen data to assess its predictive performance. Equation (9) gives the accuracy for segregation prediction (precision):

$$Accuracy = \frac{Number\ of\ Correct\ predictions}{Total\ Number\ of\ Predictions} \quad (9)$$

Use appropriate evaluation metrics like Precision for classification tasks to examine model execution.

Prediction: Deploy the trained ANN prevision input source. Keras Model:

The `Model` function is used to instantiate a Keras model given its inputs and outputs. This creates a model that includes all layers required in the computation of `outputs` given `inputs`.

F. Model Evaluation and Prediction

Criteria such as RMSE, MAE, and R^2 are commonly used for regression model evaluation, offering different insights into model performance. A model with low RMSE and high R^2 is generally deemed good, yet the choice depends on specific task requirements; minimizing large errors (RMSE) or capturing the overall trend (R^2) might be prioritized based on context.

G. Data Analysis and Visualization

Creation of graphical representation where the x-axis represents the actual AQI values and the y-axis represents the predicted AQI values. The alpha = 0.5 parameter regulates the transparency of the points, making it simpler to observe overlapping points. Figure 6 gives the model's performance by plotting the existing values against the identified values.

Figure 6 illustrates the model's AQI prediction accuracy across locations. The x-axis shows actual AQI values from sensors, while the y-axis represents model-predicted AQI. Points closer to the line of best fit indicate higher accuracy. The model performs well for low to moderate AQI but tends to underestimate high values.

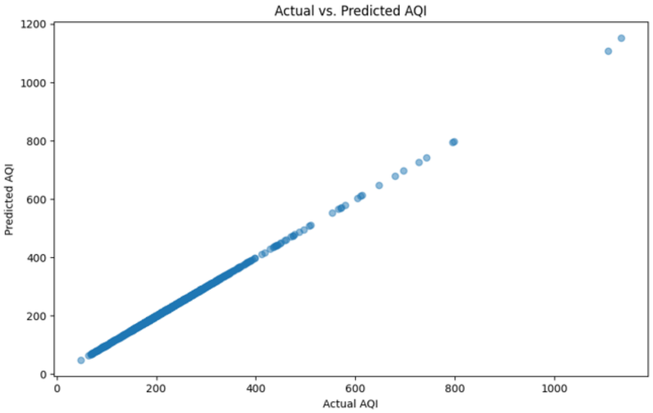

Figure 6: Scatter Plot between Actual AQI and Predicted AQI

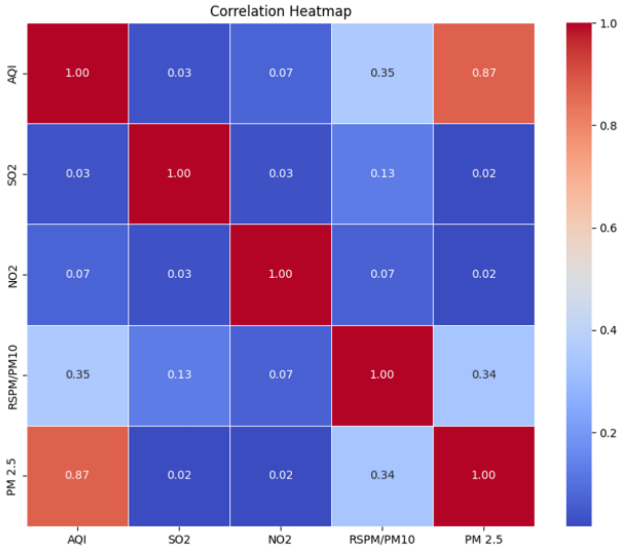

Figure 7: Correlation Heatmap

III. Results and Discussion

A. Correlation Heatmaps and Residuals

A correlation heatmap visually represents the efficiency of correlations between multiple variables using color-coded cells. Darker colors signify stronger correlations, with warm colors convey. Figure 7 represents correlation heatmap.

$$\text{Residual} = \text{Observed value} - \text{Predicted value}$$

Figure 8 shows a residual plot comparing observed and predicted values. Randomly scattered, close-to-zero residuals indicate a good model fit, with one outlier visible. A quantile-quantile plot (Q-Q) can verify if residuals follow a normal, bell-shaped distribution. Figure 9 shows residuals clustering around -2.5, suggesting slight overestimation and low error, indicating model accuracy.

B. AQI Categories

The purpose of these steps is to examine how well the model is performing aligned with categorizing AQI values. This is useful for understanding the workings of the model.
Output is finalized by:

C. Mean Absolute Error (MAE)

This is another measure of how close the identified values to the existing values. The MAE is calculated by using Equation (10).

$$MAE = \frac{1}{n}\sum_{i=1}^{n} \mid y_i - \hat{y}_i \mid \quad (10)$$

Where n is count of observations, y_i is the existing value, and $\hat{y}_i$ is identified value. A lower MAE indicates better model accuracy and is more robust to outliers than MSE. Here, the **MAE of 0.037** shows minimal prediction error.

D. *R-squared (R^2)*

TABLE 2. Result Comparison

S.No.	*Methodology*	*Output*		
		RMSE	*MAE*	*R^2*
1	Linear Regression [8]	11.9	8.89	50.8
2	LSTM [4]	12.382	6.626	-
3	Gated Recurrent Unit (GRU) [4]	14.2	6.97	-
4	CNN-LSTM [2]	5.4	11.1	-
5	Random Forest, Linear Regression [5]	0.591	20.13	11.36
6	A3T – GCN [6]	9.354	6.065	-
7	ANN [1]	26.79	-	73.021
8	ANN, CNN **[Proposed]**	0.4522	0.037	99.98

This is a measure of how well a regression model explains or predicts variation. It ranges from 0 to 1, where 0 means that x cannot explain any variation in y, and 1 means that changes disclose in y can be done by x. The formula for R^2 depends on whether you use least squares or maximum likelihood estimation to fit your model. For least squares estimation, R^2 can be calculated as in equation (11).

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}$$

...(11)

where $\hat{y}$ is mean value of y. For maximum likelihood estimation, R^2 determined as

$$R^2 = r^2$$

where r is an estimate of the correlation coefficient from x to y, given by equation (12).

$$r = \frac{\sum_{i=1}^{n} x_i y_i - n(\bar{x} y_1 - \bar{x})(\bar{y}_1 - \bar{y})}{\sqrt{\sum_{i=1}^{n} x_i^2 - n(\bar{x})^2}\sqrt{\sum_{i=1}^{n} y_i^2 - n(\bar{y})^2}}$$

...(12)

where $\hat{x}$ and $\hat{y}$ are means of x and y, respectively. A high R-squared value, like **0.9998**, suggests an excellent fit, meaning the model captures most data variation. However, R-squared alone doesn't indicate causation or potential overfitting. It simply measures how well identified align with existing data.

E. *Root Mean Squared Error (RMSE):*

The RMSE is a measure that provides standard deviation of the differences between actual and predicted values. It is an extension of Mean Squared Error (MSE) and is calculated by taking square root of the MSE. The formula for RMSE is in equation (13).

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}$$

...(13)

where n is count of observations, y_i is existing value, and $\hat{y}_i$ is the identified value. The RMSE is particularly useful because it provides an interpretable scale that is in the same units as the target variable. **RMSE of 0.4522174134442851**.

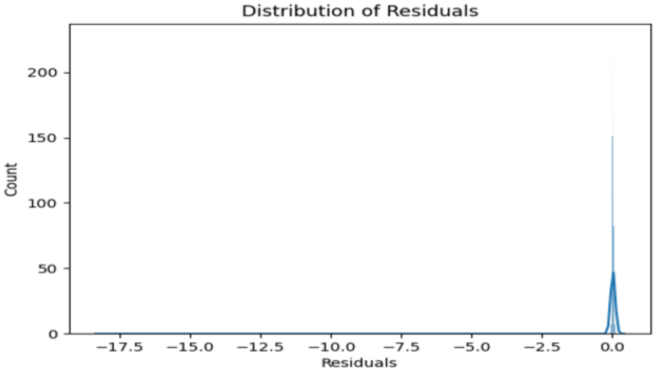

Figure 8: Residual Plot

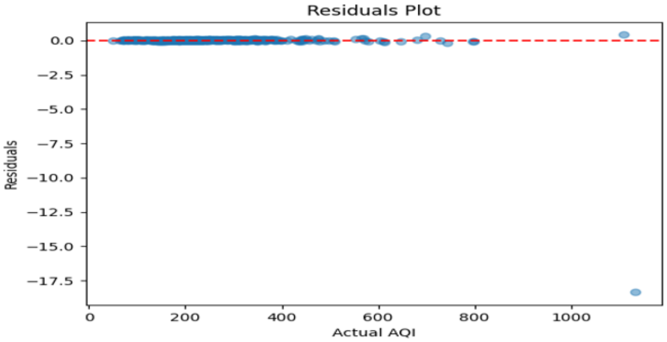

Figure 9: Distribution of Residuals

F. *Actual vs. Predicted Categories Data Frame:*

The Data Frame distinguish the actual and predicted AQI categories for the first 100 instances. Each row represents a data point, and the "Actual" and "Predicted" columns show the corresponding AQI categories. It seems like there might be some missing values in the "Actual" column, indicated by the Not a Number (NaN) value.

IV. CONCLUSION

In conclusion, addressing impressive challenges, particularly quality of air in Tamil Nadu, especially in the context of developing nations like India, demands innovative solutions. This paper proposes a groundbreaking approach utilizing ANNs and CNNs, harnessing the power of machine learning and big data analytics. By surpassing the drawbacks of traditional monitoring methods, such as ground-based stations and satellite observations, our models offer high-resolution, real-time air quality insights. The robustness and generalizability of our models, validated against existing methods which is compared with our proposed model in table 2, underscore their potential significance. Moreover, our analysis sheds light on spatiotemporal air quality patterns, revealing key factors and drivers of pollution. This comprehensive approach not only equips policymakers with valuable tools but also empowers researchers and citizens to comprehend and actively contribute to improving air quality in Tamil Nadu and analogous regions facing similar environmental challenges.

REFERENCES

[1] M. Banach, Z. Dlugosz, R. Dlugosz, and T. Talaska, "The Use of Artificial Neural Networks in Predicting Air Pollution in Cities-Hardware Implementation Issues," in *Proceedings of the International Conference on Microelectronics, ICM*, Institute of Electrical and Electronics Engineers Inc., Sep. 2021, pp. 271–274. doi: 10.1109/MIEL52794.2021.9569046.

[2] L. Jovova and K. Trivodaliev, "Air Pollution Forecasting Using CNN-LSTM Deep Learning Model," in *2021 44th International Convention on Information, Communication and Electronic Technology, MIPRO 2021 - Proceedings*, Institute of Electrical and Electronics Engineers Inc., 2021, pp. 1091–1096. doi: 10.23919/MIPRO52101.2021.9596860.

[3] R. S. Kumar, A. Arulanandham, and S. Arumugam, "Air quality index analysis of Bengaluru city air pollutants using Expectation Maximization clustering," in *2021 International Conference on Advancements in Electrical, Electronics, Communication, Computing and Automation, ICAECA 2021*, Institute of Electrical and Electronics Engineers Inc., 2021. doi: 10.1109/ICAECA52838.2021.9675669.

[4] F. Naz *et al.*, "Comparative Analysis of Deep Learning and Statistical Models for Air Pollutants Prediction in Urban Areas," *IEEE Access*, vol. 11, pp. 64016–64025, 2023, doi: 10.1109/ACCESS.2023.3289153.

[5] P S Neethu, et. al., "Performance Evaluation of SVM based Hand Gesture Detection and Recognition System using Distance Transform on different datasets for Autonomous Vehicle Moving Applications", Circuit World, ISSN: 0305-6120, Vol.: 48, Issue: 2, pp. 204-214, 2022. DOI: 10.1108/CW-06-2020-0106.

[6] D. Iskandaryan, F. Ramos, and S. Trilles, "Graph Neural Network for Air Quality Prediction: A Case Study in Madrid," *IEEE Access*, vol. 11, pp. 2729–2742, 2023, doi: 10.1109/ACCESS.2023.3234214.

[7] Chatterjee, K., Kumar, S. S., Kumar, R. P., Bandyopadhyay, A., Swain, S., Mallik, S., Al-Rasheed, A., Abbas, M., & Soufiene, B. O. (2024). Future Air Quality Prediction using Long Short-Term Memory based on Hyper Heuristic Multi-Chain Model. *IEEE Access*. https://doi.org/10.1109/ACCESS.2024.3441109

[8] S. B. Sonu and A. Suyampulingam, "Linear Regression Based Air Quality Data Analysis and Prediction using Python," in *Proceedings of the IEEE Madras Section International Conference 2021, MASCON 2021*, Institute of Electrical and Electronics Engineers Inc., 2021. doi: 10.1109/MASCON51689.2021.9563432.

[9] J. H. Crawford *et al.*, "The Airborne and Satellite Investigation of Asian Air Quality (Asia-Aq): An Opportunity for International Collaboration," in *International Geoscience and Remote Sensing Symposium (IGARSS)*, Institute of Electrical and Electronics Engineers Inc., 2022, pp. 6506–6509. doi: 10.1109/IGARSS46834.2022.9883819.

[10] V. Gupta, S. Kapadia, and C. Bhadane, "Time Series Analysis and Forecasting of Air Quality in India," in *2023 5th International Conference on Electrical, Computer and Communication Technologies, ICECCT 2023*, Institute of Electrical and Electronics Engineers Inc., 2023. doi: 10.1109/ICECCT56650.2023.10179673.

[11] L. Lyu, J. Kong, and Y. Peng, "Urban Ambient Air Quality Data Mining and Visualisation," in *Proceedings - 2022 International Conference on Artificial Intelligence of Things and Crowdsensing, AIoTCs 2022*, Institute of Electrical and Electronics Engineers Inc., 2022, pp. 616–620. doi: 10.1109/AIoTCs58181.2022.00101.

[12] M.Paranthaman, et.al., "Design of H Shaped Patch Antenna for Biomedical Devices", International Journal of Recent Technology and Engineering, ISSN : 2277-3878, Vol. No. 7, Issue:6S4, pp. 540-542, 2019.

[13] F. C. Zegarra, J. Vargas-Machuca, and A. M. Coronado, "Comparison of CNN and CNN-LSTM Architectures for Tool Wear Estimation," in *Proceedings of the 2021 IEEE Engineering International Research Conference, EIRCON 2021*, Institute of Electrical and Electronics Engineers Inc., 2021. doi: 10.1109/EIRCON52903.2021.9613659.

[14] Z. Wang, Y. Cao, and J. Li, "A Detection Algorithm Based on Improved Faster R-CNN for Spacecraft Components," in *2023 IEEE International Conference on Image Processing and Computer Applications, ICIPCA 2023*, Institute of Electrical and Electronics Engineers Inc., 2023, pp. 1804–1808. doi: 10.1109/ICIPCA59209.2023.10257992.

[15] Rajan, S.P (2020). Recognition of Cardiovascular Diseases through Retinal Images Using Optic Cup to Optic Disc Ratio. *Pattern Recognition and Image Analysis*, *30*(2), 256–263.