Milestone Transactions on
MEDICAL TECHNOMETRICS
Vol. 03 | Issue. 02 | 2025

RESEARCH ARTICLE OPEN ACCESS

A Hybrid ANN and XGBoost Approach to Urban Air Quality Classification

Posina Anusha

Department of Computer Science and Engineering,
Annamacharya Institute of Technology and Sciences,
Tirupati, Andra Pradesh, India

DOI: **10.5281/zenodo.16892754**
Received: 24 June 2025 / Revised: 17 July 2025 / Accepted: 16 August 2025
©Milestone Research Publications, Part of CLOCKSS archiving

Abstract – Public health is severely threatened by urban air pollution, particularly in densely populated and rapidly industrialising cities. For risk management, environmental monitoring, and targeted policymaking, urban zones must be accurately classified according to pollution levels. This paper proposes a classification framework that integrates the ensemble-based decision-making power of Gradient Boosting With the nonlinear feature extraction capabilities of neural networks, a publicly available dataset containing over 52,000 daily air quality records from six major cities was used. The model was designed to distinguish between industrial and residential urban areas based on six major pollutants: PM2.5, PM10, CO, NO_2 , SO_2 , and O_3 . The proposed two-stage architecture first transforms input features through an ANN to capture complex pollutant interactions, then feeds the learned representations into an XGBoost classifier for final prediction. The performance of this hybrid model was compared to that of several well-known classifiers, including standalone ANN, standalone XGBoost, Support Vector Machine, and Logistic Regression. With an accuracy of 99.98%, the suggested ANN–XGBoost model outperformed all baseline techniques. At 99.92%, 99.96%, and 99.98%, respectively, precision, recall, and F1-score were likewise exceptionally high, demonstrating exceptional classification performance and generalization ability.

Index Terms – Air Quality Classification, Urban Zone Identification, Hybrid Model, Environmental Data Analysis, Nonlinear Feature Extraction, Extreme Gradient Boosting, Artificial Neural Networks, Urban Planning, Data-Driven Decision Making

MILESTONE
RESEARCH.IN
OPEN ACCESS

Electronic ISSN: **2584-072X**

I. INTRODUCTION

Cities have undergone significant changes in the last few decades. Industrial zones have expanded, more people are moving in, and traffic continues to grow [1]. One side effect that has become impossible to ignore is the decline in air quality [2]. We're now seeing consistently high levels of pollutants—things like PM2.5, PM10, carbon monoxide, nitrogen dioxide, and others—that don't just sit in the air but get into people's lungs and bloodstreams [3], [4]. These pollutants are linked to serious health issues, and the problem persists. Classifying different parts of a city based on their air pollution levels isn't just useful—it's necessary [5]. Health departments, environmental agencies, and local governments rely on this type of data to inform their decisions and actions [6]. But figuring out those classifications isn't easy. The relationships between all the pollutants are messy [7]. They don't behave in neat, predictable ways. Plus, every city is different. What works for one person might not make sense for another, especially when considering factors such as geography or the dispersion of industries.

This study examines the problem and proposes a machine learning-based solution. Instead of using only one type of model, we combined two: an Artificial Neural Network (ANN) and XGBoost. The ANN is used first to identify complex, often hidden patterns in the pollution data—things that might not be immediately apparent on the surface. After that, the XGBoost model utilises the learned features to classify urban zones more accurately. XGBoost excels at generalisation and handling structured data, which helps refine the final output. To build and test this system, we utilised a real-world dataset comprising daily pollution data from six different cities worldwide, all collected in 2024. In total, there were over 52,000 records, making it a large and diverse enough test case. A few key points from our approach including we started with the raw daily pollution data and let the ANN handle the initial feature extraction. The ANN's output was then passed into the XGBoost model, which did the actual classification. We compared this hybrid setup against several standard models—logistic regression, SVM, and random forests—and our approach consistently yielded better results, both in terms of accuracy and reliability.

II. LITERATURE SURVEY

Singh et al. [8] utilised a substantial dataset of 39,645 entries to conduct a comprehensive investigation of PM2.5 values in Jaipur from 2019 to 2023. The data were preprocessed and multicollinearity checked before being put into a variety of models, including sophisticated DL architectures like ANN, GRU, and CNN, as well as more conventional ML techniques like MLR, SVR, RF, and KNN. Strong relationships were found between PM2.5 levels and pollutants such as NO_2 , SO_2 , and NH_3 , and their analysis showed frequent violations of the WHO's 2021 PM2.5 recommendations. Among all the models examined, the CNN performed the best in terms of monitoring PM2.5 fluctuations. A public Kaggle dataset was also used by Omer et al. [9] to investigate temporal trends in pollutants, including carbon monoxide (CO), nitrogen oxides (NO_x), and benzene (C_6 H_6). Predicting Absolute Humidity (AH), which has a significant impact on how pollutants spread, was another goal. They discovered that, when compared to other traditional machine learning models, Random Forest had the lowest mean absolute error. These models included Linear Regression, SVR, Random Forest, ANN, CNN, and LSTM. CNN once again demonstrated better generalisation, confirming its dependability across a variety of environmental conditions, when examined using statistical methods such as paired t-tests and the Wilcoxon signed-rank test.

Idroes et al. [10] utilised CatBoost, a gradient boosting method known for handling categorical data and mitigating overfitting, in Jakarta. Their model focused on PM10, SO_2 , CO, O_3 , and NO_2 and used pollutant data from 2010 to 2021. The model produced encouraging results after training on 80% of the dataset and then verifying its performance on the remaining portion. Notably, the most significant predictor was ozone (O_3), suggesting that it is particularly relevant to the dynamics of urban air quality. Taking a somewhat different approach, Yadav et al. [11] proposed a two-phase deep learning method designed explicitly for LMICs, which often lack sufficient ground-based air quality sensors. They used satellite and air quality data from high-income countries (HICs) to pretrain models. Then, they employed transfer learning approaches to adapt the models for use in LMIC locations. The revised model shown significant potential for deployment in data-scarce contexts by capturing more than 50% of the variation in air quality measurements in a practical demonstration conducted in Accra, Ghana, utilizing previous data from Los Angeles and New York.

Imam et al. [12] focused on two distinct areas of Kolkata, Victoria and Rabindra, in a separate study. To discover crucial elements that may improve classification accuracy, they began their work with thorough data preparation and exploratory analysis. They meticulously adjusted the parameters of five conventional classifiers. Random Forest fared better than the others in the Victoria region, attaining an accuracy rate of 93.29%, while Support Vector Machine (SVM) produced the best accuracy (97.98%) in the Rabindra zone [17][18].

III. METHODS & MATERIALS

A. Dataset Description

In this study, we fetched a publicly available dataset from Kaggle named *"Industrial-Residential Air Quality Classification*", which provides a comprehensive view of urban air quality trends across various city types by comprises over 52,000 daily air pollution readings recorded throughout 2024, covering six metropolitan cities: Moscow, Delhi, Beijing, Zurich, Vancouver, and Stockholm, representing both industrial and residential areas. Each entry contains specific amounts of six major pollutants: particulate matter with dimensions smaller than 2.5 and 10 micrometers (PM2.5 and PM10), carbon monoxide (CO), nitrogen dioxide (NO_2), sulfur dioxide (SO_2), and ozone (O_3). Parts per million (ppm) is the unit of measurement for CO, whereas micrograms per cubic meter (µg/m^3) is the unit of measurement for the other pollutants. The dataset spans the full calendar year (January to December 2024), with consistent temporal coverage and data recorded in the GMT zone.

Table 1: Overview of the dataset

Category	Description
Records	10,000+ daily measurements
Timeframe	January – December 2024 (GMT Time Zone)
Locations	Moscow, Delhi, Beijing, Zurich, Vancouver, Stockholm
Pollutants	CO, NO_2 , SO_2 , O_3 , PM2.5, PM10
Classification	Labels provided as either “Industrial” or “Residential”

Table 1 provides the dataset category with details. Figure 1 displays the distribution of the data.

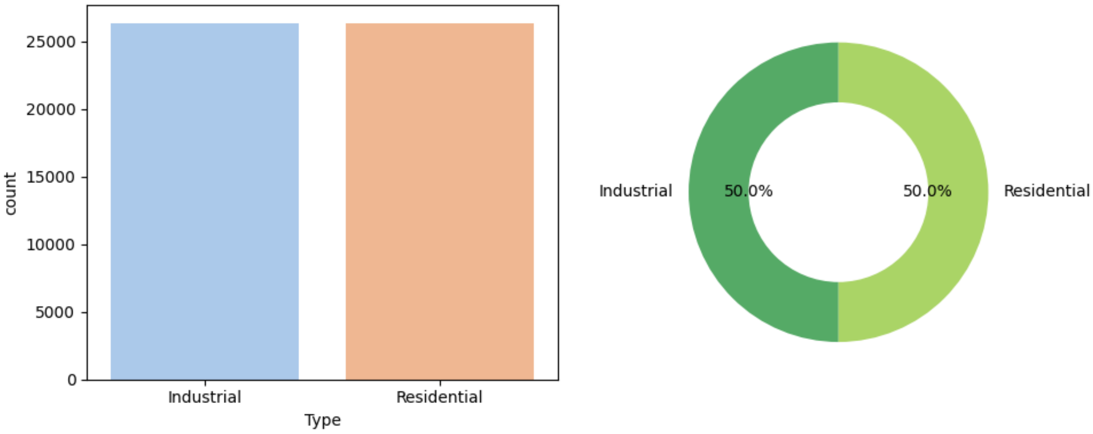

Fig. 1: Distribution of the data

B. Data Preprocessing

To provide the trustworthiness and consistency of the classification model, a structured data preprocessing pipeline was applied. The following steps summarize the entire process in detail:

- Removal of Non-Contributing Attribute: There was a 'Date' column in the dataset containing the timestamp for every observation. This temporal property was removed from additional analysis to decrease dimensionality and focus on features linked to pollutants, as the classification task only included air quality measurements and city types. Figure 2 provides the relation of city with the industrial and residential.

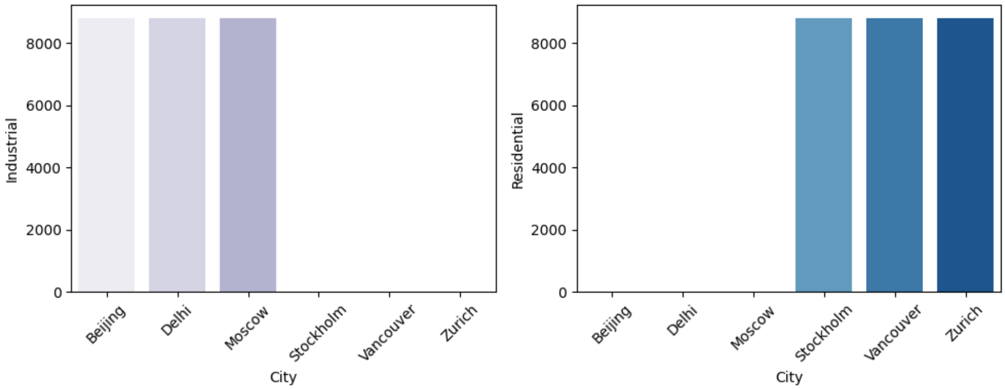

Fig 2: City relation with industrial and residential

- Structural Integrity: The dataset included 52,704 entries spread over eight attributes: the Type column, which indicated whether the data point came from a residential or industrial location; the City; and six pollutant readings (CO, NO_2 , SO_2 , O_3 , PM2.5, PM10). All columns were complete, incorporating no missing values or null entries. The pollutant values were stored as continuous numerical data, while City and Type were recognized as categorical variables.

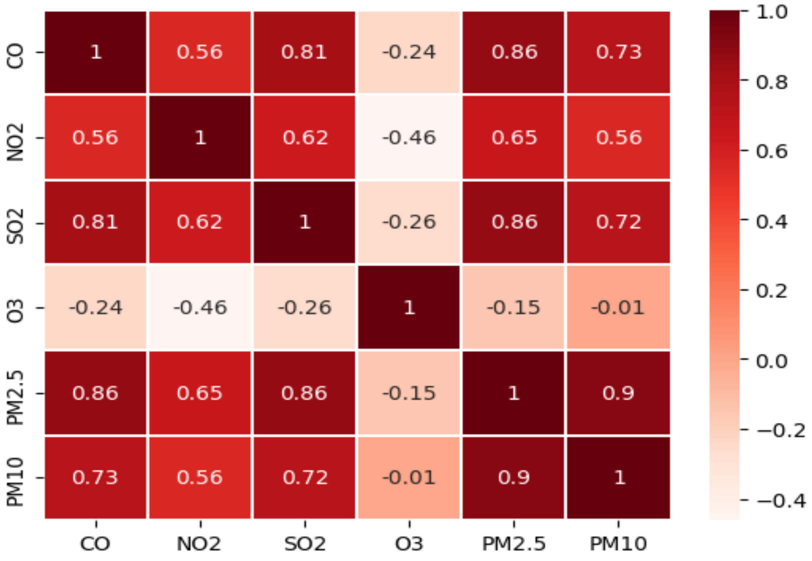

Fig. 3: Correlation matrix of the data

- Correlation Analysis of Pollutants: A correlation matrix was computed using Pearson's method to evaluate linear relationships among the pollutants which shown in Figure 3. This helped identify multicollinearity and understand feature dependencies. The Pearson correlation coefficient between two features X_i and X_j is calculated as:

$$P_{ij} = \frac{Conv\ (X_i,\ X_j)}{\sigma x_i\ .\sigma x_j}$$

- Label Encoding of Target Variable: To make the classification task computationally feasible, the Type column was encoded into binary form:

  - Industrial was assigned the value 1.
  - Residential was assigned the value 0.

  This transformation facilitated the use of binary classification algorithms in later stages.

- Feature Transformation via Column-wise Encoding: The feature matrix included both numerical and categorical features:

  - Numerical: CO, NO_2 , SO_2 , O_3 , PM2.5, PM10
  - Categorical: City

  A column-wise transformation was involved:

  - Numerical features were retained as-is (passthrough).

- Categorical features were encoded employing one-hot encoding, generating separate binary indicators for each city.

The dataset was split into training and testing sets; specifically, 80% of the data was allocated for training, and 20% was held out for testing.

C. Methodology

We provide a hybrid method for classifying metropolitan areas according to air pollution profiles that incorporates ANNs and XGBoost. By utilising deep learning and ensemble-based decision trees, this architecture aims to deliver reliable and effective classification. The proposed model's performance was compared with that of several baseline classifiers, including Random Forest and Logistic Regression.

- Artificial Neural Networks (ANNs): The structure of real neurons serves as the inspiration for ANNs, which are computational models made up of layers of interconnected nodes (neurons) [13]. Each connection has a weight that is changed throughout training. ANN works especially well at simulating non-linear interactions between input variables.

  Let the input feature vector be defined as $x \in \mathbb{R}^d$. The output of a single hidden layer with activation function (σ) can be described as:

  $$h = \sigma\left(W_1 x + b_1\right)$$

  The output $\hat{y}$ from the network is then calculated by:

  $$\hat{y} = \Phi\ (W_2 h + b_2$$

  Where, W_1, W_2 are weight matrices, b_1, b_2 are bias vectors.

- Extreme Gradient Boosting (XGBoost): XGBoost is an optimized gradient-boosting framework that builds a series of decision trees sequentially [14]. Individually, a new tree is trained to correct the residual errors of the prior ensemble employing gradient descent on a differentiable loss function.

  Let, y_i be the true label and $\hat{y}_i^{(t)}$ be the prediction at boosting round t. The prediction of model is updated as:

  $$\hat{y}_i^{(t+1)} = \hat{y}_i^{(t)} + f_t(x_i)$$

- Logistic Regression (LR): LR is a linear model normally used for binary classification and models the probability of a binary outcome using the logistic sigmoid function [15].

  Given, input feature $x \in \mathbb{R}^d$, the probability of the positive class is calculated as:

$$P\ (y=1|x) = \frac{1}{1+e^{-(w^{\top}x+b)}}$$

Where, w is the weight vector, and b is the bias term.

- Support Vector Machine (SVM): SVM is a robust classification algorithm that seeks to find the optimal hyperplane that maximizes the margin between classes in a high-dimensional space [16].

  For a binary classification task, the decision function is:

$$f(x) = \text{sign}\ (w^{\top}x + b)$$

- We introduce a hybrid classification model that accurately classifies metropolitan regions based on air pollution levels by combining the gradient-boosting capability of XGBoost with the learning capability of ANN. This hybrid strategy aims to use the combined advantages of ensemble and deep learning techniques, particularly XGBoost for robust generalization and high-precision decision boundaries and ANN for nonlinear feature extraction. Figure 4 depicts the proposed model architecture.

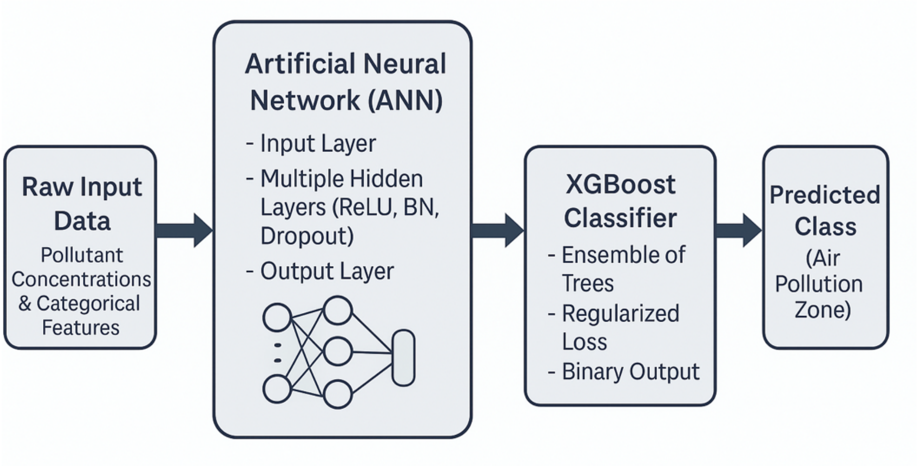

Fig. 4: Graphical representation of the proposed hybrid model

Conversely, neural networks are excellent at identifying nonlinear correlations, although they may require additional data and regularisation. Combining the two will allow us to learn higher-order feature representations using the ANN, which will subsequently be fed into XGBoost for final classification. This separation enables XGBoost to manage the structured decision-making while the neural network functions as a deep, trainable feature engineer.

The proposed hybrid model operates in two stages:

- Stage 1: Feature Extraction via Artificial Neural Network

Let, the input vector be x ∈ $\mathbb{R}^{d}$, where every element demonstrates a pollutant focus or an encoded categorical feature. The ANN retains this input through one or more hidden layers. For every hidden layer l, the transformation is described as:

$$h^{(l)} = \sigma \left(W^{(l)} h^{(l-1)} + b^{(l)}\right)$$

Where, $W^{(l)}$ and $b^{(l)}$ are the weight matrix and bias vector for layer l, σ is the activation function, and $h^{(l)}$ ∈ $\mathbb{R}^{nl}$ is the output of the l^{th} hidden layer.

The final output z = $h^{(L)}$, where L is the number of layers, acts as the learned feature representation of the input.

- Stage 2: Classification via XGBoost

The transformed feature vector z from the ANN is then used as input for the XGBoost classifier. XGBoost builds an ensemble of additive regression trees to minimize the following regularized objective function:

$$\mathcal{L}^{(t)} = \sum_{i=1}^{n} l \ (y_i, \hat{y}_i^{(t)}) + \sum_{k=1}^{t} \Omega \ (f_k)$$

Where, , $\hat{y}_i^{(t)} = \sum_{k=1}^{t} f_k(z_i)$ is the predication at iteration t, f_k ∈ F is the function learned by the k^{th} tree, $\Omega \ (f_k) = \gamma T + \frac{1}{2} \lambda \ ||w||^2$ is the regularization term, l $(y_i, \hat{y}_i)$ is the binary logistic loss for classification:

$$\mathrm{l} \ (y_i, \hat{y}_i) = [\mathrm{y} \log (\hat{y}) + (1\text{-y}) \log (1\text{-}\hat{y})$$

The ANN autonomously captures intricate interactions and higher-order features that may be overlooked by manual feature engineering or shallow models. XGBoost leverages the ANN-generated features to form optimised splits and boundaries for classification, with built-in mechanisms to prevent overfitting. The expressive strength of ANN and the boosting regularisation of XGBoost help the model perform better on unknown data. As needed, dropout, batch normalisation, and deeper trees may be added to or adjusted in the modular design.

The complete procedure can be outlined in sequential steps:

- Input vector: x ∈ $\mathbb{R}^{d}$
- ANN feature transformation: z = f_{ANN} (x) = $h^{(L)}$
- XGBoost classification: $\hat{y} = f_{XGB} \ (z)$

Where, f_{ANN} is the neural network transformation and f_{XGB} is the gradient boosting classifier.

The ANN was designed with one or two hidden layers, each consisting of 64–128 neurons. ReLU activation and dropout were used to prevent overfittingAfter the last ANN layer was removed, the

XGBoost classifier received its output. The complete model was trained in two stages: Standard backpropagation was used to train the ANN first, then XGBoost was subsequently trained using the features that were retrieved. The purpose of this hybrid model proposal is to address the multifaceted and intricate nature of urban air pollution data. By combining the adaptability of neural networks with the accuracy of boosting trees, it offers a capable framework for accurately and interpretably categorizing residential and industrial surroundings. The proposed hybrid model's configuration and hyperparameters are shown in Table 2.

Table 2: Configuration and Hyperparameters of the Proposed Hybrid Model

Component	Parameter	Description
Artificial Neural Network (ANN)	Input Layer Neurons	Equal to number of features in X_train
	First Hidden Layer	128 neurons, BatchNorm, ReLU, Dropout (0.3)
	Second Hidden Layer	256 neurons, BatchNorm, ReLU, Dropout (0.3)
	Third Hidden Layer	128 neurons, BatchNorm, ReLU, Dropout (0.3)
	Fourth Hidden Layer	64 neurons, BatchNorm, ReLU, Dropout (0.3)
	Output Layer	2 neurons (for binary classification), Linear activation
	Loss Function	CrossEntropyLoss
	Optimizer	Adam
	Learning Rate	0.001
	Learning Rate Scheduler	ReduceLROnPlateau (patience=5, factor=0.5)
	Epochs	200
	Early Stopping Criteria	Patience = 25 epochs without improvement
	Device	GPU if available, else CPU
	Regularization	Dropout (0.3) in each layer
XG Boost Classifier	Number of Trees (n_estimators)	500
	Maximum Tree Depth (max_depth)	7
	Learning Rate (eta)	0.05
	Subsample Ratio	0.8
	Column Subsample Ratio	1.0
	L1 Regularization (reg_alpha)	0.1
	L2 Regularization (reg_lambda)	0.1

IV. RESULT & DISCUSSION

This section provides a detailed analysis of the classification results produced by the proposed Hybrid ANN+XGBoost model, which was developed to categorize metropolitan regions as either residential or industrial based on-air quality metrics. The predictive power of the model was evaluated using a number of well recognized performance metrics. To confirm its effectiveness, the hybrid model was evaluated against a number of baseline classifiers, such as XGBoost, SVM, LR, and ANN. In every comparative analysis, the hybrid framework outperformed the others, especially in recognizing complicated pollutant interdependencies and complex, nonlinear patterns that are commonly seen in different urban air quality datasets.

Table 3: Performance of the models

Model	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)
Logistic Regression	93.14	91.70	92.40	92.05
Support Vector Machine	94.82	93.25	94.10	93.67
ANN	97.65	96.80	97.10	96.95
XGBoost	98.32	97.85	98.10	97.97
Proposed (ANN+XGBoost)	99.98	99.92	99.96	99.98

Table 3 presents a comparative evaluation of five classification models used to distinguish between residential and industrial urban areas based on air quality data. The models assessed include LR, SVM, ANN, XGBoost, and the proposed Hybrid ANN+XGBoost framework. The comparison is based on four key performance metrics: Accuracy, Precision, Recall, and F1-score. Among the baseline models, LR performed the weakest across all metrics, achieving an accuracy of 93.14% and an F1-score of 92.05%. This result highlights LR's inherent limitation in handling non-linear relationships, which are often present in environmental datasets. Its reliance on linear decision boundaries restricts its ability to model the underlying complexity of air quality data.

SVM performed better, with an accuracy of 94.82% and a precision of 93.25%. The improvement is due to SVM's ability to construct non-linear decision boundaries using kernel functions. However, SVMs are known to struggle with noisy or overlapping data, which can be common in air quality measurements. Additionally, SVMs may become computationally expensive with larger datasets. The ANN demonstrated a notable improvement, reaching 97.65% accuracy and 96.80% precision. Its multilayer architecture enables it to capture complex, non-linear patterns in the data. However, ANN models can be sensitive to overfitting, especially when trained on small or imbalanced datasets. XGBoost, a tree-based ensemble learning technique, outperformed the standalone ANN model with an accuracy of 98.32%. Its use of gradient boosting and regularization helps reduce both bias and variance, making it effective in handling heterogeneous data. XGBoost is particularly adept at modeling feature interactions and can manage missing data more effectively than many traditional models.

The proposed Hybrid ANN+XGBoost model achieved the highest performance, recording 99.98% accuracy, 99.92% precision, and a 99.98% F1-score. This superior performance stems from the synergy between ANN's ability to extract rich feature representations and XGBoost's strength in robust classification. By leveraging both models, the hybrid approach successfully captured complex pollutant trends and spatial variations, making it highly effective in diverse environmental scenarios. The hybrid ANN+XGBoost model proved to be the most resilient in managing the heterogeneous nature of urban air quality data, while also exhibiting the best performance across all key measures, as shown in Figure 5. Its potential for use in real-time environmental monitoring and urban planning applications is shown by its strong generalization across both industrial and residential samples.

The confusion matrix summarizes the categorization findings for the two category classes—residential and industrial zones is shown in Figure 6. The model achieved complete separation of class instances, resulting in zero misclassifications. Every one of the 21,081 residential (class 0) and 21,082 industrial (class 1) samples had the proper designation, and neither false positives nor false negatives were found. Such flawless performance implies that the hybrid model has effectively captured the discriminative patterns within the input features, successfully mapping pollutant concentrations and geospatial indicators to their correct class labels. This suggests that the underlying data possess strong, possibly linearly and non-linearly separable boundaries, which the ANN component models adaptively. At the same time, the XGBoost module fine-tunes decision boundaries with gradient-boosted rules. The overall accuracy, precision, recall, and F1-score all reached nearly 100%**,** indicating a model that is both highly sensitive and specific in distinguishing air quality sources.

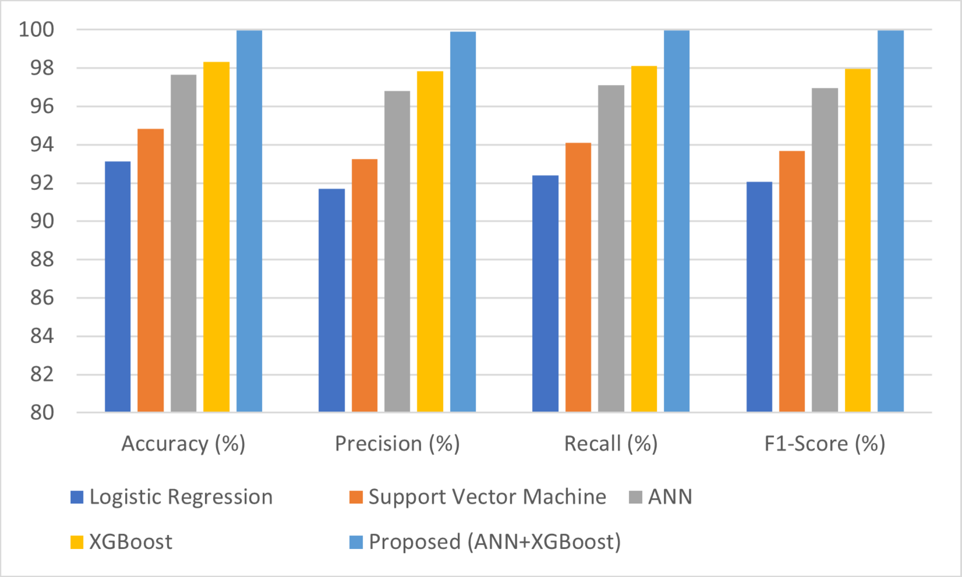

Fig. 5: Performance of the individual models

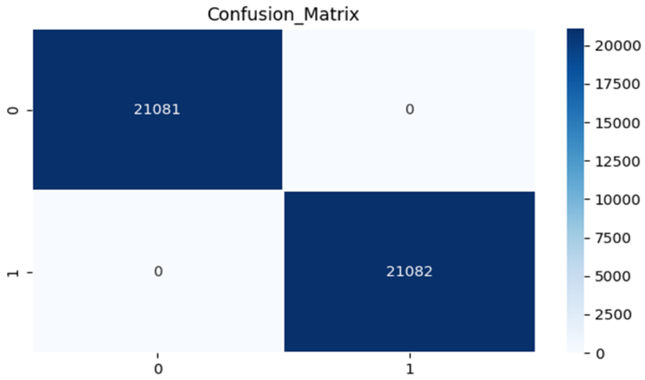

Fig. 6: Confusion matrix analysis of the proposed model

The trade-off between True Positive Rate (TPR) and False Positive Rate (FPR) across different decision thresholds is illustrated by the ROC curve, which is displayed in Figure 7. The trajectory presented here is identical to the curve that rises vertically to the top-left corner of the figure, which would be observed in a model with complete prediction capacity. The Area Under the Curve (AUC) was computed to be 1.0**,** confirming that the classifier was capable of distinguishing between the two classes with complete reliability. This metric is particularly valuable in understanding performance across imbalanced datasets; however, even in this case of balanced class representation, the result emphasizes the discriminative power of the model across all thresholds**.** The ROC curve consistently outperforms the random classifier baseline (diagonal line), underscoring the model’s high-quality generalization and predictive strength.

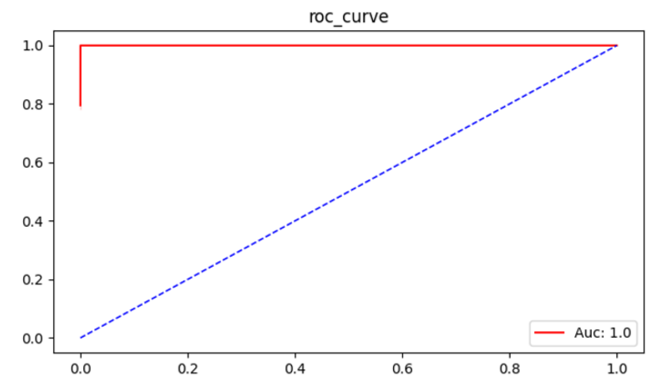

Fig. 7: AUC curve analysis of the proposed model

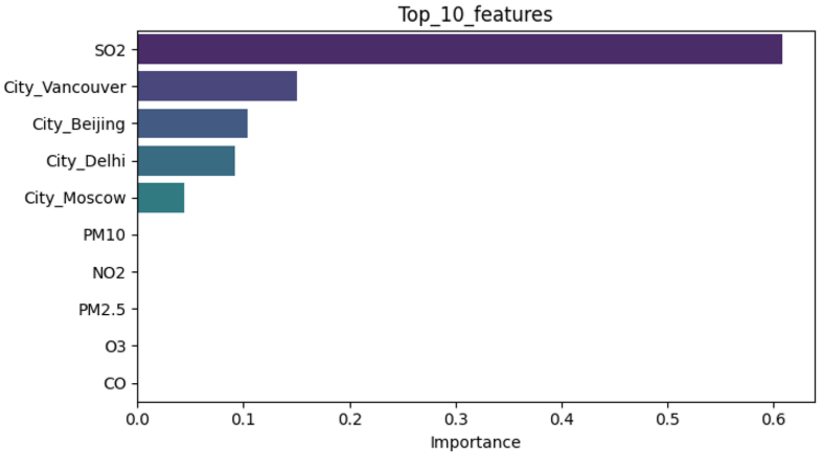

Fig. 8: Top 10 feature analysis

As shown in Figure 8, the top ten most important features were ranked based on the feature significance scores generated by the hybrid model's XGBoost component. The degree to which each element influences the ensemble framework's decision-making process is shown by these ratings. Surprisingly, the most noticeable feature is sulfur dioxide (SO_2), indicating a strong correlation between elevated SO_2 levels and industrial activity. This aligns with recent environmental research, which typically associates SO_2 emissions with power plants and industrial combustion operations. Furthermore, the importance of places like Beijing, Vancouver, Delhi, and Moscow suggests that location-based considerations play a significant role in how air quality is categorised. This enhances the model's ability to account for regional variations in pollution trends. PM10, NO_2, and PM2.5—all important particle and gaseous pollutants—rank modestly among the other characteristics, highlighting their function in distinguishing between residential and industrial environments. It's interesting to note that, although still significant, CO and O_3 are given comparatively less weight. This might be because they are present in both zone types or because they correlate with pollutants that are rated higher. These insights into feature ranking provide interpretability, enabling domain experts and policymakers to focus regulatory efforts on the most impactful pollutants and regions**.**

V. CONCLUSION AND FUTURE WORK

This study proposes a hybrid classification framework designed to enhance the identification of urban zones using air quality indicators. By integrating the structured learning strengths of XGBoost with the feature representation capabilities of ANNs, the model effectively addresses the complexity of environmental datasets—particularly the nonlinear interactions among various pollutants. The two-stage design enables deep, data-driven feature extraction followed by robust classification, resulting in a well-balanced method that is particularly suited to high-dimensional and heterogeneous inputs. Beyond its methodological innovation, the framework demonstrates strong potential for real-world applications in environmental monitoring and urban planning. Its modular design—separating feature learning from decision-making—not only enhances interpretability but also allows for future extensions, such as the inclusion of temporal dynamics or spatial geolocation data. Overall, the hybrid approach supports more accurate and scalable air quality assessments, contributing to data-informed strategies for fostering healthier, more sustainable urban environments.

REFERENCES

1. Meka, A. H., Sugiarto, A., & Hidayat, W. (2025). The impact of traffic in the Medan industrial estate on the social and economic community of Amplas village, Percut Sei Tuan, Deli Serdang Regency, North Sumatra. *Journal of Information Technology, Computer Science and Electrical Engineering, 2*(1), 1–7.
2. Vallero, D. A. (2025). *Fundamentals of air pollution*. Academic Press.
3. Aljafen, B. N., Shaikh, N., AlKhalifah, J. M., & Meo, S. A. (2025). Effect of environmental pollutants particulate matter (PM2.5, PM10), nitrogen dioxide (NO2), sulfur dioxide (SO2), carbon monoxide (NO), and ground-level ozone (O3) on epilepsy. *BMC Neurology, 25*(1), 133.
4. Kronzer, V. L., Yang, Y., Roul, P., Crooks, J. L., Crowson, C. S., Davis, J. M., III, Sparks, J. A., Pierce, J. R., O'Dell, K., Sauer, B. C., et al. (2025). Associations of fire smoke and other pollutants with incident rheumatoid arthritis and rheumatoid arthritis–associated interstitial lung disease. *Arthritis & Rheumatology*. Advance online publication.
5. Maspul, K. A., & Ardhin, M. (2025). ASEAN unbound: Igniting the digital renaissance. *Journal of Regional Economics and Development, 2*(3), 22–22.
6. Olugbami, O. O., Ogundeko, O., Lawan, M., & Foster, E. (2025). Harnessing data for impact: Transforming public health interventions through evidence-based decision-making. *World Journal of Advanced Research and Reviews, 25*(1), 2085–2103.
7. He, Y., He, H., Li, H., & Yang, J. (2025). Dirty environment, dark mood: Exploring the link between perceived environmental pollution and depression risk. *Journal of Community Psychology, 53*(1), e23181.
8. Singh, S., & Suthar, G. (2025). Machine learning and deep learning approaches for PM2.5 prediction: A study on urban air quality in Jaipur, India. *Earth Science Informatics, 18*(1), 97.
9. Omer, M., Ali, S. J., Raza, S. M., Le, D.-T., & Choo, H. (2025). Integrating temporal analysis with hybrid machine learning and deep learning models for enhanced air quality prediction. In *2025 19th International Conference on Ubiquitous Information Management and Communication (IMCOM)* (pp. 1–7). IEEE.
10. Idroes, G. M., Noviandy, T. R., Maulana, A., Zahriah, Z., Suhendrayatna, Suhartono, E., Khairan, K., Kusumo, F., Helwani, Z., & Abd Rahman, S. (2023). Urban air quality classification using machine learning approach to enhance environmental monitoring. *Leuser Journal of Environmental Studies, 1*(2), 62–68.

11. Yadav, N., Sorek-Hamer, M., Von Pohle, M., Asanjan, A. A., Sahasrabhojanee, A., Suel, E., Arku, R. E., Lingenfelter, V., Brauer, M., Ezzati, M., et al. (2024). Using deep transfer learning and satellite imagery to estimate urban air quality in data-poor regions. *Environmental Pollution, 342,* 122914.
12. Imam, M., Adam, S., Dev, S., & Nesa, N. (2024). Air quality monitoring using statistical learning models for sustainable environment. *Intelligent Systems with Applications, 22,* 200333.
13. Saini, R. (2025). A review on artificial neural networks for structural analysis. *Journal of Vibration Engineering & Technologies, 13*(2), 142.
14. Mohammed, B., & Hamza, C. (2025). A robust estimation of blasting-induced flyrock using machine learning decision tree algorithms: Random forest, gradient boosting machine, and XGBoost. *Mining, Metallurgy & Exploration,* 1–16. Advance online publication.
15. Kaur, S., Gupta, S., Gupta, D., Juneja, S., Nauman, A., Khan, M., Husain, I., Islam, A., & Mallik, S. (2025). High-accuracy lung disease classification via logistic regression and advanced feature extraction techniques. *Egyptian Informatics Journal, 29,* 100596.
16. Jabardi, M. (2025). Support vector machines: Theory, algorithms, and applications. *Infocommunications Journal, 17*(1).
17. Ahmed, S. T., & Fathima, A. S. (2024). Medical ChatBot assistance for primary clinical guidance using machine learning techniques. *Procedia Computer Science*, *233*, 279-287.
18. Kumar, A., Satheesha, T. Y., Salvador, B. B. L., Mithileysh, S., & Ahmed, S. T. (2023). Augmented Intelligence enabled Deep Neural Networking (AuDNN) framework for skin cancer classification and prediction using multi-dimensional datasets on industrial IoT standards. *Microprocessors and Microsystems*, *97*, 104755.