Prepublished copy

Nature Environment & Pollution Technology

This is a peer-reviewed prepublished version of the paper

Original Research

Deep Learning Approach for Evaluating Air Pollution Using the RFM Model

Jannah Mohammad [1], Mohammod Abul Kashem [2]

[1] Department of Computer Science and Engineering, Dhaka University of Engineering and Technology, Gazipur, Bangladesh; jmgodhuli@gmail.com
[2] Department of Computer Science and Engineering, Dhaka University of Engineering and Technology, Gazipur, Bangladesh; drkashemll@duet.ac.bd

† Corresponding author: Jannah Mohammad; jmgodhuli@gmail.com
https://orcid.org/0000-0002-0801-9553

Abstract: Air pollution is a required environmental and public health issue in India, with multiple municipalities repeatedly ranking among the most polluted in the world. This study leverages large datasets to construct a predictive model for forecasting air quality trends using a novel approach that integrates the Recency Frequency Monetary (RFM) model with deep learning. The research aims to efficiently quantify pollution events frequency and assess the impact of air quality variations on public health, offering a more flexible and adaptive system for air quality monitoring. As a result, a large volume of air quality data provided by RFM (Recency, Frequency, and Monetary) will be flexible and frequently handled and analyzed. In this research, the performance of the integrated RFM technology is examined using Python and Google Colab, and the simulation results are compared to air pollution information from neural networks for structures in additional data using existing air quality monitoring systems in India. Performance examination of both regression and classification techniques in RFM. The execution of RFM can be one of the models and its potential to enhance air quality monitoring and urban sustainability.

Key Words	Neural Network; Air pollution; India; Air Quality Index
DOI	https://doi.org/10.46488/NEPT.2025.v24i02.D1718 (DOI will be active only after the final publication of the paper)
Citation of the Paper	Jannah Mohammad 1, Mohammod Abul Kashem, 2025. Deep Learning Approach for Evaluating Air Pollution Using the RFM Model . *Nature Environment and Pollution Technology*, 24(2), D1718. https://doi.org/10.46488/NEPT.2025.v24i02.D1718

1. INTRODUCTION

Urban air pollution is an escalating problem that severely impacts public health and energy consumption. In North India, the burning of agricultural waste has degraded air quality, with pollutants like $PM_{2.5}$ affecting regions as far as the central Himalayas and even doubling air pollution levels in cities like Kathmandu due to cross-border pollution (Khanal et al., 2022). Rapid urbanization, especially in countries like India, makes air pollution issues more severe. Kanpur City, for instance, has serious air quality problems, leading the Indian government to create an Environmental Management Plan (EMP) (Gupta, 2008). This plan emphasizes the urgent need for effective strategies to reduce pollution and protect health. Many areas in India have pollution levels far above World Health Organization guidelines (Weagle and Martin, 2019), resulting in significant health problems and highlighting the need for better monitoring and prediction tools.

Effective air pollution monitoring involves measuring and analyzing pollutant levels to guide mitigation efforts. Traditional methods are being improved with advanced technologies, such as neural networks, which can predict pollution levels more accurately (Wesolowski, Suchacz, and Halkiewicz, 2006). A comparison of Indigenous Structures (Kumar et al., 2023) compared Support Vector Machines (SVM) with deep learning models, finding SVMs competitive but lagged in temporal dynamics. Research from institutions like the Spanish Ministry of Science and Education has shown that neural networks can provide detailed insights and forecasts (Ibarra-Berastegi et al., 2009). For example, the Greater Istanbul Area uses neural networks for short-term pollution predictions to help reduce emissions (Kurt et al., 2008), and research in Bilbao, Spain, uses multiple neural network models for forecasting (IbarraBerastegi et al., 2008). According to (Lisboa, 2002), Artificial Neural Networks (ANNs) are valuable for analyzing complex data and improving early disease detection.

Recent advancements in machine learning offer new ways to analyze air pollution. (Mohammad and Kashem, 2022) used the Recency, Frequency, Monetary (RFM) model with k-means clustering to compare pollution levels, achieving 50% accuracy in clustering. Integrating neural networks with the RFM model enhances data analysis and pattern recognition (Liao et al., 2022), which is useful for managing large and complex datasets (Mena et al., 2023). RFM analysis has been effective in customer segmentation for marketing (Anitha and Patil, 2022), but predicting air pollution with neural networks has faced challenges, particularly with short-term forecasts. In this research, The RFM model integrates neural networks to learn from the most relevant information, which can improve air quality prediction by evaluating regression and classification metrics.

This research paper proposes a unique approach by combining the RFM model with neural networks to enhance air pollution predictions and attain more in-depth insights into pollution conventions in India. The RFM model can evaluate pollution instances based on their frequency and intensity. Integrating neural networks is intended to increase prediction accuracy and provide a better knowledge of air quality patterns. This research evaluates the effectiveness of regression and classification techniques within the RFM model. Regression helps predict the quantity of pollution, while classification categorizes pollution levels. This combined approach allows for a thorough analysis and more manageable interpretation of pollution data.

The paper's organization is as follows: Section 2 covers data assemblage and exploratory analysis. Section 3 describes the AQI computation and the proposed methodology for operating neural networks and RFM models. Section 4 presents simulation consequences and assesses the performance of regression and classification procedures. Section 5 concludes with insights and propositions for prospective research.

2. EXPLORATORY DATA ANALYSIS

The research investigation collects air pollution data from 2015 to 2020 in India. The dataset includes the following cities: Amravati, Amritsar, Chandigarh, Delhi, Gurugram, Hyderabad, Kolkata, Patna, and Visakhapatnam. After collecting the data, an Exploratory Data Analysis (EDA) has been conducted to determine the dataset's structure and features. The subsequent efforts involved straining for null values and using data-cleaning techniques to provide the dataset's integrity and reliability for additional analysis. Latitude is followed by ° N (for the Northern Hemisphere). Longitude is followed by ° E (for the Eastern Hemisphere).

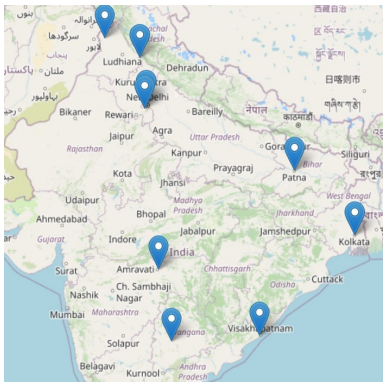

Fig. 1: Location map of India.

Table 1: LATITUDE AND LONGITUDE OF INDIAN CITIES.

City	Latitude	Longitude

Amravati	20.9374° N	77.7796° E
Amritsar	31.6340° N	74.8723° E
Chandigarh	30.7333° N	76.7794° E
Delhi	28.7041° N	77.1025° E
Gurugram	28.4575° N	77.0263° E
Hyderabad	17.3850° N	78.4867° E
Kolkata	22.5726° N	88.3639° E
Patna	25.5941° N	85.1376° E
Visakhapatnam	17.6868° N	83.2185° E

2.1 Dataset description

The dataset provides comprehensive data for various cities, with no missing values in the City and Date columns, ensuring complete information for these entries.

$PM_{2.5}$ and PM_{10}: These columns contain concentrations of fine particulate matter $PM_{2.5}$ and coarser particulate matter

PM_{10} units of microgram m 3, The $PM_{2.5}$ and PM_{10} columns, representing concentrations of fine particulate matter and coarser particulate matter respectively, are critical for understanding air quality issues in India. Due to factors like industrial operations, automobile emissions, and seasonal crop burning, $PM_{2.5}$ and PM_{10} levels are a significant concern in major Indian cities (World Health Organization, 2021). In this research, the dataset reveals 4,598 missing values in the $PM_{2.5}$ column and 11,140 missing values in the PM_{10} column.

The NO, NO_2, and NO_x columns units of parts per million (ppm)measure nitric oxide, nitrogen dioxide, and total nitrogen oxide levels. These nitrogen oxides, produced from combustion processes and atmospheric reactions, are important pollutants to monitor (U.S. Environmental Protection Agency, 2021). The data shows 3,582 missing values in the NO column, 3,585 missing values in the NO2 column, and 4,185 missing values in the NOx column. Additionally, the NH3 (ammonia) units of parts per million (ppm) columns has 10,328 missing values.

The CO column units of parts per million (ppm), which tracks carbon monoxide levels, is also vital due to the high levels of CO in India caused by traffic emissions, industrial activities, and biomass burning, leading to significant air quality and health issues in urban areas (Central Pollution Control Board, 2023). This column has 2,059 missing values.

The SO_2 column units of parts per million (ppm), measuring sulfur dioxide levels, reflect the impact of coal-fired power plants, industrial operations, and vehicular emissions on air pollution and health (Ministry of Environment, Forest and Climate Change, 2022), with 3,854 missing values. The O_3 column records ground-level ozone levels in units of parts per million (ppm), an air pollutant linked to respiratory and cardiovascular risks (European Environment Agency, 2022), and has 4,022 missing values. The Benzene, Toluene, and Xylene columns units of parts per million (ppm)represent volatile organic compounds (VOCs) that pose soundness threats and donate to ozone construction. These VOCs are primarily radiated by industrial operations, conveyance emissions, and chemical solvents in India (Agency for Toxic Substances and Disease Registry, 2021), with 5,623 missing values in the Benzene column, 8,041 in the Toluene column, and 18,109 in the Xylene column.

Lastly, the AQI (Air Quality Index) column provides a composite index reflecting overall air quality based on the concentrations of the pollutants mentioned above. There are 4,681 missing values in the AQI column, which corresponds to the same number of missing values in the *AQIBucket* column that categorizes AQI into qualitative buckets.

Data Cleaning: An air pollution dataset contains missing values for pollutant concentrations, which are set to 0 to indicate that no measurement has been taken. The dataset contains comprehensive air quality data for several Indian cities, with complete entries for the City and Date columns. However, pollutant concentration data exhibited significant missing values. It found 29,531 rows and 16 columns of air pollution data; substituting null values with a predetermined value is a standard method for dealing with missing data points.

3. METHODS

This analysis operates the Recency Frequency Monetary (RFM) model, which is commonly used in customer segmentation (Anitha and Patil, 2022), to assess air pollution occurrences. The model is adapted to measure the recency, frequency, and severity (monetary impact) of pollution occurrences. A deep learning neural network is employed to predict air quality using the RFM data, with both regression and classification techniques tested for accuracy. The Air Quality Index (AQI) is an important instrument for expressing the health effects of air pollution. It employs classifications to describe the severity of air quality, ranging from Good to Hazardous (US EPA, 2014).

.

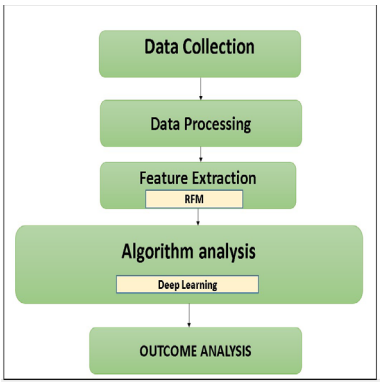

Fig. 2: Propose methodology.

Table 2: Air Quality Index.

AQI status	Index Range	Characterization of Air Quality Index
GOOD	0 to 50	The level of air pollution is acceptable.
Moderate	51 to 100	The air is in good condition. But some individuals might be concerned.
Unhealthy for sensitive groups (Caution)	101 to 150	Health issues may affect members of sensitive groups. Less likely to be impacted is the general public.
Unhealthy	151 to 200	Members of weak residents may face more intense soundness consequences than available residents' participants.
Very Unhealthy	201 to 300	Everybody is additionally at hazard for negative soundness repercussions.
Hazardous (ExtremelyUnhealthy)	301 to 500	Everyone is additionally potential to be affected by circumstances in an emergency.

3.1. RFM:

RFM (Recency, Frequency, Monetary) principles can be adapted to analyze air pollution data by considering analogous factors:

Recency: Recency of air pollution station from location. In air pollution, this could refer to the timeliness of the air quality measurements or the most recent pollution events area.

Frequency: The frequency of pollution events or measurements. This could mean how often the air quality measurements are taken or the number of times a particular pollution level has been recorded.

Monetary: Although not directly applicable, this could be adapted to represent the severity or impact of the pollution. The severity of pollution levels and their environmental and health consequences. Perhaps it is appropriate to be based on pollutant concentration levels and their subsequent health or environmental effects. The customer segmentation approach successfully assesses and categorizes customers by combining ABC and RFM strategies (Liu et al., 2019). (Panus et al.˅ 2016) employ ABC classification to identify commodities based on their

economic or technological relevance. While ABCXYZ analysis is described as an approach that integrates ABC and XYZ categories over two dimensions, similar to the BCG matrix, it is incapable of offering a detailed analysis (Teslenko et al., 2023). In this research, The RFM classification ranges help in categorizing the data into different air pollution levels of predicted pollution.

Table 3: RFM CLASSIFICATION RANGES AND DESCRIPTIONS.

Recency	Frequency	Monetary
A	0-50	Very low values or scores
B	51-100	Low values
C	101-150	Moderate-low values
X	151-200	Moderate values
Y	201-300	High values
Z	301-500	Very high values

3.2 Algorithm analysis:

Neural network

Layers of connected nodes, or neurons, make up neural networks. During training, the weights of each connection are changed. Neural network fundamental equations include the following: Weighted Sum: Each neuron computes a weighted sum of its inputs. If $x_1, x_2, \ldots, x_n$ are the inputs and $w_1, w_2, \ldots, w_n$ are the corresponding weights, the weighted sum z for a neuron is given by:

$$Z = \sum_{i=1}^{n} w_i x_i + b \qquad \ldots(1)$$

where b is the bias term (Goodfellow, Bengio, and Courville, 2016). Activation Function: The weighted sum z is then passed through an activation function to produce the neuron's output a. Common activation functions include the sigmoid function, tanh function, and Rectified Linear Unit. For a generic activation function , the output a is: $a = \sigma(z)$(Deng, 2014)

Forward Propagation: In a neural network with multiple layers, the output of each layer is passed as input to the next layer. For each layer l, the output $a^{(l)}$ is computed as :

$$\mathbf{a}^{(l)} = \sigma\left(\mathbf{W}^{(l)}\mathbf{a}^{(l-1)} + \mathbf{b}^{(l)}\right) \qquad \ldots(2)$$

where $w^{(l)}$ is the weight matrix for layer l, $\mathbf{a}^{(l+1)}$ is the output from the previous layer, and $B^{(l)}$ is the bias vector for layer l (LeCun et al. 2015). Loss Function: The performance of the neural network is evaluated using a loss function L, which measures the difference between the predicted output and the actual target. For a given set of predictions y and actual values y, the loss function L could be Mean Squared Error (MSE), Cross-Entropy, or another appropriate function (Bishop, 2006).

Back propagation: Back propagation is used to determine the gradients of the loss function for each weight and bias to train the network. Gradient descent or its derivatives are then used to update the weights and biases:

$$\mathrm{w}(l) \leftarrow \mathbf{w}(l) - \eta\frac{\partial L}{\partial \mathbf{w}I} \qquad \ldots(3)$$

$$b \leftarrow b - \eta \frac{\partial L}{\partial b} \qquad \ldots(4)$$

where η is the learning rate (Rumelhart, Hinton and Williams, 1986).

A. Predict Model performance metrics

The statistical Mean Absolute Error (MAE), Root Mean Square Error (RMSE), Mean Square Error (MSE), Relative Absolute Error (RAE) and coefficient of determination (R2).These metrics are widely recognized and utilized in various research contexts, including the development of hybrid models that integrate neural networks with traditional statistical methods for air quality forecasting (Zhang et al., 2019). Their application allows for a comprehensive evaluation of model performance, ensuring that the predictions are both accurate and meaningful (Huang et al., 2020). Moreover, in a study by (Yadav and Malik, 2020), neural network based predictive models for air quality indices demonstrated the effectiveness of these statistical performance metrics in enhancing prediction accuracy. Similarly, (Ma et al. 2020) conducted a comparative study of neural networks and traditional statistical models for air quality forecasting.

Mean Absolute Error (MAE): The average of the absolute errors is known as the mean absolute error (MAE). Understanding whether the amount of the error deserves concern or not is made easier by the fact that the MAE units are the same as the expected target

$$1/N \sum_{i=1}^{N} |w_i - w'_i| \qquad \ldots(5)$$

Mean Squared Error (MSE) - The term Mean Squared Error (MSE) refers to the average or mean of the square of the discrepancy between actual and estimated data.

$$1/N \sum_{i=1}^{N} (w_i - w'_i)^2 \qquad \ldots(6)$$

Here, n is the number of statements and w'i the indicated $w_i$true value.

The Relative Absolute Error (RAE) measures the total absolute error as a proportion of the total absolute difference between the true values and their mean. It provides a normalized view of the error, making it easier to understand the model's performance relative to a simple model that always predicts the mean of the target variable. The Relative Absolute Error (RAE) is defined as: RAE

$$= \frac{\sum_{i=1}^{N} |w_i - w'_i|}{\sum_{i=1}^{N} |w_i - \bar{w}|} \qquad \ldots(7)$$

where:

- N is the number of observations,
- w_i is the true value,
- w_i' is the predicted value,
- $\overline{w}$ is the mean of the true values.

Root Mean Squared Error (RMSE) - MSE's squared root.

$$\sqrt{1/N \sum_{j=1}^{N} (wj - w'j)^2} \qquad \ldots(8)$$

Squared Error (RSE) - The ratio of value from the square of the fundamental error.

$$\sum_{j=1}^{N}(w_j - w'_j)^2 \Big/ \sum_{j=1}^{N}(w_j - w)^2 \qquad \ldots(9)$$

R Squared Error: The proportion of the error's square.

$$R2 = 1\text{-}\left(\sum_{j=1}^{N}(w_j - w'_j)^2 / \sum_{j=1}^{N}(w_j - w)^2\right) \qquad \ldots(10)$$

Classification Metrics: When evaluating classification models, several metrics are commonly used to assess performance. These include Precision, Recall, F1-Score, Accuracy, and their corresponding averages (Macro Avg and Weighted Avg).

Precision: Precision is the ratio of correctly predicted positive observations to the total predicted positives. It determines the proportion of predicted positive cases that are correctly identified as positive.

$$\text{Precision} = \frac{\text{True Positives(TP)}}{\text{True Positives(TP)} + \text{False Positives(FP)}} \qquad \ldots(11)$$

Recall: Recall (also known as Sensitivity or True Positive Rate) is the ratio of correctly predicted positive observations to all observations in the actual positive class. The amount of positive instances are correctly identified?

$$\text{Recall} = \frac{\text{True Positives(TP)}}{\text{True Positives(TP)} + \text{False Negatives(FN)}} \qquad \ldots(12)$$

F1-Score: The F1-Score is the harmonic mean of Precision and Recall. It provides a single metric that balances both the concerns of false positives and false negatives.

$$\text{F1-Score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \qquad \ldots(13)$$

Support: Support refers to the number of actual occurrences of each class in the dataset. It is used to weigh the average of metrics such as Precision, Recall, and F1-Score. Accuracy

Accuracy is the ratio of correctly predicted observations to the total observations. It works well when the classes are well are balanced.

$$\text{Accuracy} = \frac{\text{TP + True Negatives (TN)}}{\text{Total Observations}} \quad \ldots(14)$$

Macro Average

Macro Average computes the average of Precision, Recall, and F1-Score for each class independently, without considering the class imbalance.

$$\text{Macro Avg} = \frac{1}{N}\sum_{i=1}^{N} \text{Metric}_i \quad \ldots(15)$$

where N is the number of classes.

Weighted Average

Weighted Average calculates the average of Precision, Recall, and F1-Score while considering the support (the number of true instances for each label). It accounts for class imbalance by giving more importance to classes with higher support.

$$\text{Weighted Avg} = \frac{\sum_{i=1}^{N} \text{Support}_i \times \text{Metric}_i}{\sum_{i=1}^{N} \text{Support}_i} \quad \ldots(16)$$

Confusion Matrix: The Confusion Matrix is a tabular representation of the actual versus predicted classifications. It provides insights into the RFM model's performance by showing the count of true positive (TP), false positive (FP), true negative (TN), and false negative (FN) predictions, which are then used to calculate metrics like Precision, Recall, F1-Score, and Accuracy (Kök, Şimşek, and Özdemir, 2017). These metrics are commonly used in classification tasks to assess how well the model distinguishes between different air quality states.

A Confusion Matrix for a binary classification problem is structured as:

	Predicted Positive	Predicted Negative
Actual Positive	TP	FN
Actual Negative	FP	TN

4. RESULT AND DISCUSSION

The Simulation Result analysis of the air quality data across various cities reveals significant insights into the distribution and concentration of key pollutants across different cities in India. This result includes Air pollution data on several air quality indicators such as $PM_{2.5}$, PM_{10}, NO, NO_2, NOx, NH_3, CO, SO_2, O_3, Benzene, Toluene, Xylene, and the specified Air Quality Index (AQI). The data set is used to concoct and stretch a deep-learning model for forecasting AQI based on pollution statuses in India. Utilizing Google Colab and Python is an applicable method for examining and offering regression and classification experimental data. Regression training and testing sets with 20% of the data reserved for the arbitrary form materialized to 42 testings, batch size of 32, and Training the classification model with a learning rate of 0.1, and 5000 iterations.

Table 4: RFM model predicts air pollution in India.

Recency	Frequency	Monetary
India	190	Unhealthy

As shown in Table 4, provides information on recent high pollution levels in India, the frequency of these events, and the severity of the pollution, suggesting an unhealthy air quality level in Monetary.

The neural network model has been trained to predict air pollution levels over 150 epochs, showing rapid initial improvement. As training continued, loss values decreased, indicating convergence. The final epoch recorded a training loss of 8,301.5947 and a validation loss of 8,923.1963, demonstrating the model's ability to learn and stabilize.

Table 5: RFM model hyperparameters.

Hyperparameters	Significances
Input Sequence	12
Hidden Layer	1
Output Layer	1
Number of Epochs	150

The RFM model has been trained and validated using a neural network over 150 epochs, resulting in a mean squared error (MSE) of 557.68. This relatively low MSE indicates that the model can accurately predict air pollution levels with a reasonable degree of precision. A lower MSE indicates a better fit of the model to the data.

The neural network model, trained over 150 epochs, demonstrated rapid initial improvement in predicting air pollution levels, over the first 150 epochs of training,

Table 6: Training and Validation Metrics Per Epoch (First 50 Epochs)

Training Accuracy	Training Loss	Validation Accuracy	Validation Loss
0.1333	5.7914	0.1533	4.6140
0.1543	4.5265	0.1538	4.4378
0.1560	4.3593	0.1531	4.3554
0.1513	4.2842	0.1570	4.3306
0.1571	4.2104	0.1574	4.3196
0.1549	4.1714	0.1586	4.3149
0.1625	4.0927	0.1608	4.3170
0.1587	4.0802	0.1599	4.3051
0.1579	4.0707	0.1579	4.3413
0.1615	4.0069	0.1596	4.3587
0.1619	4.0050	0.1595	4.3467
0.1623	4.0040	0.1604	4.3194
0.1629	4.0032	0.1610	4.3120
0.1630	4.0030	0.1608	4.3195
0.1636	4.0028	0.1609	4.3186
0.1640	4.0026	0.1613	4.3125
0.1635	4.0025	0.1610	4.3069
0.1637	4.0024	0.1614	4.3095
0.1638	4.0023	0.1613	4.3090
0.1639	4.0023	0.1616	4.3101
0.1642	4.0022	0.1614	4.3078
0.1641	4.0022	0.1615	4.3103
0.1643	4.0021	0.1615	4.3112
0.1644	4.0021	0.1617	4.3109
0.1645	4.0020	0.1617	4.3098
0.1646	4.0020	0.1616	4.3087
0.1647	4.0020	0.1618	4.3082
0.1648	4.0020	0.1618	4.3089
0.1648	4.0020	0.1618	4.3088
0.1649	4.0019	0.1619	4.3091
0.1649	4.0019	0.1619	4.3089
0.1650	4.0019	0.1620	4.3086

0.1650	4.0019	0.1621	4.3084
0.1651	4.0019	0.1621	4.3087
0.1651	4.0018	0.1622	4.3086
0.1652	4.0018	0.1622	4.3089
0.1652	4.0018	0.1622	4.3087
0.1653	4.0018	0.1623	4.3085
0.1653	4.0018	0.1623	4.3088
0.1653	4.0018	0.1624	4.3086
0.1654	4.0018	0.1624	4.3085
0.1654	4.0018	0.1624	4.3088
0.1654	4.0018	0.1624	4.3087
0.1655	4.0018	0.1625	4.3086
0.1655	4.0018	0.1625	4.3086
0.1655	4.0018	0.1625	4.3086
0.1656	4.0018	0.1625	4.3087
0.1656	4.0018	0.1626	4.3087
0.1656	4.0018	0.1626	4.3086
0.1656	4.0018	0.1626	4.3086

A Deep Learning model's training and validation metrics have been collected. Metrics include training accuracy and loss, as well as validation accuracy and loss. The training accuracy begins low and increases over epochs, peaking about epoch 77. Validation accuracy varies while lack of success improves significantly, hovering around 0.16 to 0.17. Training loss typically reduces, whereas validation loss swings and occasionally increases, indicating model architectural flaws. A final epoch produces a training loss of 8,301.5947 and a validation loss of 8,923.1963.

As shown in Table 7, the classification system exhibits varied performance across different classes. Class A has modest precision and recall, resulting in an F1 score of 0.64. Class B outperforms with excellent precision (0.75) and recall (0.70), yielding an F1-Score of 0.72. Alternatively, Class C has lower performance, reflected in an F1-Score of 0.51 due to its lower recall (0.43). Class X struggles with poor performance, as evidenced by an F1-Score of 0.40, indicating substantial challenges. Class Y demonstrates moderate performance with an F1-Score of 0.57. In contrast, Class Z delivers the best results, with high precision (0.68) and recall (0.88), culminating in an F1-Score of 0.77. The overall accuracy of 66% reflects the proportion of correctly classified instances, while the macro average provides an unweighted mean across all classes, highlighting balanced performance. The weighted average, which accounts for the support of each class, offers a balanced view of the system's performance.

Table 7: CLASSIFICATION AIR POLLUTION RFM MODEL IN INDIA.

Class	Precision	Recall	F1-Score	Support
A	0.67	0.62	0.64	836
B	0.75	0.70	0.72	2614
C	0.64	0.43	0.51	1507
X	0.45	0.37	0.40	694
Y	0.54	0.60	0.57	806
Z	0.68	0.88	0.77	2403
Accuracy		0.66		
Macro Avg	0.62	0.60	0.60	8860
Weighted Avg	0.66	0.66	0.65	8860

As shown in Table 6 and Table 7, the model has excellent overall accuracy and a decent fit, as evidenced by the R^2 value. However, the high error metrics indicate potential difficulties with prediction accuracy. The classification metrics reveal varied performance across different classes, though some classes show poorer results.

Table 8: Evaluation for the RFM.

Class	TP	FN	FP	TN
A	522	314	262	6793
B	1839	775	622	3655
C	641	866	1000	4384
X	256	438	572	5625
Y	484	322	419	5666
Z	2121	282	1019	4469

Despite the model's overall high accuracy (Table 8), performance issues have been observed in specific classes, particularly in distinguishing between moderate and severe pollution levels. A substantial number of false positives and false negatives indicate that the categorization process needs further improvement. The model demonstrated better accuracy in predicting air pollution for regions with high variability in pollution levels. Classes C and Z, show a large number of incorrect classifications, indicating the model's difficulty in distinguishing these classes. As shown in Table 9, a deep learning algorithm predicts the air quality index (AQI) with an accuracy of 94%. However, a large Mean Squared Error (454.09) indicates considerable errors. Low metrics such as MAE (14) and RMSE (21.31) indicate that the model's predictions are generally close to true AQI values.

Table 9: EVALUATION METRICS FOR THE RFM.

Metric	Value
Test Accuracy	94%
Mean Squared Error (MSE)	454.09
Root Mean Squared Error (RMSE)	21.31
Mean Absolute Error (MAE)	14.64
Relative Absolute Error (RAE)	0.215
Relative Squared Error (RSE)	0.055
R Squared Error (R^2)	0.945

Table 10: RFM PERFORMANCE METRICS IN BACK PROPAGATION.

Metric	Value
Test Accuracy	99%
Mean Squared Error (MSE)	0.48
Root Mean Squared Error (RMSE)	0.06
Mean Absolute Error (MAE)	0.05
Relative Absolute Error (RAE)	0.11
Relative Squared Error (RSE)	0.03
R Squared Error (R^2)	0.99

As shown in Table 10, using backpropagation, various performance metrics provide insights into the model's accuracy and efficacy. Mean Squared Error (MSE) measures the average squared difference between estimated and actual values, indicating the model's accuracy. A low MSE suggests a relatively low degree of error, reflecting good model performance. In conjunction, the Root Mean Squared Error (RMSE) assesses the average magnitude of prediction errors, and a low RMSE signifies that the error are small. In addition, the Mean Absolute Error (MAE) quantifies the average absolute difference from actual values; an MAE of 0.05 indicates model accuracy. The Relative Absolute Error (RAE) stands at 11% of the baseline error, offering that the model performs significantly better than the baseline. Furthermore, the Relative Squared Error (RSE) is 0.03, indicating a 97% reduction in error compared to the baseline. Finally, the R^2 value, representing the proportion of variance explained by the model, is 0.9, signifying an excellent fit. The RFM model using backpropagation regression achieves 99% test highest accuracy for air quality predictions, with low MSE and RMSE, excellent accuracy, and strong performance compared to baseline models. It explains 99% of the dependent variable's variance with an R^2 of 0.99, demonstrating robustness.

Table 11: EVALUATION BACK PROPAGATION FOR RFM.

Class	TP	FN	FP	TN
A	0	1209	0	2688
B	1601	0	1209	2087
C	1289	0	0	2608
X	510	0	0	3387
Y	609	0	0	4288
Z	689	0	0	4208

As shown in Table 11, the model's failure to identify Class A and high false positives for Class B raises concerns about its accuracy. The confusion matrix reveals that while the model performs well for Classes C, X, Y, and Z, it significantly underperforms in Classes A and B. The model's failure to correctly classify Class A and high false positives for Class B

As shown in Table 12, the RFM model's precision, recall, and F1-score for classes A, C, X, Y, and Z are 0.00, indicating it failed to predict any instances. Class B has a precision of 0.27, a recall of 1.00, and an F1-score of 0.43, suggesting it properly recognized all cases but assembled a large number of false positives. The overall accuracy is 0.27, indicating underperformance, especially considering imbalanced performance across different classes. The Macro average values are low, indicating poor performance across all classes without considering class distribution. They perform well for class B but struggle to generalize across other classes,

Table 12: CLASSIFICATION REPORT FOR BACK PROPAGATION.

Class	Precision	Recall	F1-Score	Support
A	0.00	0.00	0.00	1209
B	0.27	1.00	0.43	1601
C	0.00	0.00	0.00	1289
X	0.00	0.00	0.00	510
Y	0.00	0.00	0.00	609
Z	0.00	0.00	0.00	689
Accuracy	0.27			
Macro Avg	0.05	0.17	0.07	5907
Weighted Avg	0.07	0.27	0.12	5907

Table 13: RFM PERFORMANCE METRICS IN FORWARD PROPAGATION.

Metric	Value
Test Accuracy	93%
Mean Squared Error (MSE)	39085.58
Root Mean Squared Error (RMSE)	197.70
Mean Absolute Error (MAE)	137.63
Relative Absolute Error (RAE)	1.44
Relative Squared Error (RSE)	1.93
R Squared Error (R^2)	0.93

As shown in Table 13, the model has a high test accuracy of 99%, indicating its exceptional performance in classifying or predicting the target variable correctly. However, the mean squared error (MSE) is significantly higher than expected, suggesting a notable difference between predicted and actual values. Moreover, the root mean squared error (RMSE) is considerable, indicating significant inaccuracies in predictions. In addition, the mean absolute error (MAE) is relatively high, reflecting a considerable average error in predictions. Furthermore, the model's absolute error is 144% of the baseline error, implying worse performance compared to a baseline model. Similarly, the relative squared error (RSE) is 193% of the total squared error, indicating large errors relative to the baseline. Despite these issues, the R^2 value is high, suggesting that 93% of the variance in the target variable is explained by the model, which indicates a very good fit.

Table 14: EVALUATION FOR WARD PROPAGATION FOR RFM.

Class	TP	FN	FP	TN
A	1209	0	4698	6793
B	0	2107	1209	3575

C	0	1289	1209	3874
X	0	510	1209	3384
Y	0	609	1209	3384
Z	0	689	1209	6682

As shown in Table 14, the confusion matrix for ward Propagation reveals that the model is underperforming significantly, with a strong bias towards Class A and failure to classify other classes. The increased amount of false positives for Class A and zero true positives for the additional classes. Table 15, The model's performance in class A has been analyzed, with a recall of 1.00 and a precision of 0.20, indicating a correct labeling of class A instances. However, the model also misclassified many instances of classes B, C, X, Y, and Z, resulting in zero precision, recall, and F1-Score. The overall accuracy has been 20%, indicating only 20% of the total predictions are correct.

Table 15: CLASSIFICATION REPORT FOR WARD PROPAGATION.

Class	Precision	Recall	F1-Score	Support
A	0.20	1.00	0.34	1209
B	0.00	0.00	0.00	1601
C	0.00	0.00	0.00	1289
X	0.00	0.00	0.00	510
Y	0.00	0.00	0.00	609
Z	0.00	0.00	0.00	689
Accuracy		0.20		
Macro Avg	0.03	0.17	0.06	5907
Weighted Avg	0.04	0.20	0.07	5907

The model's performance across all classes has been poor, with a significant failure to predict instances of classes B, C, X, Y, and Z. Classes with relatively fewer samples have larger weights than the majority classes to balance their contribution to the loss function. The neural network model demonstrated a strong ability to capture complex patterns in the air pollution data. The RFM model integration appeared to increase forecast accuracy, particularly in metropolitan regions with varying pollution levels. The RFM model's integration with neural networks enhanced air pollution forecast accuracy, as indicated by higher metrics including MSE and R-squared compared with standard models without RFM integration. Combining RFM models A B C X Y Z with deep learning can yield better predictive performance. The model's ability to effectively account for recency, frequency, and monetary value in pollution data has been evident, particularly when predicting air quality index (AQI) values.

5. CONCLUSIONS

The integration of the RFM model with neural networks has proven effective in enhancing air quality predictions across various Indian cities. By incorporating recency, frequency, and severity, the model provides a detailed and accurate representation of air pollution trends. The neural network demonstrated significant reductions in loss and effective convergence, offering valuable insights into temporal patterns and high-risk pollution areas. Regression performs better in Air pollution levels in predicting continuous outcomes and effectively apprehending the nuanced variations in air quality metrics. However, the strength of classification in the RFM model keeps the focus on their practical application in air quality analysis. Performance issues have been observed in distinguishing between moderate and severe pollution levels, highlighting the need for further refinement of the classification model. Regularization techniques and more complex neural network architectures could improve robustness and generalization. Evaluating the RFM with neural networks complex model on prior observed assessment air pollution data and comparing training and validation errors is critical for detecting overfitting from occurring.Additionally, future research should focus on optimizing the model's performance through better tuning and exploring advanced architectures. Regression techniques show promise for determining precise pollutant levels, while classification methods are more suitable for assessing overall pollution risk. These improvements will help develop targeted interventions to mitigate urban air pollution effectively.

REFERENCES

1. Anitha, P. and Patil, M.M., 2022. RFM model for customer purchase behavior using K-Means algorithm. *Journal of King Saud University - Computer and Information Sciences*, 34(5), pp.1785-1792. DOI: 10.1016/j.jksuci.2019.12.011.
2. Mohammad, J. and Kashem, M.A., 2022. Air pollution comparison RFM model using machine learning approach. In: *Proceedings of the 2022 IEEE 7th International Conference on Convergence Technology (I2CT)*. Pune, India, 2022. pp.1-5. DOI: 10.1109/I2CT54291.2022.9824248.
3. Mena, G., Coussement, K., De Bock, K.W., De Caigny, A. and Lessmann, S., 2023. Exploiting time-varying RFM measures for customer churn prediction with deep neural networks. *Annals of Operations Research*, pp.1-23. DOI: 10.1007/s10479-023-05259-9.
4. Liao, J., Jantan, A., Ruan, Y. and Zhou, C., 2022. Multi-behavior RFM model based on improved SOM neural network algorithm for customer segmentation. *IEEE Access*, 10, pp.122501-122512. DOI: 10.1109/ACCESS.2022.3223361.
5. Kurt, A., Gulbagci, B., Karaca, F. and Alagha, O., 2008. An online air pollution forecasting system using neural networks. *Environmental International*, 34(5), pp.592-598. DOI: 10.1016/j.envint.2007.12.020.
6. Ibarra-Berastegi, G., Elias, A., Barona, A., Saenz, J., Ezcurra, A. and de Argandona, J.D., 2008. From diagnosis to prognosis for forecasting air pollution using neural networks: Air pollution monitoring in Bilbao. *Environmental Modelling & Software*, 23(5), pp.622-637. DOI: 10.1016/j.envsoft.2007.09.003.
7. Wesolowski, M., Suchacz, B. and Halkiewicz, J., 2006. The analysis of seasonal air pollution patterns with the application of neural networks. *Analytical and Bioanalytical Chemistry*, 384, pp.458-467. DOI: 10.1007/s00216-005-0197-0.
8. Ibarra-Berastegi, G., Saenz, J., Ezcurra, A., Elias, A. and Barona, A., 2009. Using neural networks for short-term prediction of air pollution levels. In: *Proceedings of the 2009 International Conference on Advanced Computing Tools for Engineering Applications (ACTEA)*. Zouk Mosbeh, Lebanon, 2009. pp.498-502. DOI: 10.1109/ACTEA.2009.5227910.
9. Kalajdjieski, J. et al., 2020. Air pollution prediction with multi-modal data and deep neural networks. *Remote Sensing*, 12(24), p.4142. DOI: 10.3390/rs12244142.
10. Brauer, M. et al., 2019. Examination of monitoring approaches for ambient air pollution: A case study for India. *Atmospheric Environment*, 216, p.116940. DOI: 10.1016/j.atmosenv.2019.116940.
11. Gupta, U., 2008. Valuation of urban air pollution: A case study of Kanpur City in India. *Environmental Resources and Economics*, 41, pp.315-326. DOI: 10.1007/s10640-008-9193-0.
12. Bishop, C.M., 2006. *Pattern Recognition and Machine Learning*. Springer. DOI: 10.1007/978-0-387-45528-0.
13. Deng, L., 2014. A tutorial survey of architectures, algorithms, and applications for deep learning. *APSIPA Transactions on Signal and Information Processing*, 3. DOI: 10.1017/ATSIP.2013.9.
14. Goodfellow, I., Bengio, Y. and Courville, A., 2016. *Deep Learning*. MIT Press. DOI: 10.7551/mitpress/10993.001.0001.
15. LeCun, Y., Bengio, Y. and Hinton, G., 2015. Deep learning. *Nature*, 521(7553), pp.436-444. DOI: 10.1038/nature14539.
16. Rumelhart, D.E., Hinton, G.E. and Williams, R.J., 1986. Learning representations by back-propagating errors. *Nature*, 323(6088), pp.533-536. DOI: 10.1038/323533a0.
17. U.S. Environmental Protection Agency (EPA), 2014. Air Quality Index: A Guide to Air Quality and Your Health. Available at: https://www.epa.gov/airquality-index [Accessed 7 Oct. 2024].
18. U.S. Environmental Protection Agency (EPA), 2021. Nitrogen Dioxide (NO2) Pollution. Available at: https://www.epa.gov/no2-pollution [Accessed 7 Oct. 2024].
19. World Health Organization (WHO), 2021. Ambient air pollution: A global assessment of exposure and burden of disease. Available at: https://www.who.int/publications/i/item/9789240061311 [Accessed 7 Oct. 2024].
20. Central Pollution Control Board (CPCB), 2023. National Ambient Air Quality Status & Trends. Available at: https://cpcb.nic.in/air-quality-status/ [Accessed 7 Oct. 2024].
21. Ministry of Environment, Forest and Climate Change (MoEFCC), 2022. Annual Report 2021-22. Available at: https://moef.gov.in/annual-report/ [Accessed 7 Oct. 2024].
22. European Environment Agency (EEA), 2022. Air quality in Europe – 2022 report. Available at: https://www.eea.europa.eu/publications/air-quality-in-europe-2022 [Accessed 7 Oct. 2024].
23. Agency for Toxic Substances and Disease Registry (ATSDR), 2021. Toxicological Profile for Benzene. Available at: https://www.atsdr.cdc.gov/toxprofiles/tp3.html [Accessed 7 Oct. 2024].
24. Zhang, D. et al., 2019. A novel hybrid model based on neural networks and statistical methods for air quality forecasting. *IEEE Access*, 7, pp.109415-109427. DOI: 10.1109/ACCESS.2019.2932995.
25. Huang, C. et al., 2020. Evaluation of neural network model performance using statistical metrics in predicting air pollutant concentrations. *IEEE Transactions on Neural Networks and Learning Systems*, 31(5), pp.1450-1460. DOI: 10.1109/TNNLS.2019.2921940.
26. Yadav, R. and Malik, H., 2020. Neural network-based predictive models for air quality index using statistical performance metrics. *Neural Computing and Applications*, 32(9), pp.4591-4603. DOI: 10.1007/s00521-019-04162-0.
27. Ma, Y. et al., 2020. A comparative study of neural networks and traditional statistical models for air quality forecasting. *Journal of Cleaner Production*, 256, p.120461. DOI: 10.1016/j.jclepro.2020.120461.
28. Zhang, S. et al., 2020. Comparative analysis of machine learning models and traditional statistical methods for air quality prediction. *Atmospheric Pollution Research*, 11(5), pp.927-935. DOI: 10.1016/j.apr.2020.02.002.

29. Kok, M.U., Şimşek, M. and Özdemir, S., 2017. A deep learning model for air quality prediction in smart cities. In: *2017 IEEE International Conference on Big Data (Big Data)*. Dec. 2017, pp.1983-1990. DOI: 10.1109/BigData.2017.8258237.
30. Liu, M., Du, Y. and Xu, X., 2019. Customer value analysis based on Python crawler. In: *2019 Chinese Control And Decision Conference (CCDC)*. Jun. 2019, pp.4345-4350. DOI: 10.1109/CCDC.2019.8832805.
31. Panus, J. et al., 2016. Customer segmentation utilization for differentiated approach. In: *2016 International Conference on Information and Digital Technologies (IDT)*. Jul. 2016, pp.227-233. DOI: 10.1109/DT.2016.7557178.
32. Teslenko, D. et al., 2023. Comparative analysis of the applicability of five clustering algorithms for market segmentation. In: *2023 IEEE Open Conference of Electrical, Electronic and Information Sciences (eStream)*. pp.1-6. DOI: 10.1109/eStream59056.2023.
33. Khanal, S., Pokhrel, R.P., Pokharel, B., Becker, S., Giri, B., Adhikari, L. and LaPlante, M.D., 2022. An episode of transboundary air pollution in the central Himalayas during agricultural residue burning season in North India. *Atmospheric Pollution Research*, 13(1), p.101270. doi:10.1016/j.apr.2021.101270.
34. Kumar, P. & Gupta, A. (2023) 'Comparative analysis of SVM and deep learning methods for predicting air quality index', *Environmental Monitoring and Assessment*, 195(6), pp. 12-24. doi: 10.1007/s10661-023-11670-1.
35. Data available at: https://raw.githubusercontent.com/jannahmohammad/Air/master/cityday(3).csv.