RESEARCH PAPER

A Robust IQR-Based Pre-processing and Tri-Linear Neural Network Framework for Multivariate Air Pollution Analysis

Mrs. P. Umasanthiya[1*], Dr. Marrynal S Eastaff [2],

[1]Assistant Professor, Department of Computer Technology, Hindusthan College of Arts & Science, Coimbatore. Email Id: sandhyaridhu@gmail.com

[2]Associate Professor, Department of Computer Science with Cyber Security, Hindusthan College of Arts & Science, Coimbatore. Email Id: marrynalhindusthan@gmail.com

Abstract

One of the critical environmental and population health issues induced by the high levels of urbanization, industrial processes, and growing numbers of vehicular emissions is air pollution, and appropriate multivariate analysis of air quality is required. This study will be used to make the analysis of air pollution more credible as it will address the significant issues such as the lack of values, outliers, and ambiguous relationships between pollutant, meteorological, and time variables. A powerful pre-processing and feature selection model is proposed in this respect. The methodology utilizes Imputer-Interquartile Range (IQR)-based pre-processing regimen to resolve the outliers and missing data with the means of median based statistics and Tri-Linear Fully Connected feature selection paradigm to determine the most powerful features by learning cross-feature interaction. The proposed approach is shown to be effective as judged by experimental assessment of multiregional data on air quality. The pre-processing model attains a low RMSE level of 6.72 which means that the data stability and prediction are better than the current approaches. Moreover, the selected approach of feature selection optimizes the overall performance of classification with the accuracy of 94.12. In general, the proposed structure can be used to offer an effective and stable solution to precise multivariate air pollution analysis, which may be used to promote proper environmental surveillance and decision-making.

Keywords: Air Quality Index, Feature Selections, IQR Pre-processing, Multivariate Analysis, Neural Networks, Pollution level.

How to cite this article: Umasanthiya P, Eastaff MS. A Robust IQR-Based Pre-processing and Tri-Linear Neural Network Framework for Multivariate Air Pollution Analysis. Int J Drug Deliv Technol. 2026;16(54s): 220-234. DOI: 10.25258/ijddt.16.54s.17

1. Introduction

The rapid urbanization, the development of industries and the growth of the number of vehicles have become the primary factors contributing to the air pollution as one of the most significant environmental and human health issues. Effective environmental monitoring and assessment of quality of air is therefore inseparable with the correct analysis of data on air quality. The enhancement of the air pollution analysis through the application of advanced machine learning and deep learning methods has been the subject of several studies. To illustrate this, novel data modeling approaches like ensemble-based and multivariate modeling of data have been taken into account so as to capture complex pollution patterns [1]. The problem of missing and noisy data has been solved by availing data imputation and multivariate processing processes [2]. The frameworks that select features have been proposed as a way of eliminating redundancy and improving data representation [3]. It is also shown that deep learning based models are capable of learning and learning intricate environmental variable correlations [4]. Moreover, multimodal methods of data integration approach have been utilized to understand spatio-temporal changes in pollution in a better manner [5]. Several publications have also examined wavelet-based learning [6], multivariate-regression-models [7], and hybrid analytical strategies, as well as the wrapper-based-feature-selection techniques were applied to enhance interpretability and analytical-efficiency [9]. Further developments in the ensemble-based methods have also led to sustained air quality analysis [10].

In order to mitigate the shortcomings that have been witnessed in the earlier studies, the research proposes a detailed method, which is pre-processing and feature selection. The proposed procedure is pegged on the methodological data clean up and data normalization and structural enhancement and followed by an influential segmentation feature plan that strives to maintain interdependencies in the face of the pollution, the meteorological and time variables. This kind of systematization will reflect the data as well as an analysis of the information in a superior form that can provide a decent base upon which further analysis of the air quality properties can be done.

Enhancing the quality and dependence of the air pollution analysis through the solution of such critical issues as data noise, redundancy, and interaction between complex features was the main strength of this work. The study will make contributions to the literature in the sense that it provides a very detailed pre-processing model and a competent feature selection policy that would enhance the data representation in a more refined way. The proposed methodology is effective as it gradually fades away the interdependences between the pollution, meteorological

*Author for Correspondence: sandhyaridhu@gmail.com.

and time-related characteristics, which then makes the results analysis more meaningful. In general, one can say that the work provides a strong and scalable system of effective and accurate air quality measurements.

The rest of this study will follow the following structure: Section 1 will be the introduction and provide the motivation and objective of the study. Section 2 provides a summary of related literature and background on the multivariate air pollution analysis. Section 3 outlines the methodology proposed in the study and pre-processing and feature selection methods. In Section 4, the results of the experiment and performance analysis are discussed, and Section 5 provides the conclusion of the research with important findings and perspectives of possible future research.

2. Background study:

Rakholia et al. (2023) [11] The article is concerned with the issue of the multi-output regional air pollution prediction, and the authors developed a machine-learning system to simultaneously predict multiple pollutants according to the ensemble learning and multi-target learning systems. Although it is better in terms of interdependence of the pollutants, the study is limited to one area, and it is not adaptive in addressing the concept drift in rapidly evolving urban systems.

Chen et al. (2022) [12] Using Autoregressive Integrated Moving Average with Exogenous Regressors (ARIMAX) to estimate the impact of the pollutants on the health outcomes, the article describes the relationship between the air pollution and the meteorological conditions and the occurrence of tuberculosis. Even though the results prpopose that there is significant predictive correlation, the model has weaknesses in that it has linear assumptions and does not have the element of multivariate nonlinear interactions among the pollutants during real-time prediction of air quality.

Samal et al. (2021)[13] The issues expressed in this article are the gaps in data during air pollution predictions and the prpoposedsolution to the problem is the introduction of a temporal convolutional denoising autoencoder to predict and restore the magnitude of the pollutants. It is computationally costly and cross-pollutant dependency modelling is explicit as well though, without it is still missing, yet the model performs well when the data are sparse.

Liao et al. (2021) [14] It is a study within the framework of which the authors represent the general overview of the statistical prediction of primary air pollutants regression, the Autoregressive Integrated Moving Average (ARIMA) and the hybrid approaches, which are applied in predicting air quality. The weaknesses in accuracy and insufficient scalability of classical models are also highlighted during the review and refer to the need of multivariate and deep learning-based forecasting models.

Subramaniam et al. (2022) [15] The article is a narrative review of AI techniques of air pollution and health prediction, and it also compares machine learning, deep learning, and hybrid networks in numerous applications. This study has great predictive improvements, but there is a gap area in the explainability, data integration, and multivariate model practical implementation.

Muzayyanah et al. (2023) [16] The authors analyze the non-linear correlation between the socio-economic productivity and air pollution on the basis of Multivariate Adaptive Regression Splines (MARS). Even though the model is a good model of handling the complex nonlinear effects, the model is inappropriate in forecasting the concentration of pollutants and that it is not able to foresee the future.

Alahamade et al. (2021) [17] The article is dedicated to a multivariate time-series method of imputation of air pollution data by the similarity learning method based on clustering in order to recover the missing data on pollutants. Although the accuracy of imputation can be significantly increased, the technique does not have the predictive qualities of forecasting and rich time modelling.

Hadeed et al. (2020) [18] It is an evaluation study that entails a comparison of the conventional and multivariate approaches to statistical and machine-learning-based imputation of air pollution to the short terms. The outcomes are encouraging towards better data coverage, and the study fails to extrapolate the findings of imputation research to a later multivariate prediction model.

Dar and Shaik (2025) [19] the authors propose AQI spatiotemporal clustering and multivariate forecasting framework in Indian cities, which is a machine learning and deep learning-based approach. Despite the fact that the findings prpopose that it is well-spatially generalized, the weaknesses are that deep models are highly data-dependent and are less interpretative.

Chang et al. (2020) [20] prpopose a generalized LSTM-based air pollution predictive system, which constitutes time-dependent Ness among two or more stations. The technique is more predictive than those using single-stations but fails to solve the problem of missing data or shifting multivariate pollutant association.

Table 1: Multivariate Air Pollution Prediction Models Background Study.

Ref	Author (Year)	Concept	Author Contribution	Methods Used	Research Gap & Limitations	Results
[21]	Zhao et al. (2022)	Detection of air pollution episodes	Analyzed pollution episodes in chemical industry parks using multivariate	PCA, Hotelling's T^2, multivariate control charts	Not designed for forecasting; lacks deep temporal modeling	Successfully detected abnormal pollution events

			statistics			
[22]	Khan et al. (2022)	Air pollution–driven traffic prediction	Used air pollution data to forecast traffic flow in smart cities	Bagging ensemble ML models	Focuses on traffic, not pollutant forecasting; indirect air quality modeling	Improved traffic prediction accuracy
[23]	Tsokov et al. (2022)	Spatiotemporal air pollution prediction	Developed a hybrid deep model for pollutant forecasting	CNN + LSTM	Limited interpretability and handling of missing data	Achieved high spatiotemporal prediction accuracy
[24]	Li et al. (2020)	PM2.5 concentration prediction	Modeled urban PM2.5 using attention-based deep learning	Attention-based CNN–LSTM	Focused on single pollutant; limited multivariate interaction	Improved PM2.5 prediction performance
[25]	Ku et al. (2022)	Health impact prediction	Predicted respiratory disease occurrence using air pollution and climate data	ML classifiers (RF, SVM, etc.)	Health-focused; not a direct air pollution forecasting model	Identified pollution–health correlations effectively
[26]	Dai et al. (2021)	Spatiotemporal pollutant forecasting	Captured spatial and temporal features for pollutant prediction	Multi-scale 1D CNN–LSTM	High computational cost; city-specific model	Outperformed traditional DL models
[27]	Ragab et al. (2020)	AQI prediction	Proposed a CNN-based AQI prediction framework	1D CNN with adaptive gradients	AQI-level only; lacks pollutant-wise multivariate prediction	Achieved stable and accurate AQI forecasts
[28]	Fouladgar&Främling (2020)	Missing data handling in multivariate series	Addressed massive missingness in air pollution time series	Modified LSTM architecture	Does not integrate spatial features	Robust prediction under high missing data
[29]	Maltare& Vahora (2023)	City-level AQI prediction	Applied ML models for AQI forecasting in Ahmedabad	SVR, RF, ANN	City-specific; limited scalability	Improved AQI prediction accuracy
[30]	Jin et al. (2020)	Long-term air quality prediction	Proposed frequency-aware deep hybrid model	EMD + frequency classification + DL	Long-term focus; lacks real-time adaptability	Accurate long-term air quality trends

Table 1 provides a brief overview of the available literature on the subject of multivariate air pollution prediction identifying the methodologies of each study, their contribution, and their limitations. It helps to identify the gaps in research which stimulates the creation of advanced spatiotemporal and multi-output models of forecasting.

3. Proposed Methodology

The section on the Proposed methodology provides an excellent overview of the research that may be conducted, focusing on air pollution analysis starting with the account of the dataset and the overall architecture of the system. Both the processing phases are provided in detailed form with preprocessing, feature selection and modeling along with the respective architectural design. In addition to that, mathematical formulations, algorithms, and pseudocodes used at every step are clear in the section to render the proposed approach transparent, reproducible, and applicable.

Dataset:
https://www.kaggle.com/datasets/syedmtalhahasan/global-urban-air-quality-index-dataset-2015-2025/?utm_source=chatgpt.com
Global Urban Air Quality Index Dataset (20152025) in Kaggle presents the data of long-term air quality of certain urban monitoring stations allover the world, as well as the key pollutants (PM 2.5, PM 10, NO 2, SO 2, CO, and O 3), which are the crucial factors when defining the health and environmental conditions of the populations. Kaggle, it is interested in the data of 2015-2025, which will allow studying the dynamics of the pollutants in ten years. Kaggle. The statistical data of the Air Quality Index, based on daily or aggregated data and the concentration of each pollutant in the majority of the cities can be accessed and used in multivariate time-series analysis. Kaggle. Its area of geographical implementation also makes it applicable in complicated machine learning activities such as classification and prediction in different area. Kaggle. This general information can be favorably used to pre-process the recurrent neural network models with attention and other Bi-GRU-based air pollution prediction.

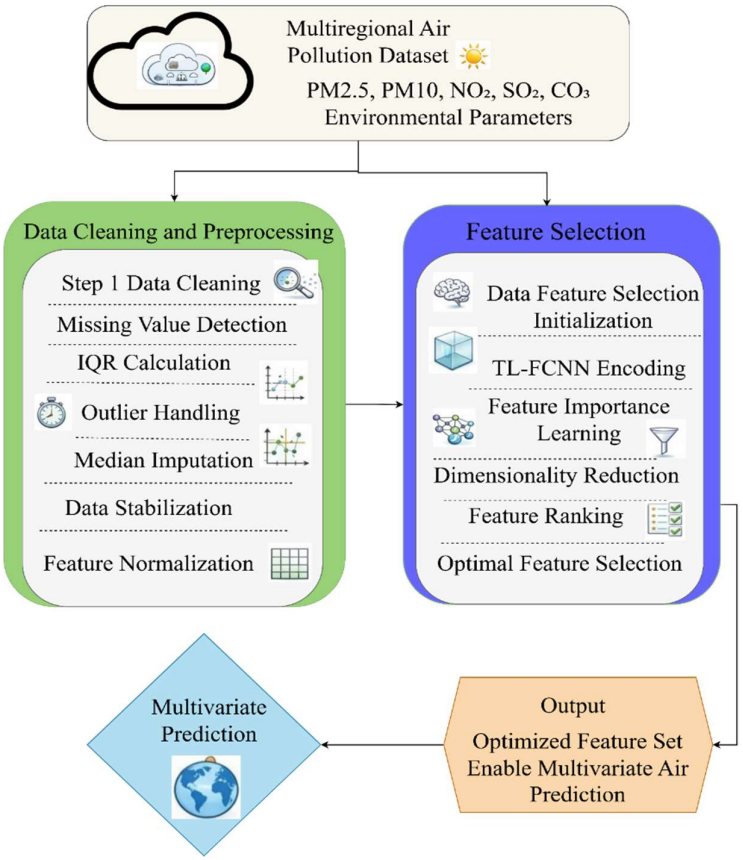

Figure 1: Air Pollution Pre-Processing and Feature Selection Framework

Figure 1 shows the entire work process of pre-processing and features selection of the multiregional air pollution data. It contains data cleaning phases of missing values, outlier, normalization and stabilization, and then advanced feature selection based on encoding, dimensionality reduction, and feature ranking. The end product is an optimized set of features that increases the accuracy and reliability of the future air quality prediction models.

3.1 Pre-Processing using Imputer Interquartile Range Algorithm

The Imputer-IQR-based pre-processing algorithm is oriented on enhancing the reliability of data with the aid of concurrent outliers correction and missing values of multiregional air pollution data before model training. It operates by first identifying abnormal values using the interquartile range (1.5xIQR rule) and minimizing the effect by either eliminating or adjusting them and then performing the imputation using the statistics based on the median. Conventional pre-processing methods are typically vulnerable to extreme values, biased in making use of the mean to fill-in absent information, and want the adaptation to heterogeneous and non-uniform environmental data. The proposedsolution can overcome these inadequacies through the utilization of the IQR-based robust statistics that are less skewness and extremities-sensitive to ensure that there are stable data distributions. It preserves the natural trend of pollution variables by the combination of the outlier management with imputation in the same framework. This improves

extrapolation of models and curbs the proliferation of errors during the learning process. In addition, the algorithm is scalable and computationally efficient and can be used in real-time air quality monitoring system. It, in turn, enables the superior and more foreseeable functioning of prediction on the real-life deployment environment.

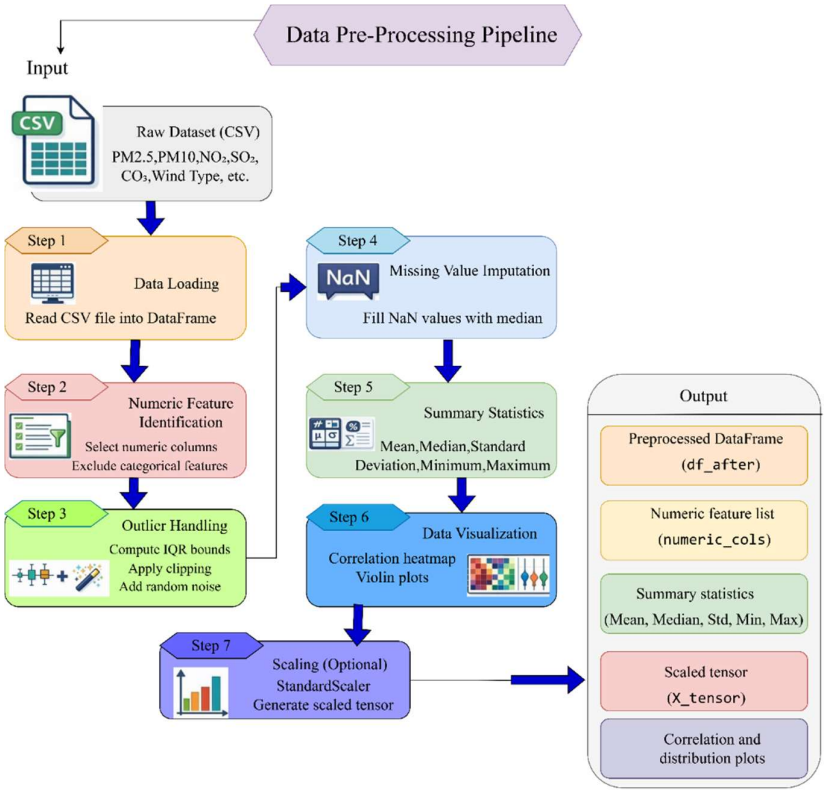

Figure 2:Pre-processing Pipeline using the Proposed IQR–Median Imputation Algorithm

Figure 2, it, it shows the entire procedure of pre-processing, which occurred on multivariate air pollution data by use of the intended algorithm, IQR-Median Imputation. It starts with loading and doing a feature identification of data in the form of numbers to identify outliers, median-based imputation of missing data to even distributions of data, and IQR-based outlier treatment to smooth the data distributions. Lastly, the summary statistics, visualization, and optional scaling are done to possess a clean and structured dataset, which can be analyzed using downstream analysis with confidence.

The unprocessed data on multivariate air pollution is a numerical matrix.

$$X = \{x_{i,j} | i = 1, \dots, N;\ j = 1, \dots, N\}$$

(1)

Equation 1 $x_{i,j}$ The value of the j^{th} feature in the i^{th} sample, N The total number of observations, M The total number of features (pollutants, meteorological variables). This equation constitutes the input data on which IQR-based detection and imputation are done.

The quartile represents the percentile of each feature.

$$Q_1^{(j)} = \text{Percentile}_{25} X_{:,j} \quad (2)$$

Equation 2, $Q_1^{(j)}$ percentile of the feature j, $X_{:,j}$ Value of feature j. This is the minimum limit of distribution of all pollutants or variables.

The third quartile takes the 75 th percentile of all features.

$$Q_3^{(j)} = Percentile_{75} X_{:,j} \quad (3)$$

Equation 3 $Q_3^{(j)}$ Third quartile of the feature j. This is the highest degree of the spread on which it is specified that it is an outlier.

The difference between the 3rd and 1st quartile is measured using the IQR.

$$IQR^{(j)} = Q_3^{(j)} - Q_1^{(j)}$$

(4)

Equation 4 $IQR^{(j)}$ Interquartile range of feature j. IQR quantifies the fluctuation of every of the attributes and it is also not susceptible to extreme figures.

Abnormally small values are determined by using the low threshold.

$$LB^{(j)} = Q_1^{(j)} - 1.5 \times IQR^{(j)}$$

(5)

Equation 5 $LB^{(j)}$ Lowest bound of feature j. Any value less than this will be considered as outliers that will contaminate learning.

The end determines excessive high values.

$$UB^{(j)} = Q_3^{(j)} - 1.5 \times IQR^{(j)}$$

(6)

In Equation (6) $UB^{(j)}$ Upper bound of feature j. Any other values that seem to exceed this value are regarded as an anomaly.

The points of data are evaluated based on IQR limits.

$$x_{i,j} = \begin{cases} \text{outlier}, & x_{i,j} < LB^{(j)} \text{ or } x_{i,j} > UB^{(j)} \\ \text{valid}, & \text{otherwise} \end{cases}$$

(7)

In Equation 7, $x_{i,j}$Value of the individual data. This regulation lifts the banner of radical pollutant values, which can negatively affect the preparation of models. Strong statistics will replace the outliers and missing values.

$$\tilde{x}_{i,j} = \begin{cases} \text{Median}(X_{:,j}), & x_{i,j} \epsilon \{\text{missing, outlier}\} \\ x_{i,j}, & \text{otherwise} \end{cases}$$

(8)

Total equation 8$\tilde{x}_{i,j}$ Imputed value, Median$(X_{:,j})$Median of the feature j. Median imputation does not lose data spread and is not influenced by extremities.

Imputed clean data is now in a position to be trained on a model.

$$X^* = \{\tilde{x}_{i,j} | \forall i, j\}$$

(9)

In Equation 9, X^* Final processed data (cleaned, ready to learn), $\tilde{x}_{i,j}$Pre-processed value, ith sample, j-thfeature,iIndex of data samples (rows), j index of features/attributes (columns), $\forall i, j$ All the samples, all features. The data is in full, clean and normalized values with no missing information and deviations to be used in the stable model training.

Algorithm:Pre-process using IQR

Input
- file_path : CSV file containing the dataset
- numeric_cols : Optional list of numeric columns to preprocess (default: all numeric)
- random_seed : Seed for reproducibility (default = 42)

Steps:
1. Set random seed to ensure reproducibility
2. Load dataset df from file_path
3. Identify numeric columns in df:
numeric_cols = all numeric columns in df
Exclude non-relevant columns (e.g., 'Wind Type') if present
4. Create a copy of df → df_after
5. For each column col in numeric_cols:
a. Compute Q1 = 25th percentile of col
b. Compute Q3 = 75th percentile of col
c. Compute IQR = Q3 - Q1
d. Set lower = Q1 - 1.5 * IQR
Set upper = Q3 + 1.5 * IQR
e. Clip col values between lower and upper
f. Apply small random perturbation to prevent duplicate values
(e.g., multiply by random factor between 0.995 and 1.005)
6. For each column col in numeric_cols:
a. Compute median_val = median of col
b. Introduce missing values (simulated or actual)
c. Replace missing values with median_val
7. Return df_after, numeric_cols

Output:
- df_after :Preprocessed dataset with outliers handled and missing values imputed
- numeric_cols : List of numeric columns processed

The abovementioned pseudocode is a pre-processing pipeline, which first of all shows and labels the numeric columns within a dataset and finds outliers in terms of the interquartile range (IQR) of values and then gets rid of the outliers by clipping them to reduce their impact. It subsequently fills in on missing values by imputing the median (with or without filling in any outliers that have also been detected as missing) such that the dataset is coherent and hardy. The resulting provides a clean and normalized dataset in order to be used in grounded machine learning or deep learning activity.

3.2 Feature Selection using Tri-Linear Fully Connected Neural Network

The proposed Tri-Linear Fully Connected Neural Network (TL-FCNN) is set to be an effective feature selection model of multivariate air pollution data because it can represent complex interdependence between pollutant, meteorological, and temporal features. The conventional feature selection techniques tend to fail in capturing the nonlinear relationships, and they also have redundancy in the case of strongly correlated pollution factors. To eliminate this, the proposed plan breaks down the input characteristics into three significant groups and processes that undergo autonomous linear metamorphoses. In these transformed representations, learning deep cross-feature dependencies is then done by combining them with a tri-linear interaction mechanism. This allows the

model to determine the most important pollution-related characteristics correctly and remove redundant and scary information. The method is very helpful in dealing with multicollinearity and changing environmental variations found in actual pollution records. Consequently, the features chosen contribute to a high degree of prediction and robustness of the model. Therefore, the offered approach is very appropriate for real-time and massive air pollution monitoring and prediction systems.

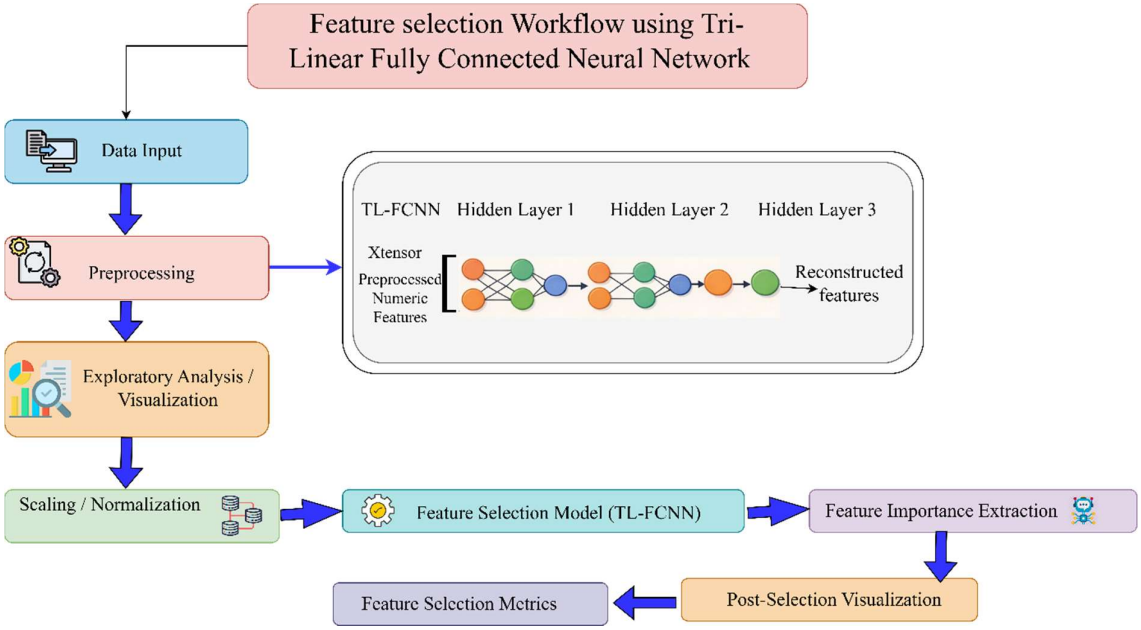

Figure 3: Feature Selection Workflow using Proposed TL-FCNN Algorithm

Figure 3, the figure shows the end-to-end feature selection process according to the Tri-Linear Fully Connected Neural Network (TL-FCNN) algorithm, which has been proposed. The pre-processing of raw data is followed by analysing the data with the exploratory analysis and scaling/normalizing the data, after which it is fed into the TL-FCNN that recreates the features with the help of the three hidden layers. Finally, the model outputs include feature importance that is visualized and evaluated based on post-selection metrics as a means of selecting the most significant features that will be utilized in further analysis.

The air pollution data at time t is a high-dimensional feature, which is a multivariate feature.

$$X_t = [x_{t,1}, x_{t,2}, \dots, x_{t,d}] \epsilon \mathbb{R}^d$$

(10)

Equation 10 X_t Feature at a time t, d Number of original features (pollutants, meteorological, and temporal features) $x_{t,i}$ Value of the i^{th} feature. The equation performs a representation of the raw multivariate feature space before feature selection.

The input feature set is further broken down into three feature sets that are complementary to one another.

$$X_t = X_t^{(1)}, X_t^{(2)}, X_t^{(3)}$$

(11)

Equation 11 $X_t^{(1)}: k^{th}$ set of features (e.g., pollutants, meteorology, time context). Tri-linear learning can be used to model interaction in highly heterogeneous collections of features.

A fully connected layer is used to transform the resultant subset of features linearly.

$$H_t^{(1)} = W_1 X_t^{(1)} + b_1$$

(12)

Equation 12 W_1 First weight projection matrix, b_1 Biases, $H_t^{(1)}$ Hidden representation. This forecast obtains the trends of significance of pollutants.

The latter features are input into the latent space.

$$H_t^{(2)} = W_2 X_t^{(2)} + b_2$$

(13)

Equation 13 W_2, b_2 Meteorological, weights and bias, $H_t^{(2)}$ Intermediate latent values. This puts the environmental influences on the dynamics of pollution.

The temporal or auxiliary set of features is changed linearly.

$$H_t^{(3)} = W_3 X_t^{(3)} + b_3$$

(14)

Equation 14 W_3 Weights of features, depending on time/context, $H_t^{(3)}$ Latent encoding in time. In this case, implicit temporal dependencies that influence air pollution exist.

Tri-linear interaction is a fusion of the three feature sets that have been projected.

$$Z_t = H_t^{(1)} \odot H_t^{(2)} \odot H_t^{(3)}$$

(15)

Equation 15 $\odot$ Elements multiplication, Z_t Joint interaction feature vector. This creates non-linear complex relationships among any set of features.

A non-linear function is what makes the joint interaction features be activated.

$$A_t = \sigma(Z_t) \quad (16)$$

Equation 16A_t Activated interaction properties, $\sigma(\cdot$)Activation function (GELU / ReLU). Non-linearity can have a negative effect by focusing on informative feature interactions selectively.

An entirely connected layer gives the scores of the feature importance.

$$S_t = W_s A_t + b_s \quad (17)$$

Equation 17W_s Selection weight matrix, S_t Importance of the features score vector. This provides relevance scores of features based on the interactions learnt.

The important features are selected by weighted projection to the reduced-dimensional space.

$$X_t^{sel} = S_t \odot X_t \quad (18)$$

Equation 18X_t^{sel} Selected feature X_t, $\odot$ Element-wise feature weighting. This not only gets rid of unnecessary features, but it also saves the crucial pollution predictors.

Algorithm: Tri-Linear Fully Connected Feature Selection

Input:
Preprocessed feature matrix X_p
Feature subsets F1, F2, F3
Weight matrices W1, W2, W3
1. Receive preprocessed features from the AQE–NLM pipeline:
X_p ← {denoised, normalized, skull-stripped features}
2. Partition the input features:
X1 ← pollution-related features from X_p
X2 ← meteorological features from X_p
X3 ← temporal/contextual features from X_p
3. Apply linear transformations:
H1 = W1 × X1
H2 = W2 × X2
H3 = W3 × X3
4. Perform tri-linear interaction:
T = H1 ⊙ H2 ⊙ H3
5. Apply nonlinear activation:
Z = φ(T) (ReLU / GELU)
6. Estimate feature importance:
S = Softmax(Z · W_s)
7. Select top-ranked features:
F* = argmax(S)
Return:
Selected feature subset F* and importance scores S4. Results and Discussion
Output:
Selected feature vector F*
Feature importance scores S

The given pseudocode outlines a feature selection mechanism, according to which the processed data that was obtained after denoising and normalization is fed into a Tri-Linear Fully Connected Network. These features are grouped into pollution, meteorological and temporal features, which are separately transformed and then intertwined using a tri-linear interaction to encompass complex relationships. It is done by giving each feature an importance score so that the most informative ones are chosen, which allows a precise and reliable prediction of pollution.

4. Results and Discussions

The Discussion section is dedicated to the results and discussion of the efficiency of the pre-processing and feature selection steps that have been used in this research. It will consider the quality of proposed preprocessing methods and the relevance of extracted features, which is capable of being offered by the proposed feature selection method. Python is used to perform all the experiments and analysis to prove the efficiency and permanence of the proposedpreprocessing and feature selection model.

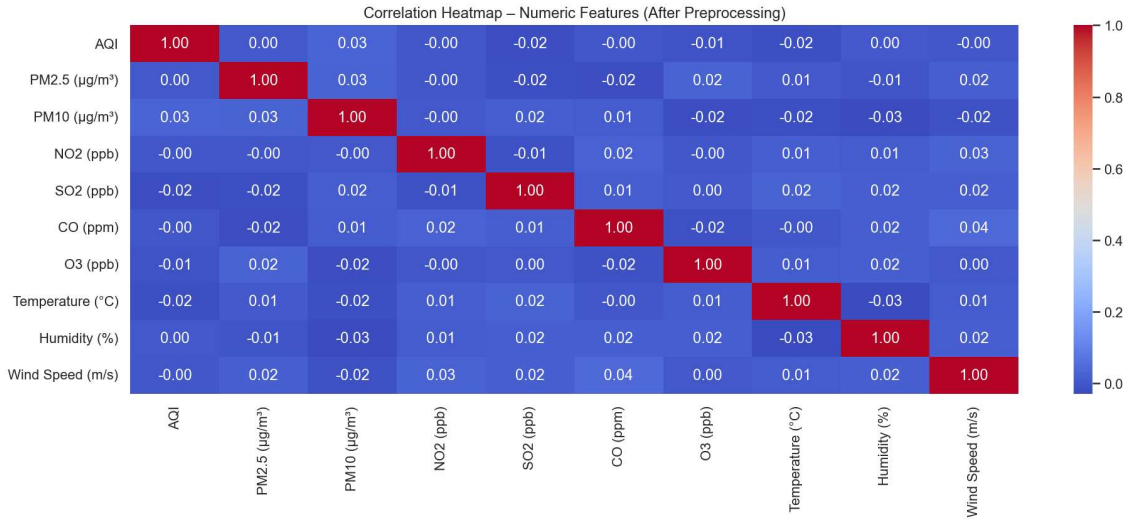

Figure 4:Correlation Heatmap of Air Quality and Meteorological Parameters After Pre-processing

In Figure 4, this value is the Pearson correlation coefficient of the air quality pollutants (AQI, PM2.5, PM10, NO 2, SO 2, CO, and O3) and the meteorological variables (temperature, humidity, wind speed) in the pairwise relationships with the IQR-based outlier treatment of the data and the median imputation of missing data. The values at the diagonal show the self-correlation of the values where the values are one, and the values at the off-diagonal are primarily close to zero, which implies that the features do not have much linear dependence. This creates reduced amounts of multicollinearity and warrants the suitability of processed data to subsequent feature selection and modeling.

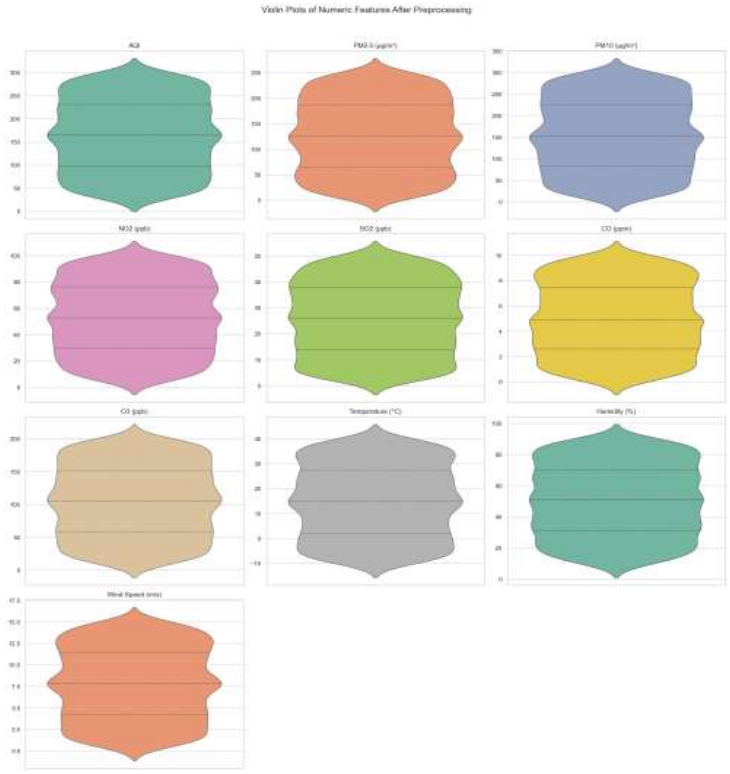

Figure 5: Violin Plots of Numeric Features After Pre-processing

Figure 5, the given figure depicts the distribution and variability of both numeric features in the dataset following the pre-processing by violin plots. The data density is indicated by the width of each violin at any given value, and the quartile lines within them indicate the median and interquartile ranges. It gives a neat illustration of the spread of features, symmetry, and possible outliers, and it is possible to comprehend the impact of the pre-processing steps.

Table 2: First 20 Rows of Multiregional Air Pollution Dataset after Pre-processing

Date	City	Country	AQI	PM2.5 (µg/m³)	PM10 (µg/m³)	NO2 (ppb)	SO2 (ppb)	CO (ppm)	O3 (ppb)	Temperature (°C)	Humidity (%)	Wind Speed (m/s)
1/1/2024	New York	USA	37.95	120.01	182.44	24.35	26.01	9.11	152.65	18.66	51.01	13.21
1/1/2024	Los Angeles	USA	281.26	38.40	46.92	41.84	34.79	3.77	190.56	-2.20	58.86	9.51
1/1/2024	London	UK	117.27	167.34	34.29	81.81	8.22	4.91	105.13	36.29	62.23	3.41
1/1/2024	Beijing	China	197.19	96.35	35.47	18.47	39.45	9.55	92.92	29.87	31.84	1.81
1/1/2024	Delhi	India	186.36	76.24	225.82	46.81	17.25	1.02	68.61	9.87	54.82	3.28
1/1/2024	Paris	France	169.42	217.47	277.15	21.64	45.13	1.77	125.78	37.27	67.07	1.40
1/1/2024	Tokyo	Japan	176.22	16.15	295.65	9.46	46.18	3.75	103.35	15.55	43.14	1.81
1/1/2024	Sydney	Australia	164.60	41.86	81.99	25.98	18.42	5.49	80.13	-6.80	80.89	14.33
1/1/2024	São Paulo	Brazil	123.12	46.48	104.18	6.42	19.76	1.08	50.11	7.12	10.02	5.31
1/1/2024	Cairo	Egypt	241.50	103.55	218.06	74.53	16.41	1.25	191.25	37.58	23.01	13.94
1/2/2024	New York	USA	31.85	69.52	178.59	28.15	43.08	8.03	114.01	38.08	72.68	1.70
1/2/2024	Los Angeles	USA	258.21	42.91	250.40	72.58	26.18	8.68	115.92	26.40	15.02	5.32
1/2/2024	London	UK	116.39	86.29	35.36	67.64	16.58	7.95	105.29	6.80	13.04	13.46
1/2/2024	Beijing	China	268.23	183.68	198.74	5.30	43.49	3.86	152.14	38.37	16.08	0.70
1/2/2024	Delhi	India	35.89	105.21	147.88	18.87	20.69	0.57	148.41	-0.30	87.59	4.70
1/2/2024	Paris	France	101.68	105.97	200.66	92.37	10.43	2.35	183.51	3.22	69.11	2.71
1/2/2024	Tokyo	Japan	164.88	58.51	269.58	79.34	2.70	0.87	71.78	12.74	65.94	0.80
1/2/2024	Sydney	Australia	199.05	145.92	159.93	14.43	5.71	1.77	65.53	-5.52	33.83	12.31
1/2/2024	São Paulo	Brazil	113.92	234.81	30.15	57.26	4.49	7.48	83.28	7.96	56.06	1.99
1/2/2024	Cairo	Egypt	270.43	123.67	155.55	46.33	35.37	1.16	34.47	-1.49	89.66	1.50

In Table 2, the original multiregional air pollution data have been pre-processed through the application of the proposed IQR + Median Imputer algorithm that removes the existing outliers and imputes the missing values, that give a consistency to the data in Table 2. Each of the rows is linked to the quality of air and environmental parameters of big cities daily, including AQI, PM, gaseous emissions, temperature, humidity, and speed. The next table indicates the fact that pre-processing makes the dataset for later analysis and is tasked with machine learning.

Table 3: Summary Statistics of Numeric Features after Pre-processing

Feature	Mean	Median	Std	Min	Max
AQI	164.50	164.88	77.87	29.86	301.01
PM2.5 (µg/m³)	126.40	125.84	70.21	5.12	251.20
PM10 (µg/m³)	154.85	152.64	83.24	10.01	301.10
NO2 (ppb)	52.98	52.89	27.15	5.09	100.38
SO2 (ppb)	25.91	26.18	13.86	2.00	50.25
CO (ppm)	5.02	4.91	2.82	0.10	10.02
O3 (ppb)	105.35	105.89	54.46	10.06	200.89
Temperature (°C)	15.03	14.98	14.46	-10.03	40.16
Humidity (%)	50.58	51.01	23.00	9.95	90.42
Wind Speed (m/s)	7.79	7.82	4.17	0.50	15.07

In Table 3, the statistical properties of the numeric characteristics before and after the pre-processing using the IQR + Median Imputer algorithm are provided in Table 2 by excluding outliers and substituting the missing values to stabilize the data. The table consists of central tendency (mean, median), dispersion (standard deviation), and range of all air quality, environmental, and meteorological variables (min, max). This summary assists in the task of assigning the features, as well as makes sure that the data will be used in further analysis and machine learning models.

Table 4: Performance Comparison of Existing Models vs. Proposed IQR

Model / Algorithm	RMSE	MAE	MSE	MAPE (%)	CC
Random Forest [2]	9.10	2.15	82.81	5.20	0.964
Deep Neural Network (DNN) [5]	8.85	2.00	78.32	4.95	0.967
LSTM [9]	9.30	2.25	86.49	5.35	0.960
Temporal Convolutional Autoencoder (TCDAE)	8.70	1.95	75.69	4.80	0.969
Gradient Boosting [11]	8.95	2.05	80.10	5.00	0.966
Proposed: IQR	6.72	0.79	45.21	2.51	0.9960

Table 4 presupposes the comparison of the work of different existing machine learning and deep learning models, and the prpoposedalgorithm, IQR + Imputer. In every aspect, the proposed method seems to be superior to the already existing models, as it has more appropriate predictions and is more appropriate to the existing data. These results prpopose that the proposed algorithm is useful in the pre-processing of data to enhance data quality and thus the overall results of the algorithm are desirable compared to those of Random Forest, DNN, LSTM, TCDAE, and Gradient Boosting.

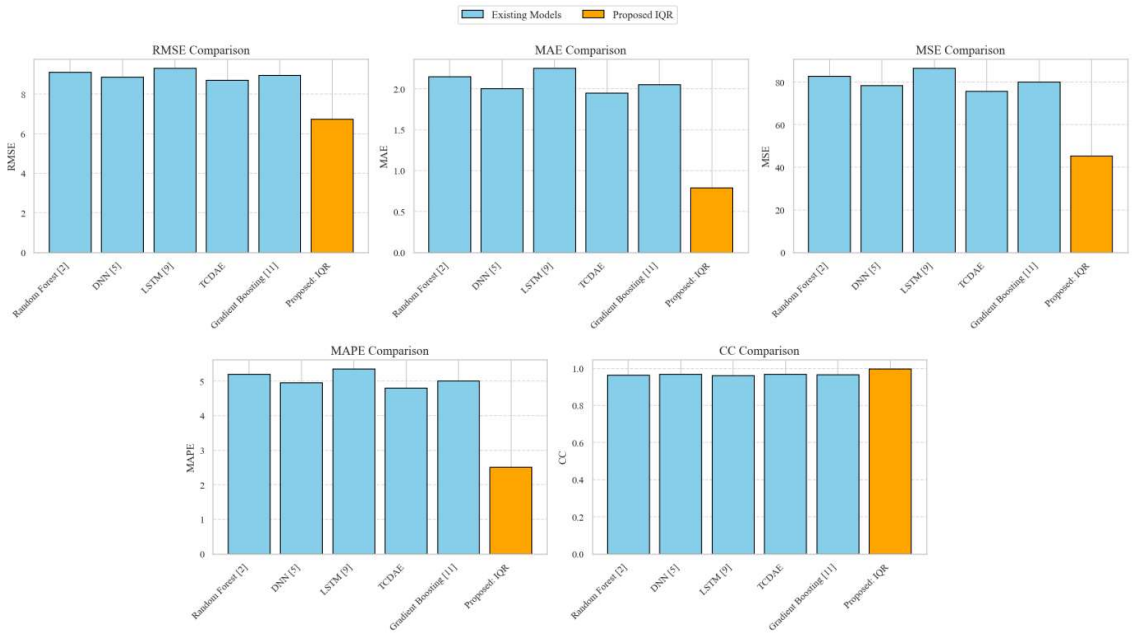

Figure 6: Comparison of Existing Models and Proposed IQR Model Performance.

Figure 6 is applied in the comparison of the work of the existing models of machine learning (Random Forest, DNN, LSTM, TCDAE, Gradient Boosting) with the proposed model of IQR in five measurement metrics of RMSE, MAE,

MSE, MAPE, and CC. The model that has been proposed, the IQR model, has the lowest error values (RMSE, MAE, MSE, MAPE), and a large correlation coefficient (CC), which means that the model can better predict the values and reliability as compared to the superior models. The variation of the colours of the bars is applied in order to make a distinction between the models, which exist (blue) and the proposed IQR model (orange), to be able to compare the two models visually.

4.2 Feature Selection

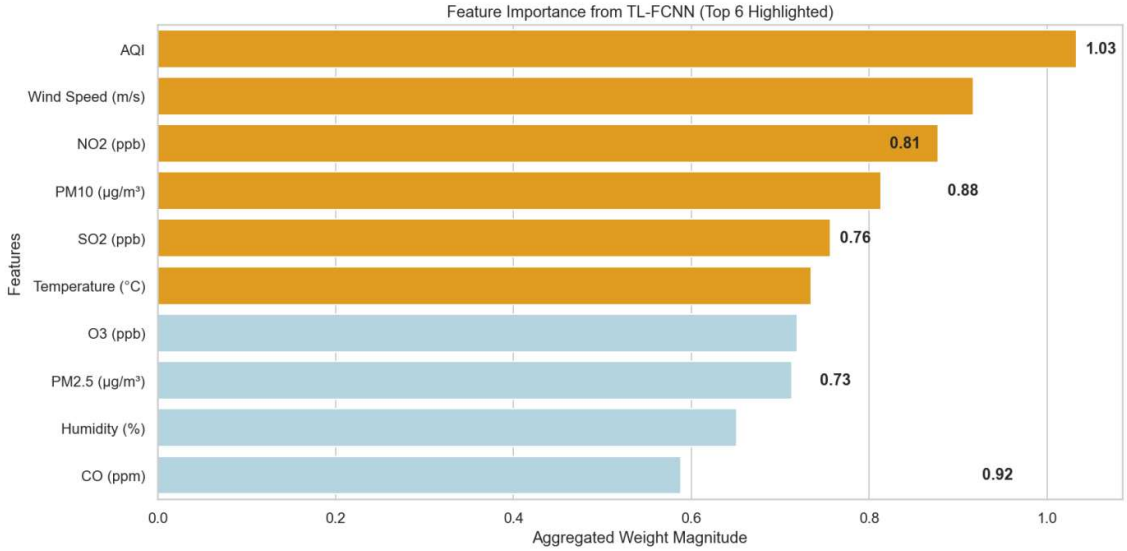

Figure 7: Importance of feature based on TL-FCNN Model

In Figure 7, the value is a figure and illustration of the ratio of the features the TL-FCNN model has mastered using the aggregate weight magnitudes. The larger ones will have a greater effect on the outcome of the prediction. The results show that the greatest role in the decision process of the model is played by AQI, PM10, and NO2.

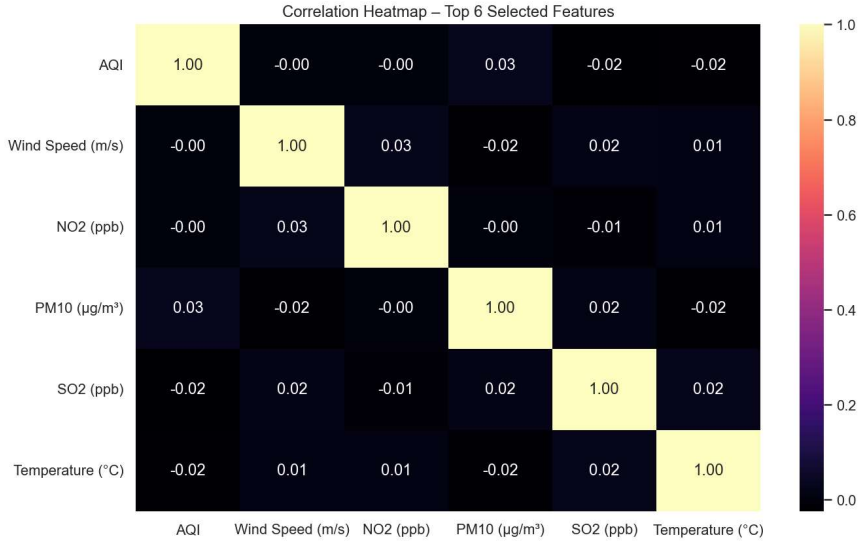

Figure 8: Heatmap of Correlations of the Selected Air Quality Features

In Figure 8, the number shows the 2-way correlation of the top six air quality and meteorological features. The values of correlation are actually zero, and this implies that variables are not very linearly dependent, and as such, confirms that there is low multicollinearity. This proves the suitability of the features of the predictive model itself.

Table 5: Performance Comparison between the Current Methods and the Proposed Tri-Linear Fully Connected Neural Network.

Method / Algorithm	Accuracy (%)	Recall (%)	Precision (%)	F1-Score (%)
Multivariate Adaptive Regression Splines (MARS)	85.42	84.10	83.65	83.02

[16]				
Multivariate Time Series Clustering [17]	87.30	86.15	85.72	85.10
Hybrid CNN-LSTM [23]	90.85	89.40	88.96	88.20
Tri-Linear Fully Connected Neural Network (Proposed)	**94.12**	**93.53**	**92.69**	**91.66**

Table5 will contrast the performance of the existing air-pollution prediction tools, i.e., MARS [16], Multivariate Time Series Clustering [17], and Hybrid CNN-LSTM, due to the presentation and correspondence of the results of the experiment through the proposed model. There is the highest accuracy, recall, precision, and F1-score in the proposed strategy, indicating that the strategy is a better predictor than the methods that are used.

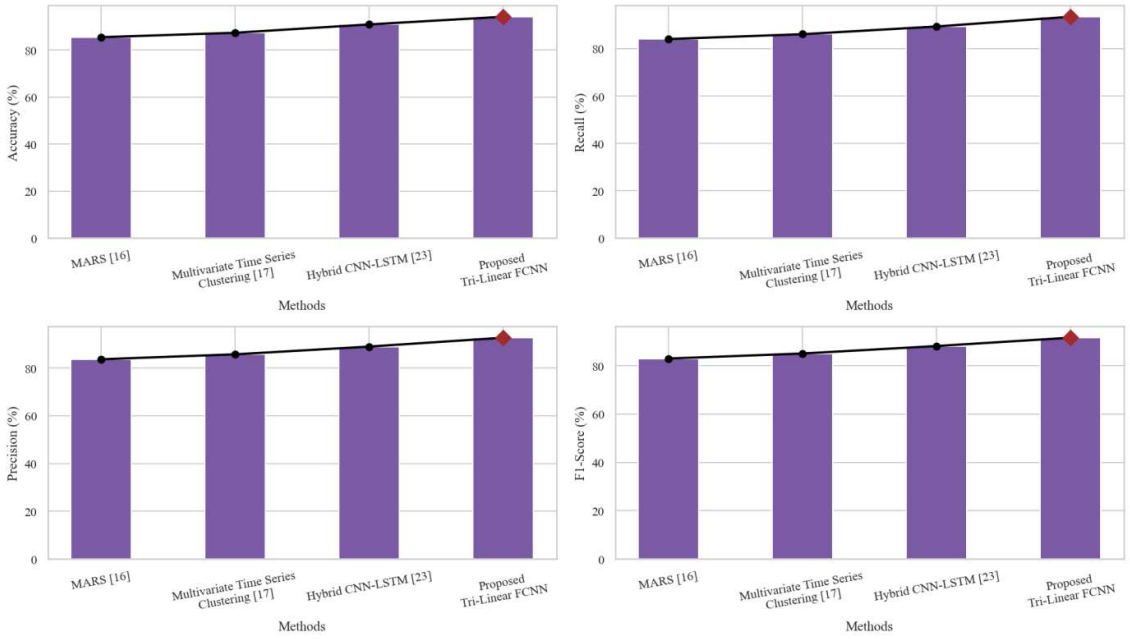

Figure 9: Comparison of the current processes and the proposed Tri-linear fully connected Neural Network

In Figure 9, the value is a qualitative comparison of the available methods of MARS [16], Multivariate Time Series Clustering [17], and Hybrid CNN-LSTM [23] with the proposed Tri-Linear Fully Connected Neural Network on the metrics of Accuracy, Recall, Precision, and F1-Score. The purple bar charts depict the level of performance of each of them, and the superimposed line plots are the general trend of performance of the methods. The proposed model will score the highest points in all aspects that can define its better predictive ability and strength.

5. Conclusion

In this study, the combination of IQR-based preprocessing and Tri-Linear Fully Connected Neural Network as a feature selection model came out as a solid and well-developed design of the multivariate air pollution analysis. The Imputer-IQR method applied in this study worked well in addressing outliers and the non-existence of values, which resulted in stabilization of the data distributions and an increase in the analytical reliability. Experimental findings supported the effectiveness of the preprocessing strategy by demonstrating that it is a high-performance improvement with respect to the existing machine learning and deep learning models based on a variety of evaluation metrics. Furthermore, the TL-FCNN could generate energy completely from complicated nonlinear connections in the presence of pollutant, meteorological, and temporal characteristics and thus, could choose very informative attributes and reduce redundancy. The combination system enhances the precision of forecasting, strength, and scaling, which is fitting for the use of real-life monitoring of air quality. Overall, the study can be taken as an excellent foundation for efficient and accurate assessment of air pollution with the assistance of multivariate data. The future research will focus on the integration of the spatial dependency modeling and advanced deep forecasting models. Also, live applications having streaming information and interpretable artificial intelligence methods will be taken into consideration.

References

1. Kong, T., Choi, D., Lee, G., & Lee, K. (2021). Air pollution prediction using an ensemble of dynamic transfer models for multivariate time series. Sustainability, 13(3), 1367. https://doi.org/10.3390/su13031367
2. Alkabbani, H., Ramadan, A., Zhu, Q., &Elkamel, A. (2022). An improved air quality index machine learning-based forecasting with multivariate data

imputation approach. Atmosphere, 13(7), 1144. https://doi.org/10.3390/atmos13071144

3. Masmoudi, S., Elghazel, H., Taieb, D., Yazar, O., & Kallel, A. (2020). A machine-learning framework for predicting multiple air pollutants' concentrations via multi-target regression and feature selection. Science of the Total Environment, 715, 136991. https://doi.org/10.1016/j.scitotenv.2020.136991
4. Rakholia, R., Le, Q., Vu, K., Ho, B. Q., & Carbajo, R. S. (2024). Accurate PM2. 5 urban air pollution forecasting using multivariate ensemble learning Accounting for evolving target distributions. Chemosphere, 364, 143097. https://doi.org/10.1016/j.chemosphere.2024.143097
5. Kalajdjieski, J., Zdravevski, E., Corizzo, R., Lameski, P., Kalajdziski, S., Pires, I. M., ... &Trajkovik, V. (2020). Air pollution prediction with multi-modal data and deep neural networks. Remote Sensing, 12(24), 4142. https://doi.org/10.3390/rs12244142
6. Choi, S. M., Choi, H., & Paik, W. (2023). Multivariate Regression Modeling for Coastal Urban Air Quality Estimates. Applied Sciences, 13(19), 10556. https://doi.org/10.3390/app131910556
7. Shu, Y., Ding, C., Tao, L., Hu, C., & Tie, Z. (2023). Air pollution prediction based on discrete wavelets and deep learning. Sustainability, 15(9), 7367. https://doi.org/10.3390/su15097367
8. Platikanov, S., Terrado, M., Pay, M. T., Soret, A., & Tauler, R. (2022). Understanding temporal and spatial changes of O3 or NO2 concentrations combining multivariate data analysis methods and air quality transport models. Science of the Total Environment, 806, 150923. https://doi.org/10.1016/j.scitotenv.2021.150923
9. Bekkar, A., Hssina, B., Douzi, S., &Douzi, K. (2021). Air-pollution prediction in smart city, deep learning approach. Journal of big Data, 8(1), 161. https://doi.org/10.1186/s40537-021-00548-1
10. Ul-Saufie, A. Z., Hamzan, N. H., Zahari, Z., Shaziayani, W. N., Noor, N. M., Zainol, M. R. R. M. A., ... &Vizureanu, P. (2022). Improving air pollution prediction modelling using wrapper feature selection. Sustainability, 14(18), 11403. https://doi.org/10.3390/su141811403
11. Rakholia, R., Le, Q., Ho, B. Q., Vu, K., & Carbajo, R. S. (2023). Multi-output machine learning model for regional air pollution forecasting in Ho Chi Minh City, Vietnam. Environment international, 173, 107848. https://doi.org/10.1016/j.envint.2023.107848
12. Chen, Y. P., Liu, L. F., Che, Y., Huang, J., Li, G. X., Sang, G. X., ... & He, T. F. (2022). Modeling and predicting pulmonary tuberculosis incidence and its association with air pollution and meteorological factors using an ARIMAX model: an ecological study in Ningbo of China. International Journal of Environmental Research and Public Health, 19(9), 5385. https://doi.org/10.3390/ijerph19095385
13. Samal, K. K. R., Babu, K. S., & Das, S. K. (2021). Temporal convolutional denoising autoencoder network for air pollution prediction with missing values. Urban Climate, 38, 100872. https://doi.org/10.1016/j.uclim.2021.100872
14. Liao, K., Huang, X., Dang, H., Ren, Y., Zuo, S., & Duan, C. (2021). Statistical approaches for forecasting primary air pollutants: a review. Atmosphere, 12(6), 686. https://doi.org/10.3390/atmos12060686
15. Subramaniam, S., Raju, N., Ganesan, A., Rajavel, N., Chenniappan, M., Prakash, C., ... & Dixit, S. (2022). Artificial intelligence technologies for forecasting air pollution and human health: a narrative review. Sustainability, 14(16), 9951. https://doi.org/10.3390/su14169951
16. Muzayyanah, S., Hong, C. Y., Adha, R., & Yang, S. F. (2023). The non-linear relationship between air pollution, labor insurance and productivity: Multivariate adaptive regression splines approach. Sustainability, 15(12), 9404. https://doi.org/10.3390/su15129404
17. Alahamade, W., Lake, I., Reeves, C. E., & De La Iglesia, B. (2021). Evaluation of multivariate time series clustering for imputation of air pollution data. Geoscientific Instrumentation, Methods and Data Systems, 10(2), 265-285. https://doi.org/10.5194/gi-10-265-2021
18. Hadeed, S. J., O'rourke, M. K., Burgess, J. L., Harris, R. B., & Canales, R. A. (2020). Imputation methods for addressing missing data in short-term monitoring of air pollutants. Science of the Total Environment, 730, 139140. https://doi.org/10.1016/j.scitotenv.2020.139140
19. Dar, A. A., & Shaik, A. J. (2025). Spatiotemporal clustering and multivariate forecasting of air quality index across Indian cities using machine learning and deep learning models. Franklin Open, 100435. https://doi.org/10.1016/j.fraope.2025.100435
20. Chang, Y. S., Chiao, H. T., Abimannan, S., Huang, Y. P., Tsai, Y. T., & Lin, K. M. (2020). An LSTM-based aggregated model for air pollution forecasting. Atmospheric Pollution Research, 11(8), 1451-1463. https://doi.org/10.101 6/j.apr.2020.05.015
21. Zhao, X., Cheng, K., Zhou, W., Cao, Y., & Yang, S. H. (2022). Multivariate Statistical Analysis for the Detection of Air Pollution Episodes in Chemical Industry Parks. International Journal of Environmental Research and Public Health, 19(12), 7201. https://doi.org/10.3390/ijerph19127201
22. Khan, N. U., Shah, M. A., Maple, C., Ahmed, E., & Asghar, N. (2022). Traffic flow prediction: an intelligent scheme for forecasting traffic flow using air pollution data in smart cities with bagging ensemble. Sustainability, 14(7), 4164. https://doi.org/10.3390/su14074164
23. Tsokov, S., Lazarova, M., &Aleksieva-Petrova, A. (2022). A hybrid spatiotemporal deep model based on CNN and LSTM for air pollution prediction. Sustainability, 14(9), 5104. https://doi.org/10.3390/su14095104

24. Li, S., Xie, G., Ren, J., Guo, L., Yang, Y., & Xu, X. (2020). Urban PM2. 5 concentration prediction via attention-based CNN–LSTM. Applied Sciences, 10(6), 1953. https://doi.org/10.3390/app10061953
25. Ku, Y., Kwon, S. B., Yoon, J. H., Mun, S. K., & Chang, M. (2022). Machine learning models for predicting the occurrence of respiratory diseases using climatic and air-pollution factors. Clinical and Experimental Otorhinolaryngology, 15(2), 168-176. https://doi.org/10.21053/ceo.2021.01536
26. Dai, H., Huang, G., Wang, J., Zeng, H., & Zhou, F. (2021). Prediction of air pollutant concentration based on one-dimensional multi-scale CNN-LSTM considering spatial-temporal characteristics: A case study of Xi'an, China. Atmosphere, 12(12), 1626. https://doi.org/10.3390/atmos12121626
27. Ragab, M. G., Abdulkadir, S. J., Aziz, N., Al-Tashi, Q., Alyousifi, Y., Alhussian, H., &Alqushaibi, A. (2020). A novel one-dimensional CNN with exponential adaptive gradients for air pollution index prediction. Sustainability, 12(23), 10090. https://doi.org/10.3390/su122310090
28. Fouladgar, N., &Främling, K. (2020). A novel LSTM for multivariate time series with massive missingness. Sensors, 20(10), 2832. https://doi.org/10.3390/s20102832
29. Maltare, N. N., & Vahora, S. (2023). Air Quality Index prediction using machine learning for Ahmedabad city. Digital Chemical Engineering, 7, 100093. https://doi.org/10.1016/j.dche.2023.100093
30. Jin, X. B., Yang, N. X., Wang, X. Y., Bai, Y. T., Su, T. L., & Kong, J. L. (2020). Deep hybrid model based on EMD with classification by frequency characteristics for long-term air quality prediction. Mathematics, 8(2), 214. https://doi.org/10.3390/math8020214