Noname manuscript No.
(will be inserted by the editor)

Autoencoder multimodal framework for the classification and prediction of air quality in indian cities

Routhu Srinivasa Rao · Lakshmana Rao Kalabarige · Alwyn R Pais · Aditya Kumar Sahu

the date of receipt and acceptance should be inserted later

Abstract Due to the rapid increase of industries around the cities, monitoring and predicting the quality of air is much needed to control the adverse health effects especially in developing countries like India. Majority of the recent works adapted into using machine learning algorithms for predicting the air quality index (AQI) and classifying the air quality into AQI buckets. In this paper, we proposed a multimodel autoencoder model which uses XGBoost machine learning algorithm for the prediction and classification of AQI. We used various imputation techniques to fill the missing values. Also, multimodel autoencoder is generated using MICE, KNN and SVD imputed data to extract the features for the prediction of AQI. SMOTE is used to convert the data into balanced data. From the experimental analysis, it is observed that multimodal autoencoder with multimodal imputed data achieved a significant accuracy of 97.14% and 93.87% with and without SMOTE method respectively. Similarly, the proposed model also achieved an R2 score of 0.9578 and RMSE of 27.59 in predicting the quality of air in Indian cities.The classification and regression results of the proposed model outperformed baseline models with significant difference in various evaluation metrics.

Keywords AQI · India · Environmental sustainability · $PM_{2.5}$ · PM_{10} · Imputation · autoencoder

✉ Routhu Srinivasa Rao
Department of Computer Science and Engineering, Gandhi Institute of Technology and Management, Visakhapatnam, Andhra Pradesh 530045, India. E-mail: srouth@gitam.edu

Lakshmana Rao Kalabarige
AI Research Lab, Computer Science and Engineering, GMR Institute of Technology, Rajam, India. E-mail: lakshmanarao.k@gmrit.edu.in

Alwyn R Pais
Information Security Research Lab, Department of Computer Science and Engineering, National Institute of Technology, Surathkal, Karnataka, India 575025 E-mail: alwyn@nitk.ac.in

Aditya Kumar Sahu
Amrita School of Computing, Amrita Vishwa Vidyapeetham, Amaravati, Andhra Pradesh, 522503, India. E-mail: adityasahu.cse@gmail.com

1 Introduction

Recently released WHO (2022) air quality report presents the air quality statistics of entire globe. The report analyzed air quality of 6743 cities of 117 countries for the last 10 years from 2010 to 2019. This report studied the density of harmful air pollutants such as CO, CO_2, Particulate Matter (PM)(PM_{10} and $PM_{2.5}$), NO_2, SO_2, O_3, NH_3, Pb, etc. The report statistics shows an incremental growth in harmful air pollutants such as $PM_{2.5}$, PM_{10},and NO_2. It is reported that over 1.7 million Indian population lost their lives due to high concentration of particulate matter(PM_{10} and $PM_{2.5}$) in air. The report also listed top 20 highly polluted cities in 117 countries in which 18 cites are from India. These statistics shows how severely India effected with air pollution.

IQAir[1] is a Swiss based air quality assessment company that protects and assess concentration of airborne pollutants concentration in air. It released air quality report 2021 which covers 6475 locations in 118 countries. The report statistics says that the India ranked 5^{th} position in global air pollution ranking. The annual average concentration of $PM_{2.5}$ is 58.1μg/m3 which is 11 times higher than the recommendation of WHO. The Delhi is in top position with 85μg/m3 annual average concentration of $PM_{2.5}$ which is 17 time higher than

[1] https://www.iqair.com/world-air-quality-report

the recommendation of WHO. The IQAir report also released most polluted country and city wise rankings among South Eastern Asia Countries. In which, India is ranked 4^{th} and 13 Indian cities are identified as most polluted with $PM_{2.5}$ of annual average concentration of 85μg/m3 among top 16 air polluted cities and remaining three are from Pakistan.

TheHealth-Effects-Institute (2019) released a report on Burden of Disease Attributable to Major Air Pollution Sources in India. It specifies various combustion procedures namely residential bio masses, open burning, industrial coal, power plant coal, transportation, brick production, and distributed diesel are main causes to increase the concentration of airborne pollutants $PM_{2.5}$, PM_{10}, CO, CO_2, NO_2, SO_2, O_3, NH_3, Pb, etc,. The report says that approximately 109.04 million Indian people were died till the year 2015 due to high concentration of airborne pollutants PM_{10} and $PM_{2.5}$. Assuming that there is no implementation of stringent air pollution control measures and basing on statistics shown in (Health-Effects-Institute, 2019; Swiss-Air-Quality-Technology-Company, 2021; WHO, 2022) it is forecasted that over 0.0017 billion people may die in the year 2030 and it may be over 0.0036 billion by the year 2050. The particles like dirt, dust, smoke, soot and gas liquid droplets with 10 micrometres mixed with air causes various diseases such as skin infections, eye problems, heart diseases, and mainly breathing problems since these can be easily inhaled into and deposited in lungs. Similarly, the fine particulates with 2.5 micrometres are very less in size. Hence, more dangerous and harmful which causes sever lung infections since it is deposited deeply into lungs.

The annual average concentration of PM2.5 and PM10 of Indian cities in the year 2019 is shown in Figure 1. The annual average concentration of PM2.5 and PM10 are given in Figure 1.a) and Figure 1.b) respectively. It gives a clear picture that the density of PM2.5 in air is higher than the standard value 35 μg/m3 as shown in Table 1. Similarly, PM10 is also very high in all Indian cities. In India, the recommended annual average concentration of each pollutant(CPCBCCR, 2014) is given in Table 1

From Figure 1 and Table 1 it is learned that all cites of India are in between satisfactory to moderate level. Hence, with no proper measures to control air pollution, it is estimated that within soon all Indian cities may easily reach either poor or very poor category because of rapid increase of air pollutants concentration due to rapid urbanization, industrialization and increased usage of motor vehicles etc.

The quality of air is classified into six categories as shown in Table 1 according to WHO (2022), Central Pollution Control Board(CPCB)[2], and The U.S. Environmental Protection Agency (EPA)[3]. Each AQI class is as follows:

1. Good: AQI value from 0 to 50 is an indication for good quality of air which may not create any health issues.
2. Satisfactory: AQI value from 51 to 100 indicated as Satisfactory which causes breathing issues to sensitive groups.
3. Moderate: AQI ranging from 101 to 200 indicated as Moderate level which causes for respiratory problems to children or elderly people.
4. Poor: AQI ranging from 201 to 300 indicated as Poor category which causes for respiratory problems to all category of people when they were exposed to air for longer duration.
5. Very Poor: AQI ranging between 301 to 400 classified as Very poor category and this certainly leads to illness of lungs.
6. Severe: AQI ranging from 401 to 500 indicated as severe which definitely causes various health hazards.

The high level of AQI is more dangerous to the mankind. In this connection, the forecasting of AQI levels in advance motivated the scientists to build a forecasting model for continuous monitoring of AQI levels. The monitoring and forecasting of AQI in urban and industrial areas is highly essential and challenging task with increasing industrial development, motor field, and transportation. To achieve the same, research community focussed in developing appropriate AQI prediction model using Artificial Intelligence methods (Rybarczyk and Zalakeviciute, 2021).

Till now, statistical, probability, deterministic, and physical approaches were used to calculate AQI index. These approaches were less efficient in handling large amount of data and complex in nature. Hence, in recent past, research community found that the Artificial approaches such as Machine Learning (ML) and deep learning (DL) are best to mitigate complexity in AQI calculation and prediction. The ML and DL models are efficient in handling large amount of data, reliable, adaptable, and consistent. Hence, we have attempted to use combination of ML and DL models to predict AQI value and classification AQI level. The proposed model considers six year of AQI data of all Indian cities with twelve air pollutants provided by CPCB. The proposed model applies pre-processing techniques such as data cleaning and balancing. The removal of unwanted data, management of outliers, and use of appropriate imputation models as part of data cleaning. Other relevant

[2] https://app.cpcbccr.com/ccr

[3] https://www.epa.gov/sites/default/files/2016-04/documents/2012_aqi_factsheet.pdf

Table 1: The annual average concentration of air pollutants

AQI Category (Range)	PM10 24-hr	PM2.5 24-hr	NO2 24-hr	O3 24-hr	CO 8-hr (mg/m3)	SO2 24-hr	NH3 24-hr	Pb 24-hr
Good (0-50)	0-50	0-30	0-40	0-50	0-1.0	0-40	0-200	0.05
Satisfactory (51-100)	51-100	31-60	41-80	51-100	1.1-2.0	41-80	201-400	0.6-1.0
Moderate (101-200)	101-250	61-90	81-180	101-168	2.1-10	81-380	401-800	1.1-2.0
Poor (201-300)	251-350	91-120	181-280	169-208	10.1-17	381-800	801-1200	2.1-3.0
Very poor (301-400)	351-430	121-250	281-400	209-748	17.1-34	801-1600	1201-1800	3.1-3.5
Severe (401-500)	430+	250+	400+	748+	34+	1600+	1800+	3.5+

approaches like data scaling to maintain equal degree of weight to all the features, balancing of all class labels to improve classification accuracy and use of ML and DL based models on cleansed data to predict and classify AQI.

The contributions of the proposed work are given as follows. Firstly, the proposed model uses autoencoders to extract the features from the given dataset. Secondly, multimodal imputation is developed using various imputation techniques. Thirdly, autoencoder multimodal framework is designed with various imputed techniques applied on the dataset. Finally, the proposed piece of work analysed six years of day wise all Indian cities air pollution data collected from central pollution control board (CPCB) with twelve features.

The remaining section of the paper is described as follows. Section 2 provides the recent literature that uses various machine earning algorithms for the AQI prediction. Proposed work of this paper is given in Section 3. Various experiments on the dataset and the corresponding results with proposed model is given in Section 4. Finally, we conclude the paper in Section 5.

2 Related Work

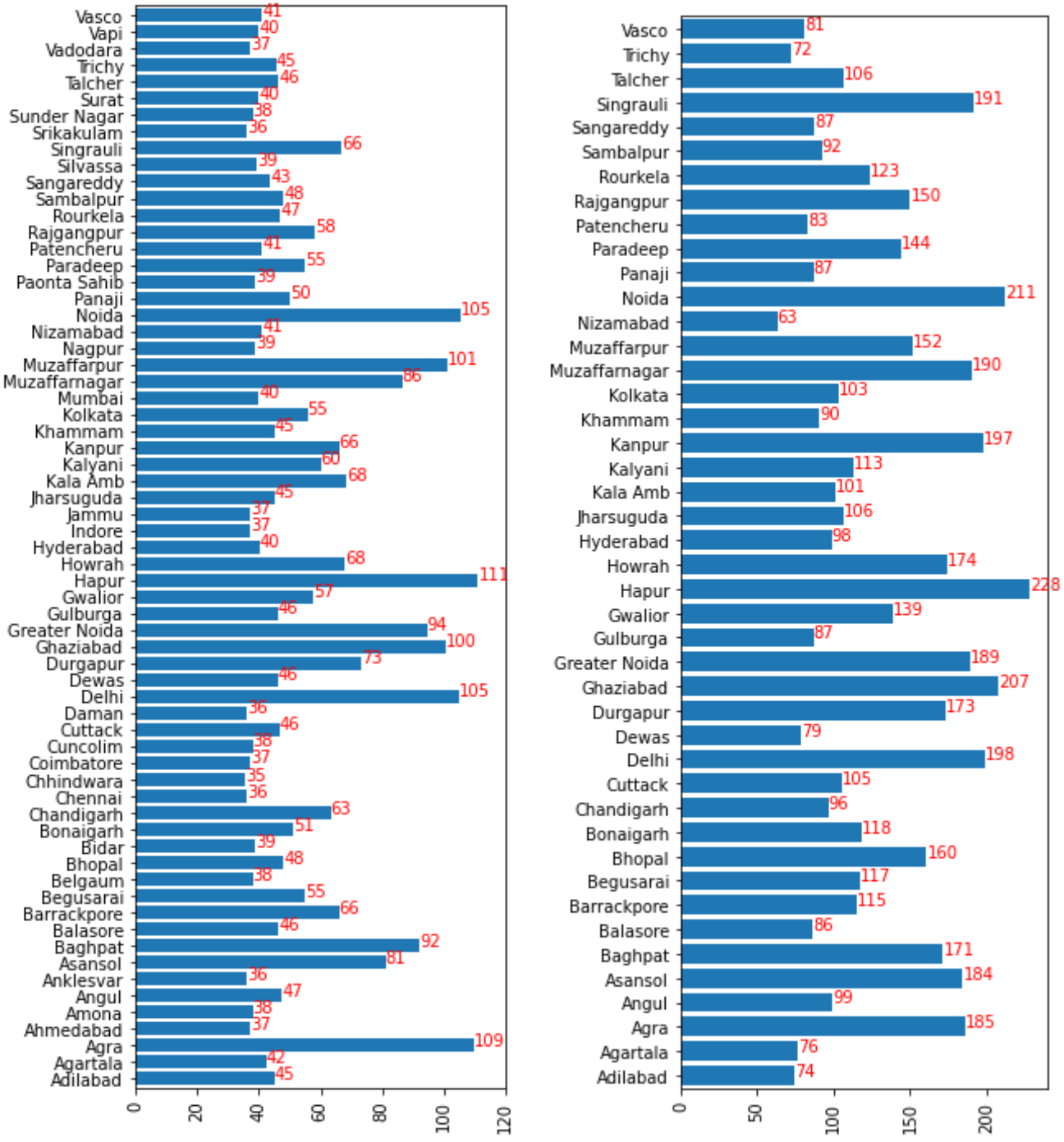
This section discuss the existing models which employs machine learning approaches to either predict and classify AQI. Some of the ML and DL models are summarized below:

Liang et al (2020) work considers air quality data of three regions of Taiwan such as south, north, and central with 91672, 94453, and 94145 instances respectively. Each dataset considered six features namely O_3, particulate matter 2.5 ($PM_{2.5}$), PM_{10}, CO, SO_2, and NO_2. It employs support vector machine (SVM) with three different kernels (polynomial, radial basis function (RBF), and Linear), AdaBoost with three different kernels (Square, Linear, and Exponential), Random forest (RF), Stacking ensemble, and Artificial neural network models to analyse air quality data of three regions. The model is evaluated using R^2-score, root mean square error (RMSE), and Mean absolute error (MAE). The results shows that the Stacking ensemble model gave greater R^2-score, lesser MAE and RMSE.

Castelli et al (2020) analyzes air quality data of California. The dataset consists 102090 instances with eight features such as $CO, SO_2, NO_2, Ozone$, $PM_{2.5}$, temperature, and humidity collected from the US Environmental Protection Agency (EPA).The work used data pre-processing techniques such as imputation of missing data, deletion of outliers, data transformation, and feature extraction. Finally, feature selection methods are applied to identify the relevant features. The authors also applied principal component analysis (PCA) to reduce its dimensionality and then employed SVM with RBF kernel (SVM-RBF) to analyze air quality at California. It is observed that the PCA SVM-RBF performed better than SVM-RBF.

Taylan et al (2021) proposed three models such as hybrid auto driven Artificial Neural Network (ANN), nonlinear autoregressive with external (exogenous) input (NARX) with a NN, and adaptive neuro-fuzzy inference (ANFIS) to estimate the air quality in the Jeddah city. The results says that the NARX model with RMSE value 0.0578 performed better than among all other models.

Kumar and Pande (2022) applied ML models such as K-Nearest Neighbour (K-NN), Gaussian Naive Bayes (GNB), Support Vector Machine (SVM), Random Forest (RF), and XGBoost on air quality data(from January 2015 to July 2020) of all Indian cities collected from central pollution control bard of India. The dataset has 29,531 instances with 12 features. The cleaning and balancing methods are applied before model training and observed that the XGBoost performed well for both classification and prediction. It obtains 91% of testing and 90% training accuracy and predicts AQI score with a R^2-Score of 0.834.



(a) The annual average concentration of PM2.5

(b) The annual average concentration of PM10

Fig. 1: The annual average concentration of PM2.5 and PM10 of Indian cities in the year 2019

Wood (2022) compares prediction performance of nine ML and three DL models on Dallas(USA) AQI dataset. The AQI data from the year 2015 to 2020 is collected for six pollutants namely Benzene (C6H6), Carbon monoxide (CO), Nitrogen dioxide (NO2), Ozone (O3), Particulate matter <2.5 μm (PM2.5), and Sulfur dioxide (SO2). Eleven, ground level metereological variables are combined with these six pollutants to identify the rise in air pollutants.

Siwek and Osowski (2016) employs RF and NN models to analyze Warsaw university region air quality dataset for the period of 13 years from 2001 to 2014. It also em-

ployed genetic algorithm and linear method of stepwise fit as feature section methods. The features selected using GA and stepwise fit are given as input to the RF for the prediction. Also, the selected features are fed to MLP, RBF and SVR to generate a ensemble.These two models are compared and the results demonstrated that model with RF performed better.

Fan et al (2017) proposed a deep recurrent neural network (DRNN) based spatiotemporal framework model for Air Quality Prediction. This work used air quality data of Jingjinji and country level meteorological data of China. The pre-processing methods such as forward-fix, mean-fix and decay-fix to impute missing values are employed in the model. The proposed work uses gradient boosting decision trees (GBDT), deep feedforward neural network (DFNN), and DRNN models to predict AQI. The DFNN performed in two variants such as DFNN1 and DFNN2. DRNN also has two variants such as DRNN1 and DRNN2. DRNN1 is a combination of one long term short memory (LSTM) layer and two dense layers, DRNN2 is built with two LSTM and two Dense layers, DFNN1 consists of three dense layers and DFNN2 has four dense layers. Each model tested for each imputation method and observed that the DRNN models outperformed GDBT annd DFNN models.

Chen et al (2019) employed sixteen different machine learning algorithms to predict the annual average of fine particulate matter (PM2.5) and nitrogen dioxide (NO2) concentrate across Europe. The performance of these sixteen models evaluated through R^2-score, RMSE and mean square error based R^2 MSE-(R^2) and concluded that the RF outperformed all other models.

Tao et al (2019) work used Beijing $PM_{2.5}$ dataset available in UCI repository. It applied regression and Deep leaning models to predict AQI. The regression models used are support vector regression (SVR), gradient boosting regressor (GBR), and decision tree regressor (DTR). The deep learning models such as RNN, LSTM, GRU, bidirectional GRU (BGRU), and convolutional based bidirectional gated recurrent unit (CBGRU) are applied on dataset for the prediction. The performance metrics such as RMSE, MAE, and symmetric mean absolute percentage error (SMAPE) are used to evaluate the model. The results demonstarated that the CBGRU outperformed all others models.

Zhu et al (2018) proposed a multi-task learning (MTL) approach with various regularization models. The regularization approaches such as Frobenius norm regularization, nuclear norm regularization, consecutive-hour-related regularizations, and l2,l1-norm regularization are applied on the combination of air quality and metrological data. These models considered an hourly based concentration of air pollutants such as ozone, particulate matter PM2.5 and sulfur dioxide. The models evaluated with RMSE and concludes that the consecutive-hour-related regularizations outperformed all other models.

Bougoudis et al (2016) proposed hybrid computational intelligence system for combined ML (HISYCOL) to analyze air pollution. In this work, the Self-organizing Map (SOM) clustering technique divides Athens air pollution data into four clusters such as SOM0, SOM1, SOM2, and SOM3. It applied Feed-forward Neural Network and RF on the dataset to generate the models. These two models applied on both original and clustered parts and applied Fuzzy inference rules on the outcome of RF and FFNN models. These models evaluated through R^2-Score and RMSE and concluded that the RF performs better than FNNN.

Shaban et al (2016) used a dataset which consists of 8833 instances with nine features and analyzed through ML algorithms SVM, M5P, and ANN. The data cleaning, pre-processing and feature engineering methods were applied before model training. The prediction trend accuracy and Normalized RMSE(NRMSE) are used to evaluate model performance and observed that the M5P outperforms all other models.

Gopalakrishnan (2021) used Linear Regression, Ridge Regression (RR), Elastic Net (EN), RF, and XGBoost machine learning models to analyze air quality of entire Oakland area. This piece of work applied data cleaning methods, and correlation based pre-processing steps on the data. The low RMSE and high R^2-Score is obtained by XGBoost compared to other models.

Kothandaraman et al (2022) proposed air quality prediction and forecasting based on $PM_{2.5}$ pollutant through six models Linear Regression, RF,KNN, Ridge and Lasso Regression, XGB, and AdaBoost. The dataset consists of 2044 instances and nine features of hourly data of Delhi city. It combines data of nine air pollutants with metrological data. The performance metrics such as MAE, MAPE, Mean Square Error (MSE), and RMSE of the models demonstrates that the XGB, AdaBoost, KNN and RF performs better among all six models.

Rybarczyk and Zalakeviciute (2021) analyzed and compared quality of air before and during COVID-19 lock-down through gradient boosting model. The model performance evaluated through RMSE and Pearson coefficient correlation. From the results it is observed that there is huge decrease in air pollution during lock-down.

Sanjeev (2021) proposed AQI prediction model through RF, SVM, and ANN. The proposed work applied data cleaning, feature selection and normalization techniques prior to the model training and it is concluded that the

RF performs better with 99.4% accuracy than SVM and ANN.

Harishkumar et al (2020) reported seven ML models such as LR, Lasso Regression, Ridge Regression, Random Forest Regression, Gradient Boost Regressor (GBR), K-Nearest Neighbour Regressor (KNNR), Multi-Layer Perceptron Regressor (MLPR), and Decision Tree Regressor (DTR) to investigate and analyze concentration of $PM_{2.5}$ throughout Taiwan. The performance of these models evaluated through RMSE, MAE, MSE and R^2-score and concluded that the GBR outperforms all other six models with low error rate and high R^2-score.

Monisri et al (2020) proposed IoT based ML model to predict quality of air. The sensor networks are prone to failures. Hence, collected data may have missing values. In this connection, this work applies techniques to handle missing values and Imputation methods to fill missing values. It also applied data normalization methods to assign equal weight to each feature and finally, the ML models such as RF, DT, and SVM are trained with pre-processed data. The MSE and RMSE metrics are used to evaluate and compare model performance and concluded that the RF gives best results than other models.

Nahar et al (2020) reported ML models of simple, medium and complex variants of DT, SVM, KNN, RF, and LR to classify quality of air. The mean value is used to impute missing values in a dataset. The results shows that the all DT variants gives 99.96% of accuracy.

Bhalgat et al (2019) reported Autoregressive and AutoRegressive Moving Average models to predict the concentration of SO_2 SO2 and $PM_{2.5}$ in air. It considers air quality data of Maharastra state of India and concluded that the $PM_{2.5}$ and SO_2 concentration is high in cities like Nagapur, Pune, and Mumbai of Maharastra. The data cleaning approaches like handling null values, and redundant instances applied before training the model.

Soundari et al (2019) reported an prediction model to find/forecast concentration of NO2, So2, Respirable Suspended Particulate Matter (RSPM), and Suspended Particulate Matter (SPM) in air through Neural Networks. The boundary value analysis is applied on the dataset to remove outliers. The proposed model achieved 96% accuracy in AQI classification.

Rybarczyk and Zalakeviciute (2018) follows a machine learning based approach to predict concentration of pollutants such as PM2.5, SO2, CO, NO3, and O3 in air. It used DTR, RFR, GBR and ANN based Multi-Layer Perceptron Regression models for the prediction of air quality. The model performance is evaluated through correlation coefficient and RMSE. The results shows that the RFR performed better than other regression models.

Ameer et al (2019) analyzed air quality data of five cities of China. Each dataset is different in its size and features. It employs regression models such as DT, RF, GBR, and MLP on the datasets. These models are trained and tested on all five cities and evaluated through MAE and RMSE. From the results, it is concluded that the performance of RF is better than all other regression models. Li et al (2019) reported an Amazon Web Services(AWS) pipe line to store, process, and make predictions through ML models such as Logistic regression and Random Forest. These models applied on ten years AQI data of California to classify AQI levels.

Wang and Kong (2019) reported an improved decision tree to predict air quality of three different cities of China. The proposed work performs seven steps such as attributes saving, pre-processing (data cleaning, handling outliers and default values), discretize the vales of each feature, calculation of the weighted information gain rate of each attribute value, selects the maximum weighted information gain rate as root nodes, generation of new branch node when a candidate attribute value is not empty and finally, generation of new branch nodes continues until the candidate node is empty. It employees ID3, C4.5, new C4.5, and Back propagation Neural Network (BPNN) models. The performance of the model evaluated through AUC, ROC and PR curves. The results shows that the performance of C4.5 is better than all other models.

Juarez and Petersen (2021) proposed an air quality prediction approach using seven ML models such as Linear Regression, KNN, SVM, RF, DT, AdaBoost, XGB, and one deep learning approach such as bidirectional LSTM (BD-LSTM). It considered hourly based air quality data of Delhi city in India. It trains and tests all these models in two phases such as in first phase it trains all the models with one year data and secondly it takes five years data to train the model. The comparative results says the model with five years data performs better with highest R^2-score.

Wibowo et al (2021) reported LSTM based air quality prediction approach. This work uses air data of different locations of Jakarta during COVID-19 outbreak. The LSTM with different optimizer and epochs are used to train and test the model and its performance evaluated through RMSE. The results showed that the model achieved RMSE of 11.45 and 11.43 with PM10 and O3 pullutant respectively using Adam optimizer.

Jiang et al (2021) reported an Empirical Mode Decomposition (EMD)-hybrid model for short-term AQI forecasting. The EMD approach extracts decomposed components as features and then, the decomposed fea-

tures of AQI and other air pollutants are given as input to the two parallel 1D Convolutional neural networks (1DCNN). The output of the 1DCNN is adopted as input features for training a Long short-term memory (LSTM) network. The proposed EMD-hybrid model is compared with other models such as 1DCNN, LSTM, Contextual LSTM(CLSTM), EMD-1DCNN, EMD-LSTM, and EMD-CLSTM. The results shows that the EMD-hybrid outperforms other models with 87.9% accuracy on AQI dataset of Guangzhou China and 85.28% accuracy on AQI dataset of Changchun.

Zhan et al (2022) proposed a decomposition ensemble model based on broad learning system considering the air quality data of Huainan and Fuyang cities of China. The proposed model compared with eighteen models and concluded that the proposed model outperforms all other models by obtaining RMSE=14.3584, MAE=9.9473, and R^2=0.9282 on air quality dataset of Huainan city and RMSE=14.9268, MAE=9.0483, and R^2=0.9228 on Fuyang city air quality dataset respectively.

Liu and Zhang (2021) reported A hybrid AQI time series prediction model such as empirical wavelet transform(EWT)-Sample Entropy(SE)-variational mode decomposition(VMD). The proposed EWT-SE-VMD model uses imperialist competitive algorithm(ICA) and echo state network(ESN). The ICA approach is used to select optimal subseries feature subset and ESN is an high-precision complex prediction system used as a predictor in the time series prediction structure. The proposed model performs better with low error rate among all other models.

Wu et al (2022) proposed an ensemble learning model such as complete ensemble empirical mode decomposition with adaptive noise (CEEMDAN) with Fuzzy entropy(FE) and LSTM (CEEMDAN-FE-LSTM). The CEEMDAN is used to decompose the AQI data series onto more stable subseries, the FE is used to recombine the subseries with similarity as one to avoid over-decomposition, and finally, fed to LSTM for training. The results shows that the proposed model outperforms all other models such as autoregressive integrated moving average(ARIMA), LSTM, and EEMD-LSTM with RMSE=9.45 and high R^2-score=0.9334.

3 Proposed Work

The model proposed utilizes XGBoost algorithm to predict the Air Quality Index (AQI) score and AQI category based on various pollutant values as input. The design of the model is outlined in Figure 2, and includes actions such as filling in missing values, eliminating outliers, balancing the data, and feature generation using autoencoders to produce a final prediction or classification.

3.1 Dataset

The Air Quality Index (AQI) data from Central Pollution Control Board of India (CPCBI) from 2015 to 2020, a total of six years, has been collected (CPCBI, 2022; VOPANI, 2019). The dataset, consisting of 16 features and 29532 instances, is described in Table 4. These features include levels of pollutants such as $PM_{2.5}$, PM_{10}, NO, NO_2, NO_x, NH_3, CO, SO_2, O_3, Benzene, Toluene, and Xylene. The dataset is divided into training and testing data, with the training data used to fit the model and the testing data used to evaluate the model's performance.

3.2 Data Cleaning

This section employs various techniques to clean the data, including removing unwanted features or observations, imputing missing values, and eliminating outliers. The cleaning process is crucial in constructing an effective model.

3.2.1 Unwanted Observations or attributes

The dataset was initially cleaned by removing the city and date features, which were determined to be irrelevant for the prediction or classification of AQI scores. Before removing outliers, observations with empty values in the AQI or AQI bucket were removed, because the machine learning model uses a supervised learning approach that relies on target attributes for fitting the data.Furthermore, observations with empty values in all features were also removed. The dataset was left with 24801 observations and 12 features, excluding the target attributes. Descriptive statistics such as mean, standard deviation, minimum value, maximum value, 25th percentile, 50th percentile, and 75th percentile were calculated and presented in Table 2.

3.2.2 Handling Outliers

Outliers are data points that are scattered or isolated from the rest of the data points, as shown in Figure 3. The figure displays the boxplot of all pollutants in the given dataset. Outliers can have a negative impact on machine learning models, such as longer training time, decreased accuracy, and poor results. Therefore, outlier detection and treatment is crucial. Methods commonly used for outlier detection include the difference between

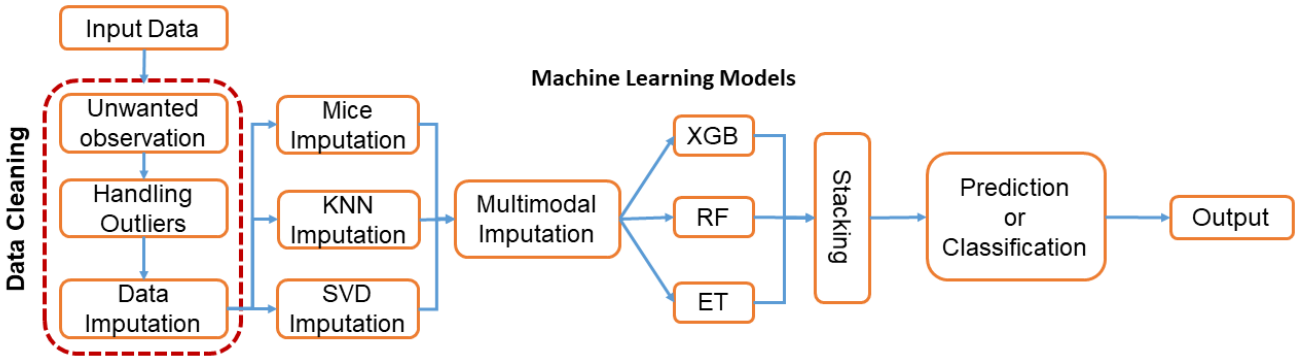


Fig. 2: The architecture of proposed model

Table 2: Air Quality Index (AQI) dataset

Stats	PM2.5	PM10	NO	NO2	NOx	NH3	CO	SO2	O3	Benzene	Toluene	Xylene
count	24172	17764	24463	24459	22993	18314	24405	24245	24043	21315	19024	9478
mean	67.48	118.45	17.62	28.98	32.29	23.85	2.34	14.36	34.91	3.46	9.52	3.59
std	63.07	89.49	22.42	24.63	30.71	25.87	7.07	17.43	21.72	16.04	20.88	6.75
min	0.04	0.03	0.03	0.01	0	0.01	0	0.01	0.01	0	0	0
25%	29	56.78	5.66	11.94	13.11	8.96	0.59	5.73	19.25	0.23	1.02	0.39
50%	48.78	96.18	9.91	22.1	23.68	16.31	0.93	9.22	31.25	1.29	3.57	1.42
75%	80.92	150.18	20.03	38.24	40.17	30.36	1.48	15.14	46.08	3.34	10.18	4.12
max	914.94	917.08	390.68	362.21	378.24	352.89	175.81	186.08	257.73	455.03	454.85	170.37

max and min values of a feature, interquartile range (IQR), and data skewness. To treat outliers, techniques such as condition-based removal, trimming, capping, and statistical outlier replacement are recommended. The proposed work uses IQR to identify outliers, where IQR is calculated by evaluating the difference between the 75^{th}(Q3) and 25^{th} (Q2) percentiles. Each feature is analyzed for outliers using IQR. The boundaries outside Q1 and Q3 are calculated using UL= Q3 + 1.5*IQR and LL = Q1 - 1.5*IQR, where UL is the upper limit and LL is the lower limit. Values below LL and above UL are considered outliers and are replaced with blanks or NaN (treated as missing values). Note that the missing values at the time of data collection and missing values due to the outlier analysis are different and are fixed using various imputation methods. The count of missing values in each pollutant at the time of data collection and after outlier analysis is given in Table 3. From the table, it is clearly observed that there is a significant increase in missing values after the outlier analysis, indicating the presence of a large number of outliers in the given data. The missing values are fixed at the Data Imputation component.

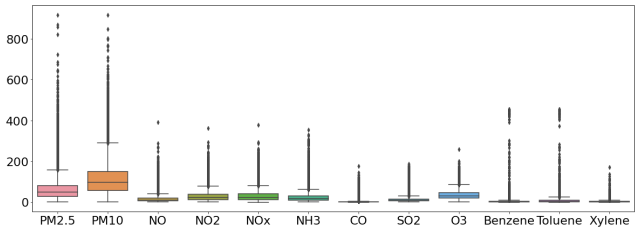


Fig. 3: The boxplot of all pollutants in the AQI dataset

3.2.3 Data Imputation

Missing values are a common problem in real-life datasets that are built through sensor networks, due to sensor malfunctioning, misconfiguration, and non-functioning. The accuracy of machine learning models is reduced by missing values, and some machine learning models may not be able to handle datasets with missing values. Therefore, it is highly recommended to employ appropriate methods for missing data imputation(Adnan et al, 2022; Tamboli, 2022). One way to handle missing values is to remove observations that have missing values in any of the attributes, but this can lead to the risk of losing important information. Therefore, missing

Table 3: Columnwise missing values before and after outlier analysis

Features	# of missing values	
	Before Outlier analysis	After Outlier analysis
PM2.5	629	2540
PM10	7037	8060
NO	338	2680
NO2	342	1414
NOx	1808	3460
NH3	6487	7445
CO	396	2685
SO2	556	2965
O3	758	1422
Benzene	3486	4944
Toluene	5777	7900
Xylene	15323	16182
AQI_Bucket	0	0

values should be replaced with an optimal value that is closer to the original value.

Imputation techniques such as replacing missing values with mean, mode, or median, replacing them with previous or next values, and classifier-based imputation are widely used approaches. Univariate techniques such as mean, median, or mode only focus on single features(column level) but do not take into account the correlation between features. Based on experimental studies, it is observed that these methods are fast in computation but do not result in significant performance improvements. Therefore, we consider multivariate imputers such as KNN imputer, Multiple Imputation by Chained Equation (MICE), and SVD imputer. KNN imputer replaces missing values for each sample based on the average of the values of nearest neighbors. MICE performs multiple regression over a random sample of data and takes an average of multiple regression values. It attempts to impute each feature with missing values by using the function of others. Finally, SVD imputer uses singular value decomposition to extract a set of mutually orthogonal patterns to estimate features with missing values in the given dataset. Each imputation technique is applied to each attribute to identify and replace missing values in the given dataset. The imputed data generated with KNN, SVD, and MICE are fed to various machine learning algorithms for both classification and prediction of the AQI score.

3.3 Multimodal Autoencoder

An Autoencoder is a type of neural network that is an unsupervised learning algorithm. It uses backpropagation to generate output values that are similar to the input values. The main goal of autoencoder is to learn an efficient and low-dimensional representation of the input data, which can be used for tasks such as dimensionality reduction, data compression, anomaly detection, and feature learning. Autoencoder consist of an encoder and a decoder, where the encoder maps the input data into a lower dimensional space and the decoder tries to reconstruct the original input from the encoded representation. The Autoencoder is trained to minimize the difference between the input and output by backpropagation. During the training, the network learns to extract the most important features from the input data and discard the less important ones, in order to obtain a compressed representation. Once trained, the encoder can be used to encode new data and the decoder can be used to reconstruct the original data from the encoded representation. In this work, We use multiple Autoencoders to extract most important features from different imputed data(MICE,KNN and SVD). The multimodal autoencoder is constructed by combining features from various sources(imputed data) to generate a feature vector which is further fed to machine learning algorithms for the classification or prediction.

3.4 Balancing methods

The data balancing(Kalabarige and Maringanti, 2022) of an imbalanced dataset is one of the most widely used data pre-processing approach to make equal number of instance for both minority and majority class labels of an dataset. The model training with imbalanced class labels may increase mis-classification rate. Hence, it is advised to feed balanced data as input to train the model. In the proposed piece of work it is observed that the considered AQI dataset is imbalanced. Hence, the Synthetic Minority Oversampling Technique(SMOTE) is applied to make AQI dataset as balanced.

3.5 Machine Learning algorithms

The proposed model used XGBoost algorithm for both classification and prediction of AQI. The classification models classify the quality of air into five labels as illustrated in Table 1 and the prediction models forecast AQI based on the values given 11 features as described in Table 4. The imputed data, autoencoder based data

and both combined vectors are fed to XGBoost algorithm for predicting the AQI score. The predicted outputs are compared with ground truths using various evaluation metrics for both classification and prediction.

4 Experimentation Results

In this section, several experiments are conducted to evaluate the proposed model that classifies and predicts the quality of air. The dataset for the experimentation is taken from the source of Central Pollution Control Board (CPCB)[4]. The dataset is divided into two parts where one part is used for training the model and the other part is used for testing the model. 80% of data is used for training and remaining portion is used for the evaluation. The dataset undergoes various preprocessing techniques to improve the quality of data. In this work, we mainly concentrate on building a multimodal with data arriving from various imputation methods such as KNN, MICE and SVD imputers. These sources are used to generate new features using autoencoders thereby improving the performance of the model. Finally, features from multiple encoders with different imputed inputs are used to generate multimodal autoencoder to achieve the significant performance in both classification and prediction of AQI. Note that, XGBoost algorithm is used for both classification and prediction as it belongs to family of ensemble algorithms and also refered as effective algorithm compared to weak learner. Various traditional metrics such as Recall, Precision, F1-Score are used to evaluate the model. Due to the multi-classification task, macro and weighted averages of the above metrics are calculated. The macro average is the arithmetic mean of all the per class specific measure(f1, recall or precision). For example, macro averaged recall score is the arithmetic mean of all per class recall scores. Similarly, weighted average is calculated by taking the mean of all per class recall scores along with the class support. The support is defined as the number of occurrences of the specific class in the given dataset. Due to the presence of imbalanced data, SMOTE is applied on the dataset to perform over sampling resulting balanced data. For the regression experiments, traditional measures such as Mean Squared Error (MSE), Root Mean Squared Error(RMSE), R2 Score and Mean Absolute Error(MAE). To ease the discussion and understanding we use following terms.

- **KNN_D:** Dataset preprocessed from KNN Imputer
- **MICE_D:** Dataset preprocessed from MICE Imputer
- **SVD_D:** Dataset preprocessed from SVD Imputer
- **KNN_AE_D:** Auto encoder features extracted from KNN_D
- **MICE_AE_D:** Auto encoder features extracted from MICE_D
- **SVD_AE_D:** Auto encoder features extracted from SVD_D
- **All_AE_D:** multimodal of autoencoders(KNN_AE_D, MICE_AE_D and SVD_AE_D)
- **All_Imp_D:** multimodal of imputers(KNN_D, MICE_D and SVD_D)

[4] https://app.cpcbccr.com/ccr/#/caaqm-dashboard-all/caaqm-landing/data.

4.1 Exp 1: Evaluation of model with KNN Imputed data using autoencoder

In this section, we firstly conduct experiment on KNN_D using XGBoost algorithm to identify the effectiveness of the KNN Imputer. Secondly, autoencoder is applied on the KNN_D to extract the features, followed by feeding the features to XGBoost for the classification and predicting the quality of the air. Finally, KNN imputed data i.e. KNN_D combined with autoencoder extracted features to generate new dataset which is fed to XGBoost for the classification and regression. The results of the above experiments are given in Table 5. Also, SMOTE is applied on each dataset to balance the data and the results with balanced data also given in Table5. From the results, it is observed KNN imputed data with XGBoost outperformed other experiments in both imbalanced and balanced environment with an accuracy of 81.46% and 88.16%. The experimental setup for classification is adopted for the predicting the quality of air using XGBoost Regressor. The results with regression model is shown in 9. The results demonstrates that KNN imputed data (KNN_D) combined with autoencoder features performed better with R2 Score of 0.8274 and RMSE of 55.83.

4.2 Exp 2: Evaluation of model with MICE Imputed data using autoencoder

Similar to the Exp 1, we conduct three experiments with MICE_D, MICE_AE_D and both combined data. The results are shown in Table 6 with and without SMOTE. From the results, it demonstrates that combined data of MICE_D, MICE_AE_D with XGBoost outperformed other models with a significant accuracy 83.17% and 90.54% using imbalanced and balanced configuration. Note that, features extracted autoencoder

Table 4: Air Quality Index (AQI) dataset

Features	Description
City	The air quality data of 26 Indians cities are considered.
Date	day wise air quality data of five years from 1-1-2015 to 1-7-2020 of each city is captured.
$\boldsymbol{PM_{2.5}}$	mixing of ultra fine particles with liquid droplets in air know as Particulate Matter with size of 2.5 microns or smaller in size.
$\boldsymbol{PM_{10}}$	mixing of fine particles in air is known as Particulate Matter with size of 10 microns.
NO	Nitrogen monoxide. It is released due to industrial combustion process, motor vehicles, and power stations
$\boldsymbol{NO_2}$	Nitrogen dioxide. It is released through oxidation of NO through combustion process
$\boldsymbol{NO_x}$	It is group of highly reactive gasses which include NO, NO_2 and other forms of Nitrogen
$\boldsymbol{NH_3}$	is called as Ammonia. It is released from agricultural activities, animal husbandry, fertilizers, etc.
CO	Carbon Monoxide is an color less gas released from fires, industrial processes, kitchen chimneys, etc.
$\boldsymbol{SO_2}$	Sulfur dioxide released from automobiles, chemical industries, etc.
$\boldsymbol{O_3}$	Ozone consisting of 3 atoms. It is mainly released from industries
Benzene	The coal and oils burning and Tobacco smoking are causes for air pollutant Benzene
Toluene	The motor vehicles are main emission resources for air pollutant Toluene
Xylene	The burning of coal, wood, and petroleum products are the main source for air pollutant Xylene
AQI	It is calculated based on available air pollutants. AQI calculation needs at least three pollutants in which either $PM_{2.5}$ or PM_10 should be one.
AQI_Bucket	Based on AQI value the Indian cites are indicated in one of the five categories such as Good, Satisfactory, Moderate, Poor, Very Poor, and Severe.

Table 5: AQI classification on KNN Imputed data with and without SMOTE

	Classifiers	Precision		Recall		F1 Score		Accuracy
		Macro	Weighted	Macro	Weighted	Macro	Weighted	
Without SMOTE	**KNN_D+XGB**	80.35	81.28	78.64	81.46	79.37	81.28	81.46
	KNN_AE_D+XGB	73.33	75.90	70.71	76.15	71.92	75.97	76.15
	KNN_D+KNN_AE_D+XGB	79.34	80.38	76.90	80.57	78.02	80.41	80.57
With SMOTE	**KNN_D+XGB**	88.20	88.17	88.19	88.16	88.18	88.16	88.16
	KNN_AE_D+XGB	82.32	82.34	82.29	82.30	82.28	82.30	82.30
	KNN_D+KNN_AE_D+XGB	87.93	87.95	87.88	87.89	87.89	87.90	87.89

Table 6: AQI classification on MICE Imputed data with and without SMOTE

	Classifiers	Precision		Recall		F1 Score		Accuracy
		Macro	Weighted	Macro	Weighted	Macro	Weighted	
Without SMOTE	**MICE_D+XGB**	72.32	75.64	67.38	75.97	69.40	75.60	75.97
	MICE_AE_D+XGB	81.21	81.42	78.46	81.48	79.76	81.41	81.48
	MICE_D+MICE_AE_D+XGB	83.40	83.15	79.87	83.17	81.51	83.08	83.17
With SMOTE	**MICE_D+XGB**	83.39	83.38	83.47	83.47	83.40	83.40	83.47
	MICE_AE_D+XGB	87.86	87.84	87.85	87.82	87.85	87.82	87.82
	MICE_D+MICE_AE_D+XGB	90.56	90.54	90.56	90.54	90.55	90.53	90.54

Table 7: AQI classification on SVD Imputed data with and without SMOTE

	Classifiers	Precision		Recall		F1 Score		Accuracy
		Macro	Weighted	Macro	Weighted	Macro	Weighted	
Without SMOTE	**SVD_D+XGB**	82.66	86.26	80.52	86.43	81.53	86.32	86.43
	SVD_AE_D+XGB	83.66	86.83	81.15	86.92	82.25	86.81	86.92
	SVD_D+SVD_AE_D+XGB	87.01	89.55	85.13	89.62	86.00	89.55	89.62
With SMOTE	**SVD_D+XGB**	91.69	91.69	91.68	91.69	91.67	91.68	91.69
	SVD_AE_D+XGB	89.66	89.68	89.69	89.68	89.65	89.66	89.68
	SVD_D+SVD_AE_D+XGB	93.50	93.52	93.52	93.51	93.50	93.50	93.51

Table 8: AQI classification using Autoencoder Multimodal with and without SMOTE

	Classifiers	Precision		Recall		F1 Score		Accuracy
		Macro	Weighted	Macro	Weighted	Macro	Weighted	
Without SMOTE	**All_AE_D+XGB**	91.75	91.90	90.69	91.90	91.20	91.88	91.90
	All_AE_D+All_Imp_D+XGB	93.84	93.88	92.82	93.87	93.31	93.87	93.87
With SMOTE	**All_AE_D+XGB**	95.85	95.85	95.86	95.85	95.85	95.84	95.85
	All_AE_D+All_Imp_D+XGB	97.13	97.13	97.15	97.14	97.14	97.13	97.14

Table 9: AQI prediction with KNN Imputed data

Techniques	**MSE**	**RMSE**	**MAE**	**R2**
KNN_D+XGB	3600.08	60.00	30.26	0.8007
KNN_AE_D+XGB	3931.81	62.70	30.29	0.7824
KNN_D+KNN_AE_D+XGB	3117.49	55.83	27.49	0.8274

Table 10: AQI prediction with MICE Imputed data

Algorithms	**MSE**	**RMSE**	**MAE**	**R2**
MICE_D+XGB	1750.52	41.83	25.76	0.9031
MICE_AE_D+XGB	1582.37	39.77	23.22	0.9124
MICE_D+MICE_AE_D+XGB	1200.90	34.65	20.12	0.9335

performed significant role in the classification with an accuracy of 81.48% compared to model with MICE_D. The MICE_AE_D complements the MICE_D achieving the significant macro f1 score, precision and recall of 81.51%, 83.40% and 79.87% respectively. Also, attempt to predict the quality of air is made with regression model and the results are given in Table 10. From the results, it clearly indicates that combination of MICE_D and MICE_AE_D achieved better R2 score of 0.93335 and RMSE of 34.65 compared to KNN imputed data.

4.3 Exp 3: Evaluation of model with SVD Imputed data using autoencoder

The autoencoder based features are concatenated with SVD Imputed data to generate the comprehensive data. The results of the proposed model with comprehensive data is shown in Table 7. From the results, it is clearly observed that model with comprehensive data (SVD_AE_D + SVD_D) achieved better performance with an accuracy of 89.62% and 93.51%. The performance in this experiment is the highest so far compared with previous experiment sections. Also, XGBoost with autoencoder features (SVD_AE) achieved significant results with an accuracy of 86.92% without SMOTE and 89.68% with SMOTE demonstrating the importance of autoencoder based features in the classification model. The experiment to predict the AQI Score using regression model is conducted and the results are given in Table 11. The results in the Table shows the model with combined data (SVD_AE_D + SVD_D) achieved an R2 score of 0.9270 and RMSE of 36.29 which has lower performance compared to model with SVD imputed data. Note that, Model with SVD imputed data(SVD_AE_D + SVD_D) performed better than models with other imputed data in classification task but stood in second position in regression task.

Table 11: AQI prediction with SVD Imputed data

Algorithms	MSE	RMSE	MAE	R2
SVD_D+XGB	1721.68	41.49	24.99	0.9047
SVD_AE_D+XGB	1638.47	40.47	21.53	0.9093
SVD_D+SVD_AE_D+XGB	1317.51	36.29	19.64	0.9270

Table 12: AQI prediction with multimodal Imputed data

Techniques	MSE	RMSE	MAE	R2
All_AE_D+XGB	973.76	31.20	17.54	0.9461
All_AE_D+All_Imp_D+XGB	761.44	27.59	15.69	0.9578

Table 13: Classification comparison of proposed model with existing work

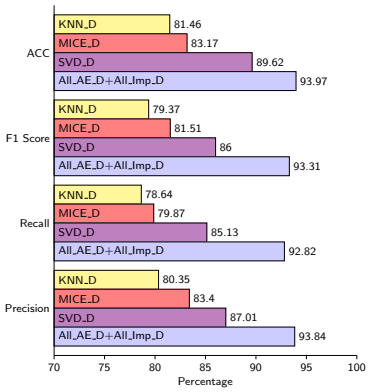
Metrics	W1	Proposed Model
Precision	96	97.13
Recall	95	97.15
F1 Score	91	97.14
Accuracy	90	97.14

4.4 Exp 4: Evaluation of model with multimodal autoencoder

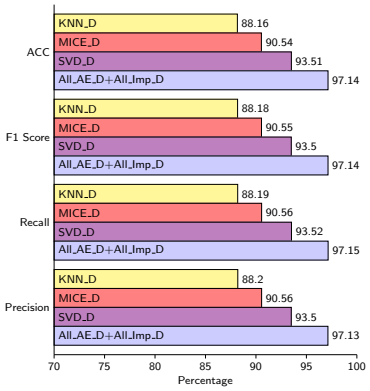
In this section,autoencoder multimodal is used for the generation of dataset where KNN_AE_D, MICE_AE_D and SVD_AE_D are combined. XGBoost algorithm is applied on the generated data for the classification and prediction of quality of air. The results are shown with and without SMOTE in Table 8. From the results, it is observed that autoencoder multimodal achieved an accuracy of 91.9% outperforming all previous models. Also, an experiment with autoencoder multimodal combined with multimodal imputed data is conducted, which resulted in highest classification accuracy with 93.87% and 97.14% using SMOTE and without SMOTE. The multimodal autoencoder combined with different data from various imputation techniques are also used for performing a regression task in predicting the AQI score. The results with regression model are given in Table 12. From the results, it is clearly seen that the proposed model with newly generated data(All_AE_D + All_Imp_D) performed better than other models with R2 score of 0.9578 and RMSE of 27.59 and is considered as final model for predicting the AQI in Indian cities.

Table 14: Regression comparison of proposed model with existing work

Metrics	W2	Proposed Model
MSE	7447.69	761.44
RMSE	86.3	27.59
MAE	23	15.69
R2 Score	0.864	0.9578



(a) Evaluation of models with exclusion of SMOTE

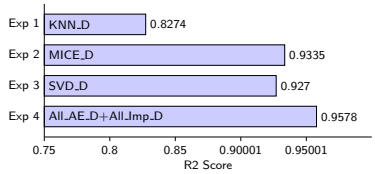


(b) Evaluation of models with inclusion of SMOTE

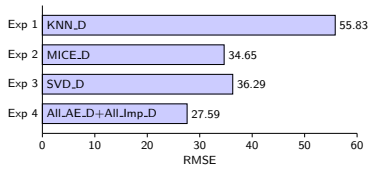
Fig. 4: Comparison of classification models with different metrics

4.5 Exp 5: Comparison of proposed work with existing works

In this section, we compare our proposed model with different experimental data carried out in previous experiments. From the results shown in Figure 4, it is clearly observed that multimodal autoencoder with multimodal imputation data outperformed other models with respect to accuracy, F1 score, Recall and Precision with inclusion and exclusion of SMOTE technique. Similarly, from the results in Figure 5, it is clealry observed



(a) Evaluation of models with R2 Score



(b) Evaluation of models with RMSE

Fig. 5: Comparison of regression models with different metrics

that the proposed multimodal autoencoder performed better than other models with significant R2 score and RMSE. We also compared our proposed model with Kumar and Pande (2022) work as the authors used same dataset for the experimentation. The comparison results with classification and regression shown in Table 13 and 14. From the results, it is demonstrated that there exists a significant difference of 7.14% in accuracy compared to existing model W1 in classifying the AQI bucket. Similarly, the proposed model achieved significant R2 score of 0.9578 and RMSE of 27.59 compared to 0.864 R2 score and 86.3 RMSE of existing work W2 in predicting the quality of air in Indian cities.

5 Conclusion

The proposed multimodal autoencoder performed better than other models with varying data in predicting the AQI. The autoencoder is designed with different imputed data using various imputers such as KNN, MICE and SVD. The SVD imputed data with XGBoost outperformed other imputation techniques in classifying the AQI buckets with an accuracy of 89.62% and 93.51% using exclusion and inclusion of SMOTE. Similarly, MICE imputed data with XGBoost outperformed other imputation techniques in predicting the air quality with an R2 score of 0.9335 and RMSE of 34.65. Finally, the proposed multimodal autoencoder with multimodal imputed data outperformed other models and existing models with a significant accuracy of 97.14% in classification and R2 score of 0.9578 in predicting the quality of air. In the future, we intend to use various deep learning algorithms for the feature extraction and filters, wrappers for the feature selection to improve the performance of the model.

6 Declarations

Ethics approval: Not applicable.

Consent to participate: Not applicable.

Consent to publication: Not applicable.

Funding: The author declares that there is no funding for this paper.

Conflict of Interest: The authors declare that they have no conflict of interest.

Authors Contributions:RSR and KLR conceptualized, performed experimentation and wrote the manuscript; ARP conceptualized and supervised the whole experiment; AKS supervised, reviewed, and edited the manuscript.

Availability of data and materials: The data used are included in the manuscript.

References

Adnan FA, Jamaludin KR, Wan Muhamad WZA, Miskon S (2022) A review of the current publication trends on missing data imputation over three decades: direction and future research. Neural Computing and Applications pp 1–16

Ameer S, Shah MA, Khan A, Song H, Maple C, Islam SU, Asghar MN (2019) Comparative analysis of machine learning techniques for predicting air quality in smart cities. IEEE Access 7:128,325–128,338

Bhalgat P, Bhoite S, Pitare S (2019) Air quality prediction using machine learning algorithms. International Journal of Computer Applications Technology and Research 8(9):367–390

Bougoudis I, Demertzis K, Iliadis L (2016) Hisycol a hybrid computational intelligence system for combined machine learning: the case of air pollution modeling in athens. Neural Computing and Applications 27(5):1191–1206

Castelli M, Clemente FM, Popovič A, Silva S, Vanneschi L (2020) A machine learning approach to predict air quality in california. Complexity 2020

Chen J, de Hoogh K, Gulliver J, Hoffmann B, Hertel O, Ketzel M, Bauwelinck M, Van Donkelaar A, Hvidtfeldt UA, Katsouyanni K, et al (2019) A comparison of linear regression, regularization, and machine learning algorithms to develop europe-wide spatial models of fine particles and nitrogen dioxide. Environment international 130:104,934

CPCBCCR (2014) National air quality index report. URL `https://app.cpcbccr.com/ccr`

CPCBI (2022) Central control room for air quality management - all india. https://app.cpcbccr.com/ccr//caaqm-dashboard-all/caaqm-landing/data

Fan J, Li Q, Hou J, Feng X, Karimian H, Lin S (2017) A spatiotemporal prediction framework for air pollution based on deep rnn. ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences 4:15

Gopalakrishnan (2021) Hyperlocal air quality prediction using machine learning. towards data science. https://towardsdatascience.com/hyperlocal-air-quality-prediction-using-machine-learning-ed3a661b9a71

Harishkumar K, Yogesh K, Gad I, et al (2020) Forecasting air pollution particulate matter (pm2. 5) using machine learning regression models. Procedia Computer Science 171:2057–2066

Health-Effects-Institute (2019) Burden of disease attributable to major air pollution sources in india. URL https://www.healtheffects.org/publication/gbd-air-pollution-india

Jiang W, Fu Y, Lin F, Liu J, Zhan C (2021) Empirical mode decomposition based deep neural networks for aqi forecasting. In: International Conference on Neural Computing for Advanced Applications, Springer, pp 757–769

Juarez EK, Petersen MR (2021) A comparison of machine learning methods to forecast tropospheric ozone levels in delhi. Atmosphere 13(1):46

Kalabarige LR, Maringanti H (2022) Symptom based covid-19 test recommendation system using machine learning technique. Intelligent Decision Technologies 16:181–191

Kothandaraman D, Praveena N, Varadarajkumar K, Madhav Rao B, Dhabliya D, Satla S, Abera W (2022) Intelligent forecasting of air quality and pollution prediction using machine learning. Adsorption Science & Technology 2022

Kumar K, Pande B (2022) Air pollution prediction with machine learning: a case study of indian cities. International Journal of Environmental Science and Technology pp 1–16

Li L, Li Z, Reichmann L, Woodbridge D (2019) A scalable and reliable model for real-time air quality prediction. In: 2019 IEEE SmartWorld, Ubiquitous Intelligence & Computing, Advanced & Trusted Computing, Scalable Computing & Communications, Cloud & Big Data Computing, Internet of People and Smart City Innovation (SmartWorld/SCALCOM/UIC/ATC/CBDCom/IOP/SCI), IEEE, pp 51–57

Liang YC, Maimury Y, Chen AHL, Juarez JRC (2020) Machine learning-based prediction of air quality. Applied Sciences 10(24):9151

Liu H, Zhang X (2021) Aqi time series prediction based on a hybrid data decomposition and echo state networks. Environmental Science and Pollution Research 28(37):51,160–51,182

Monisri P, Vikas KRN RK, Chethan Varma M (2020) Prediction and analysis of air quality using machine learning. International Journal of Advanced Science and Technology 29(05):6934–6943, URL http://sersc.org/journals/index.php/IJAST/article/view/18138

Nahar KM, Ottom MA, Alshibli F, Shquier MMA (2020) Air quality index using machine learninga jordan case study. Compusoft 9(9):3831–3840

Rybarczyk Y, Zalakeviciute R (2018) Regression models to predict air pollution from affordable data collections. Machine Learning—Advanced Techniques and Emerging Applications

Rybarczyk Y, Zalakeviciute R (2021) Assessing the covid-19 impact on air quality: A machine learning approach. Geophysical Research Letters 48(4):e2020GL091,202

Sanjeev D (2021) Implementation of machine learning algorithms for analysis and prediction of air quality. International Journal of Engineering Research & Technology (IJERT) 10(3):533–538

Shaban KB, Kadri A, Rezk E (2016) Urban air pollution monitoring system with forecasting models. IEEE Sensors Journal 16(8):2598–2606

Siwek K, Osowski S (2016) Data mining methods for prediction of air pollution. International Journal of Applied Mathematics and Computer Science 26(2):467–478

Soundari AG, Jeslin JG, Akshaya A (2019) Indian air quality prediction and analysis using machine learning. Int J Appl Eng Res 14(11):181–186

Swiss-Air-Quality-Technology-Company (2021) Interactive global map of 2021 pm2.5 concentrations by city. URL https://www.iqair.com/world-air-quality-report

Tamboli N (2022) All you need to know about different types of missing data values and how to handle it. https://www.analyticsvidhya.com/blog/2021/10/handling-missing-value/

Tao Q, Liu F, Li Y, Sidorov D (2019) Air pollution forecasting using a deep learning model based on 1d convnets and bidirectional gru. IEEE access 7:76,690–76,698

Taylan O, Alkabaa AS, Alamoudi M, Basahel A, Balubaid M, Andejany M, Alidrisi H (2021) Air quality modeling for sustainable clean environment using anfis and machine learning approaches. Atmosphere 12(6):713

VOPANI (2019) Calculating aqi (air quality index) tutorial. https://www.kaggle.com/code/rohanrao/calculating-aqi-air-quality-index-tutorial/notebook

Wang Y, Kong T (2019) Air quality predictive modeling based on an improved decision tree in a weather-smart grid. IEEE Access 7:172,892–172,901

WHO (2022) Who air quality database 2022. URL https://www.who.int/publications/m/item/who-air-quality-database-2022

Wibowo F, et al (2021) Prediction of air quality in jakarta during the covid-19 outbreak using long short-term memory machine learning. In: IOP Conference Series: Earth and Environmental Science, IOP Publishing, vol 704, p 012046

Wood DA (2022) Local integrated air quality predictions from meteorology (2015 to 2020) with machine and deep learning assisted by data mining. Sustainability Analytics and Modeling 2:100,002

Wu Z, Zhao W, Lv Y (2022) An ensemble lstm-based aqi forecasting model with decomposition-reconstruction technique via ceemdan and fuzzy entropy. Air Quality, Atmosphere & Health pp 1–13

Zhan C, Jiang W, Lin F, Zhang S, Li B (2022) A decomposition-ensemble broad learning system for aqi forecasting. Neural Computing and Applications pp 1–12

Zhu D, Cai C, Yang T, Zhou X (2018) A machine learning approach for air quality prediction: Model regularization and optimization. Big data and cognitive computing 2(1):5