LGURJCSIT Volume 06, Issue 02, Article ID 270, 15 pages
https://doi.org/10.54692/lgurjcsit.2022.0602270

Comparative Analysis of Machine Learning Techniques for Predicting Air Pollution

Farwa Akram[1], Sardar Usman[2], and M. Usman Ashraf [3], *
[1] Department of Computer Science & IT, Leads University, Lahore, Pakistan
[2] Department of Computer Science and SE, Gran Asian University Sialkot, Pakistan.
*, [3] Department of Computer Science, GC Women University Sialkot, Pakistan

Received 25 May 2022; Accepted 24 June 2022; Published 26 June 2022

ABSTRACT ***The modern and motorized way of life has cultured air pollution. Air pollution has become the biggest rival of robust living. This situation is becoming more lethal in developing countries and so in Pakistan. Hence, this inquiry was carried out to propose an architecture design that could make real-time prediction of air pollution with another purpose of scanning the frequently adopted algorithm in past investigations. In addition, it was also intended to narrate the toxic effects of air pollution on human health. So, this research was carried out on a large dataset of Seoul as an adequate dataset of Pakistan was not attainable. The dataset consisted of three years (2017-2019) including 647,512 instances and 11 attributes. The four distinctive algorithms termed Random Forest, Linear Regression, Decision Tree and XGBoosting were employed. It was inferred that XGB is more promising and feasible in predicting concentration level of NO_2, O_3, SO_2, PM_{10}, $PM_{2.5}$ and CO with the lowest RMSE and MAE values of 0.0111, 0.0262, 0.0168, 49.64, 41.68 and 0.1856 and 0.0067, 0.0096, 0.0017, 12.28, 7.63 and 0.0982 respectively. Furthermore, it was found out as well that the Random Forest was preferred mostly in the previous studies related to air pollution prophecy while many probes supported that air pollution is very detrimental to human health especially long-lasting exposure causes lung cancer, respiratory and cardiovascular diseases.***

Keywords: Machine Learning, Air Pollution Prediction, Seoul, Particulate Matter, Random Forest, Decision Tree, Linear Regression, XGBoosting

1. Introduction

Promising air quality is crucial for humans as well as for other creatures in the atmosphere for a decent living. Air quality means the air is without destructive pollutants. But in the modern era, the air is infected with several lethal and fatal pollutants that affect the quality of air and make it toxic to robust living. The most treacherous pollutants are SO_2, NO_2, O_3, CO, $PM_{2.5}$ and PM_{10}. These and many other pollutants make the air poisoned called air pollution. Air pollution refers to the contamination of air by different physical, chemical and biological factors. It is the problem faced by 99% of people all over the world [1].

Air pollution is boosting manifold in this modern era due to urbanization, excessive population and industrialization [2]. In recent years, this problem has turned into a hazardous one. Now, it is becoming one of the major causes of mortality and premature death on earth because it affects the respiratory system bitterly and reduces the age and function of the lungs. According to WHO, almost ten million people died annually due to air pollution [3]. So, concerned departments and academics are striving hard to develop systems to cope with this fatal problem of air pollution. Machine learning has assisted a lot in this matter. Systems-based machine learning has deep developed to predict the air pollution level so that timely measures to be taken to minimize the level of air pollution [4]. Machine and deep learning are useful as it gives real-time solution [5]. Many studies in the literature indicated that machine learning is very convenient to predict the air pollution level with the help of different algorithms [6], , [92-96]. The pollutants termed carbon dioxide, Sulfur dioxide, ozone, particulate matter, carbon dioxide, nitrogen oxide, ozone and hydrocarbon are the main causes of air pollution in the environment.

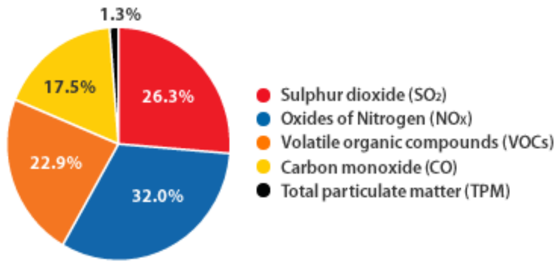


Figure 1. Most Common Pollutants (askiitians.com, 2018)

Air is the most fundamental element for all organisms on earth but air pollution has become the most critical and alarming problem in Pakistan. All the newspapers and news channels keep claiming that smog has crossed all the limits in Pakistan whose main cause is the burning of agriculture remains especially in rice harvesting season. Contaminated air is the sole cause of many respiratory and heart diseases. So, there is a terrible need that certain steps should be taken to avoid or minimize the level of air pollution. Pakistan staggering behind in using information technology like different techniques of the machine and deep learning for distinct motives like predicting air quality. Some investigations have been found that pointed out the air quality in Pakistan, but all those analyses had been carried out to know the current air quality. Only one study [7] has been found in which air prediction has been made using the machine learning techniques but this investigation was carried out on the dataset of China (2010-2015) consisted of just one pollutant named PM2.5. So, this research is being carried out on six pollutants with latest dataset of Seoul.

2. Literature Review

This section narrates the account of 32 investigations carried out in the past (2016-2021). These studies have been divided into two categories termed measurement of particulate matter and measurement of multiple pollutants.

Category# 1: Measurement of Particulate Matter ($PM_{2.5}$ & PM_{10}).

Enebish et al., [8] aimed to improve the evaluation of $PM_{2.5}$ exposure for Ulaanbaatar from 2010 to 2018 and applied six machine algorithms named as RF, GBM, SVM, MARS, GLMNET and GAM to predict the concentration level of $PM_{2.5}$. They concluded that RF and GBM performed better than others using LOLO and CV with R^2 of 0.82 while in the study of Masood & Ahmad [9] ANN gave better results. Usmani [10] asserted that the Partitioning & Regression Tree was more efficient and accurate in prediction. Bozdag [11] wanted to predict PM_{10} by employing LASSO, RF, KNN, XGB and ANN. The performance of all metrics was evaluated and the best performance was obtained on station 6 with ANN (R^2=0.58, RSME=20.8, MAE=14.43).

Sethi et al. [12] proposed a method to predict the concentration of $PM_{2.5}$ employing a feature selection approach known as “Causality Based Linear” with the help of the Delhi dataset. At first DT, RF, LR and NN were applied on the whole dataset in which LR gave promising results in predicting $PM_{2.5}$. Lee et al. [13] used the data of different 77 air monitoring stations and 560 weather stations. The experiments used RMSE, normalized RMSE (NRMSE), and R^2 as prediction performance metrics. The proposed method significantly enhanced the coefficient of determination (R^2) from 0.58 to 0.71 and reduced the RMSE from 8.56 to 7.06.

Doresawamy et al. [14] researched the $PM_{2.5}$ level in Taiwan which affects human health severely. They used the dataset of Taiwan AQM consisted of five years (2012-2017) comprised of 76 air stations and used RF, DTR and MLP regression. The findings claimed that the forecasting results of this model that were measured in the form of RMSE, MAE, MSE and R^2 were more valid than previous models. Ma et al. [15] proposed model used XGBoost which was validated on the datasets of three years (2015-2018). The results of this model were compared with another model termed (WRF-Chem) that proved that the proposed model performed better with higher 50-100% R^2 and lower standard deviation by 14-24ug m3. Joharestani et al. [16] concluded that Gradient boosting gave better performance with R^2 of 0.81, MAE of 09.92 and RMSE 13.58.

Zhang et al. [17] explored that RF Spark cluster consisted of one main node and three worker nodes performed better than traditional methods. Karimian et al. [18] used data of 9 stations for the period of four years (2013-2016) provided by AQCC. The results of this study exhibited that the LSTM model achieved the lowest RSME=8.91 Mgm-3 and MAE=6.21 Mg m-3 and 75% accuracy in forecasting air pollution. Delavar et al. [19] used 24 hours’ data related to pollutants obtained from AQCC and meteorological data provided by IMO for the period of ten years from 2006 to 2016. They applied different machine learning methods called SVM, GVM, ANN, Autoregressive non-linear neural network to predict air pollution. The comparative findings claimed that the NARX method with refined data gave the most accurate results in the prediction of $PM_{2.5}$ and PM_{10} concentrations.

Category # 2: Measurement of Multiple Pollutants

Air pollution has become the most pressing issue around the world. Sethi et al. [20] collected data from OGD (Open Government Data) India consisted of various pollutants namely $PM_{2.5}$, PM_{10} and ammonia and ozone to train the model They compared the accuracy of MLR, RFR, DTR, SVR and XGBoost. The results showed that the RFR model had minimal errors with almost 91.25% accuracy.

The study of Kiftiyani & Nazhifah [21] used the dataset consisted of three years from 2017-2019 related to NO_2, SO_2,

https://doi.org/10.54692/lgurjcsit.2022.0602270

CO, O_3, $PM_{2.5}$ and PM_{10}. They employed three deep learning techniques known as LSTM, CNN, CNN-LSTM and gave results with normalization and without normalization. In the end, it was inferred that CNN gave lower RMSE values of 4.707 without normalization while with normalization CNN-LSTM offered the lowest RMSE value of 7.137.

Sharma et al. [22] and [61-64] developed a model that could predict the concentration of different pollutants (SO_2, NO_2, RSPM, O_3) using historical and present data. They applied six classifiers which included LBR, SVM, RF, DT, KNN and ANN with the help of a dataset labelled as "Air Pollution Geocodes Dataset" containing data from 2016-2018 of 196 Indian cities. The findings suggested that RF had more accuracy than other algorithms.

Juarez & Petersen [23] developed software to analyze the hourly record of 12 air pollution and 5 weather variables per year in Delhi, India. They collected five years (January 2015 to June 2020) hourly pollutant data of Delhi from the CPCB of India. They applied eight machine learning algorithms such as XGBoost, SVR, KNN, DT, LR, RF, Adaboost and LSTM to forecast the next 1 to 24 h ozone concentration level. The result showed that the XGBoost and RF performed better with R^2 of 0.61.

Bhalgat et al. [24] used Linear Regression and Multilayer Perceptron (ANN) protocol for the prediction of next-day pollution by using the dataset comprised of 60383 records of Maharashtra. The ARIMA and AR models were used for predicting the values of SO_2. They concluded that Nagpur has a higher SO_2 level than other cities. Shen et al. [25] employed PFM (Prophet Forecasting Model) model to estimate both short-term and long-term air pollution in Seoul. The results revealed that PFM had the unique potential to predict both short-term and long-term air pollution in terms of climate which other forecasting models fail to address.

Khan et al. [26] forecasted air pollution in the four most polluted areas of Delhi. They collected data of the previous four years (2015-2019) from the website of the CPCB. It consisted of eight pollutants named PM_{10}, $PM_{2.5}$, CO, NO, NO_2, NOx, Ozone and SO_2. After implementing multiple techniques, they concluded that Anand Vihar was the most polluted area of Delhi having the worst AQI.

Kanjo [27] developed a pollution foretelling system. He collected data of one year (1 July 2017 to 1 August 2018) from two cities namely Istanbul and Bursa. The dataset contained different pollutants termed O_3, NO_2, $PM_{2.5}$ and PM_{10}. The data was processed by employing ANN, NARX, and ANFIS. The model was trained and tested using the aforementioned models. It was concluded that the model namely ANFIS offered better performance with training and validation RMSE values of 0.0022, and 0.0038 respectively.

Rubal et al. [28] evaluated the hybrid method in predicting air pollution levels. They used the differential evaluation method with RF to predict the concentration of seven pollutants (C_6H_6, NO_2, O_3, SO_2, CO, $PM_{2.5}$ and PM_{10}) with the connotation of a dataset of Delhi and Patna from 2015-2017 consisted of 946 records. The findings of the proposed method were validated in an experiment in which the proposed method outperformed.

Lepperod [29] executed to predict PM_{10}, $PM_{2.5}$ and NO_2. This dataset was of three types namely traffic data, wood burner data and historical observations of weather data from different stations of the target city. They applied several ensemble techniques known as RR, RF, GB, MP and RNN to forecast air pollution and found out that gradient boosting offered more promising results than other ensemble techniques.

Fu et al. [30] used an air quality prediction model termed as Bayesian network to predict the air quality of Hangzhou. They collected data of Hangzhou from 01 March 2018 to 30 April 2021 from the Zhenqi website. The dataset consisted of six air pollutants named $PM_{2.5}$, PM_{10}, NO_2, SO_2, CO, and O_3 which were used as evaluation factors. The results indicated that air quality prediction accuracy was more than 80%.

Sharma et al. [31] developed a model to forecast AQI including the impurities named $PM_{2.5}$, PM_{10}, O_3, NO_2, and SO_2. and calculated results for 196 cities of India on different classifiers. The performance of the five most accurate results giver classifiers namely SVM, KNN, DT, RF and ANN which discovered that Decision Tree (DT) gave more accurate results with an accuracy of 99.7% which was further maximized by 0.02% with the use of another classifier termed as Random Forest Classifier.

Castelli et al. [32] develop a model using SVR to forecast the hourly Air Quality Index (AQI) for the state of California. They collected hourly data of California from EPA between 01 January 2016 to 01 May 2018 consisted of pollutants named CO, SO_2, NO_2, and $PM_{2.5}$. It was found that SVR with RBF kernel (Redial Basis Function) provided more accurate results in the hourly prediction of pollutant concentrations with 94.1% accuracy. Asgari et al. [33] employed Apache Spark on the Hadoop cluster to boost processing speed. They found out that Logistic Regression demonstrates the best estimator with 0.68 accuracy and Naïve Bayes with 0.48.

Chen et al. [34] collected daily AQIs of 16 large cities of China included three fatal pollutants namely $PM_{2.5}$, PM_{10} and SO. They used PMI based separate IVS scheme for predictors (pollutants) selection and Ensemble Neural Network for prediction. The outcomes proved that the predictability of PBK-based machine learning methods has closely related to quality. Gocheva-Ilieva et al. [35] used time-series data included hourly measurements of air pollutants named O_3, NO_x, NO, CO, SO_2 and PM_{10}. It was concluded that the RF-ARIMA methodology offered the opportunity to develop high-performance models and achieve excellent quality of predicting concentrations of air pollutants.

Bouzoukis et al. [36] also aimed to generalize the findings of their exploration based on the large-scale data collection consisted of eleven stations by applying an ensemble technique consisted of FFNN, CLNN, FIS, SOM, and RF to get the best performance regarding prediction. It was disclosed that the system performed well in predicting the quality of air.

Masmoudi et al. [37] introduced a novel method known as ERCFR which was the amalgamation of two valid approaches termed as Ensemble of Regression Chains and Random Forest. The findings of the study were validated through experiments in which it was noticed that ERCFR perfumed better than other previous approaches but the authors asserted that more research should be conducted to purify the findings. Peng et al. [38] carried out a study that tended to overcome the deficiencies and limitations of previous traditional linear and nonlinear approaches. The data of six stations consisted of O_3, $PM_{2.5}$ and NO_2 from 2009-2014 taken by UMOS-AQ forecast system of "Environment Canada" was used in this investigation. The model was validated by making a comparison between MLRM, OSMLR, MLPNN, and OSELM in which OSELM gave better performance than others.

Liu et al. [39] developed two regression models by using SVR and RFR to predict the air pollution of Beijing and the nitrogen oxides (NO_X) concentration in an Italian city. The experimental results showed that both models achieved good results but the RFR model performed better in the experiments. Ma et al. [40] proposed a non-linear framework to investigate the most important factors of air quality with the perspective of big data on using U.S. counties dataset of four years (2012-2016). They applied Extreme Gradient Boosting (XGBoost) to model the non-linear relationships and measure the importance of features. It was concluded that this methodology uncovered the important factors of air quality skillfully and found six major factors that affected the air quality.

Table 1: Air pollution Forecasting Techniques

Sr. No	Author Name	Year	Algorithms	Findings
1	Sethi et al. [20]	2021	MLR,DTR, SVR, XGBR,RFR	The results showed that the RFR model had minimal errors with almost 97% accuracy.
2	Enebish et al., [8]	2021	RF, GBM, SVM, MARS, GLMNET and GAM	RF and GBM performed better than others using LOLO and CV with R^2 of 0.82.
3	Kiftiyani & Nazhifah [21]	2021	LSTM, CNN, CNN-LSTM	It was inferred that CNN gave lower RMSE values of 4.707 without normalization while with normalization CNN-LSTM offered the lowest RMSE value of 7.137
4	Masood & Ahmad [9]	2019	SVR and ANN	The ANN gave better results regarding the prediction $PM_{2.5.}$
5	Usmani [10]	2019	DT,ANN, SVM,RF,GLM	The findings revealed that Partitioning & Regression Tree was more efficient and accurate in prediction.
6	Bozdag [11]	2020	LASSO, RF, KNN, XGB,ANN	The ANN indicated the best performance with low RMSE and MAE values of 20.8 and 14.43 respectively.
7	Sethi et al. [12]	2019	DT, RF, LR and NN	LR gave promising results and after that CBL was applied on selected features in which RF improved the veracity in predicting $PM_{2.5}$
8	Sharma et al. [22]	2021	LBR, SVM, RF, DT, KNN and ANN	The findings suggested that RF had more accuracy than other algorithms.
9	Juarez & Petersen [23]	2021	XGBoost, SVR, KNN, DT, LR, RF, Adaboost and LSTM	The result showed that the XGBoost and RF performed better with R^2 of 0.61.
10	Sharma et al. [31]	2020	SVM, KNN, DT, RF, ANN	Decision Tree gave more accurate results with an accuracy of 99.7%
11	Castelli et al. [32]	2020	SVR with different kernel	It was found that SVR with RBF kernel provided more accurate results in the hourly prediction of pollutant concentrations with 94.1% accuracy
12	Asgari et al. [33]	2017	Multinomial Naïve Bayes and Multinomial Logistic Regression	It was found out that Logistic Regression demonstrating best estimator with 0.68 accuracy and Naïve Bayes with 0.48
13	Chen et al. [34]	2018	PMI based separate IVS scheme and Ensemble Neural Network	The outcomes proved that the predictability of PBK-based machine learning methods has closely related to quality.
14	Gocheva-Ilieva et al.	2020	ARIMA & RF	It was concluded that the RF-ARIMA achieved excellent

LGURJCSIT Volume 06, Issue 02, Article ID 270, 15 pages
https://doi.org/10.54692/lgurjcsit.2022.0602270

	[35]			quality of predicting concentrations of air pollutants.
15	Lee et al. [13]	2020	GB and extract feature-based machine learning	The proposed method based on gradient boosting demonstrated promising results.
16	Doresawamy et al. [14]	2019	RFR, DTR and MLPR	The findings claimed that the forecasting results of this model that were measured in the form of RMSE, MAE, MSE and R^2 were much valid than previous models.
17	Ma et al. [15]	2020	XGBoost	The proposed model performed better with higher 50-100% R^2 and lower standard deviation by 14-24ug m3.
18	Bhalgat et al. [24]	2019	LR, ANN	They concluded that Nagpur has a higher SO2 level than other cities
19	Shen et al. [25]	2020	PFM model, RMSE, MAE ,MSE and coverage	The results revealed that PFM had the unique potential to predict both short-term and long-term air pollution.
20	Ma et al. [40]	2020	XGBoost	After the experiments, it was concluded that this methodology uncovered the important factors of air quality skillfully.
21	Bouzoukis et al. [36]	2016	FFNN, CLNN, FIS, SOM, and RF And a system namely HISYCOL was developed.	It was disclosed that the system performed well in predicting the quality of air.
22	Masmoudi et al. [37]	2020	Ensemble of Regression Chains and Random Forest.	The findings of the study were validated through experiments in which it was noticed that ERCFR perfumed better than other previous approaches
23	Khan et al. [26]	2019	Multiple Regression Technique	After the experiment, they concluded that Anand Vihar was the most polluted area of Delhi having the worst Air Quality Index
24	Kanjo [27]	2019	ANN, NARX, and ANFIS	It was concluded that the model namely ANFIS offered a better performance with training and validation RMSE values of 0.0022, and 0.0038 respectively.
25	Lepperod [29]	2019	RR, RF, GB MP and RNN	. It was found out that gradient boosting offered more promising results than other ensemble techniques.
26	Joharestani et al [16]	2019	RF, Gradient Boosting	Comparatively, it was concluded that GB gave better performance with R^2 of 0.81, MAE of 09.92 and RMSE 13.58.
27	Peng et al. [38]	2017	MLRM, OSMLR, MLPNN, and OSELM	MLRM, OSMLR, MLPNN, and OSELM in which OSELM gave better performance than others.
28	Zhang et al. [17]	2016	random forest algorithm on Spark cluster	It was noted that Spark based method performed better in predicting the level of $PM_{2.5}$ in real-time.
29	Rubal et al. [28]	2018	RF	The findings of the proposed method were validated in an experiment in which the proposed method outperformed.
30	Liu et al. [39]	2019	SVR and RFR	The experimental results showed that both models achieved good results but the RFR model performed better in the experiments.
31	Fu et al. [30]	2021	Bayesian network model	The results indicated that air quality prediction accuracy was more than 80%.
32	Karimian et al. [18]	2019	MART, DFNN approaches and a hybrid (LSTM)	The results of this study exhibited that LSTM model achieved the lowest RSME = 8.91 and MAE=6.21 and 75% accuracy
33	Delavar et al. [19]	2019	SVM, GVM, ANN, Autoregressive non-linear NN	The comparative findings claimed that the NARX method gave the most accurate results in the prediction of $PM_{2.5}$ and PM_{10} concentrations.

3. Methodology

This section interprets the details of the dataset, attributes, pollutants, detail of their creation and their adverse impacts on humans and the environment. In addition, the detail of used algorithms and their application in the arena of machine learning has also been debated.

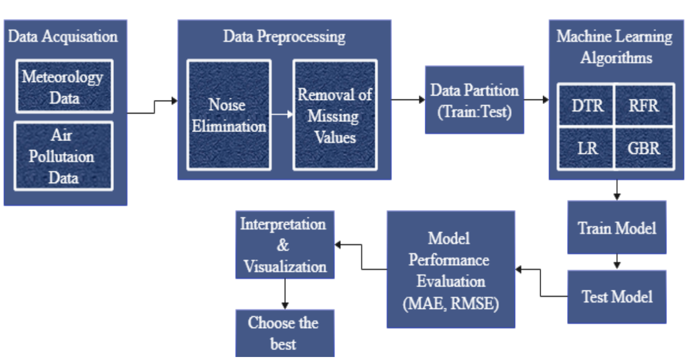
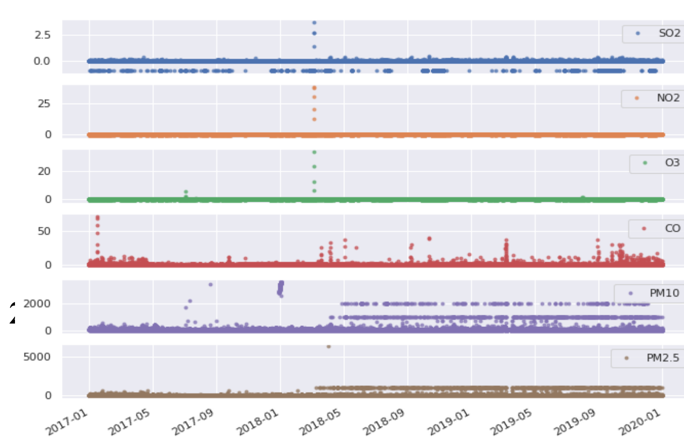


Figure 2. The proposed architecture model for predicting air



pollution

3.1. Dataset Description

The dataset that was utilized in this probe was concerned with the capital city of South Korea termed Seoul is an open data which was downloaded from the website of Kaggle and Seoul Metropolitan Government public data. It possessed six distinct pollutants which included SO2, NO2, O3, CO, PM2.5, PM10. It was collected from twenty-five different stations in which values of all pollutants were measured hourly for three years (2017-19). It embodied 647,512 instances and the following 11 attributes namely Measurement date, Station code, Latitude, Longitude, SO2, NO2, O3, CO, PM10 and PM2.5.

3.2 Data Splitting

In this study, the dataset of Seoul consists of three years (2017-19) has been used. It was split into two parts for training and testing the model. Training is the process where the model is trained to get the required outcome for a particular purpose. The part of the dataset that is selected to train the model generally consisted of a large amount of data in the dataset. In this study, 80% proportion of the dataset was selected for training the data on different regression techniques. After that remaining 20% of the data of the dataset has been used for testing

3.3 Data Processing

All the processing related to data was carried out in the operating system namely the Window-10 i5 machine. Python Programming Language was used for data development. Pandas was utilized to perform preprocessing relevant to time series evolution while machine learning algorithms were executed using a library called scikit to learn library that is an open-source ML library for the purpose of python programming language. While plotting of graph was carried out in plotly library. Sklearn metrics was adopted for evolution purpose and all the code were written on Google Colab. After that null values of the dataset were removed and experiments were made employing the four different algorithms known as DTR, LR, RFR and GBR. To evaluate the performance of all regression algorithms in the prediction of each pollutant concentration the evaluation metrics MAE and RMSE were used.

Figure 3. Hourly data distribution of each air pollutant in Seoul

3.4 Estimation Model/Regression Techniques

3.4.1 Random Forest

Machine learning is the most important branch in the domain of artificial intelligence in which a variety of algorithms are employed to execute unique assignments but the accuracy of result and time of execution was not up to the mark while using traditional algorithms. Random Forest has a lot of potentials such as classification accuracy, ability to cope with outliers and noise and lack of overfitting. RF has been some of the most widely used research approaches in the field of data mining and machine learning, and information to the field of biology [41]. Random Forest can cope with micro information data. In addition, its accuracy of results is much higher than other various algorithms. [42], [64-70].

3.4.2 Linear Regression

The model termed linear regression is linked with supervised machine learning. It is one of the unique models in the domain of data analysis which is mainly used for the motives of prediction. It is contemplated easier and more famous algorithms than other machine learning algorithms. The model reveals the best fit linear line between variables termed dependent and independent variable where "X" is considered independent while "Y" is assumed an independent variable [43][97-100]. The general form of the equation of multiple linear regression is:

$y = \beta 0 + \beta 1x1 + \beta 2x2 + \cdots \beta kxk$

Multiple linear regression models described how a single response variable Y is linearly dependent on a number of predictor variables.

3.4.3 Decision Tree

The decision tree is another important technique that is adopted to solve the problems that come in the orbit of classification and regression. It is an algorithm where decisions are leaves and data is split in the nodes. There are many advantages of using this algorithm such as data can be handled handily, data can be interpreted with considerable comfort [101]. It gives real-time solutions by anticipating the solution of some problem. All types of values like categorical and quantitative are susceptible to handle where missing values are replaced with the most suitable ones. But the decision tree may encounter the problem of overfitting that can be unravelled by employing random forest. [44], [77-84].

3.4.4 XGboost Algorithm

This algorithm is an advanced form of gradients boosting [45]. It is a highly appreciated and adored algorithm due to its best performance in solving problems about classification, ranking and regression. In addition, its execution speed is very high and gives real-time solutions. So, data analysts adore this algorithm [46]. The reason for accurate results given by this algorithm is that it produces results in the form of a tree structure with the parallel approach by remembering in mind the specification and configuration of the model [42]. It can produce state-of-the-art outcomes with minimum sources [47], [71-77].

3.5 Evaluation Criteria

It is a criterion that is used to evaluate the performance of the model. Many statistical techniques are used for evaluation. In this study, Root Mean Square Error (RMSE) and Absolute error (MAE) have been used to know the performance of the model.

3.5.1 Mean Absolute Error (MAE)

Mean Absolute Error is the standard that measures the average intensity of errors in a set of predictions values, regardless of direction [48]. It is the average of the absolute differences between actual and predicted values. It is calculated as in the equation.

	Station code	Latitude	Longitude	SO2	NO2	O3	CO	PM10	PM2.5
count	647511.000000	647511.000000	647511.000000	647511.000000	647511.000000	647511.000000	647511.000000	647511.000000	647511.000000
mean	113.000221	37.553484	126.989340	-0.001795	0.022519	0.017979	0.509197	43.708051	25.411995
std	7.211315	0.053273	0.078790	0.078832	0.115153	0.099308	0.405319	71.137342	43.924595
min	101.000000	37.452357	126.835151	-1.000000	-1.000000	-1.000000	-1.000000	-1.000000	-1.000000
25%	107.000000	37.517528	126.927102	0.003000	0.016000	0.008000	0.300000	22.000000	11.000000
50%	113.000000	37.544962	127.004850	0.004000	0.025000	0.021000	0.500000	35.000000	19.000000
75%	119.000000	37.584848	127.047470	0.005000	0.038000	0.034000	0.600000	53.000000	31.000000
max	125.000000	37.658774	127.136792	3.736000	38.445000	33.600000	71.700000	3586.000000	6256.000000

$$MAE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \acute{y}_i)$$

Where,
n = Number of observations
y_i = Actual Values
$\acute{y}_i$ = Predicted Values

3.5.2 Root Mean Square Error (RMSE)

Root mean square error is a standard deviation of the prediction errors. It is also used to measure the model performance [48] , [84-92]. It is obtained by taking the average of squared differences between actual and predicted values and then taking the square root of final results. It is calculated as in the equation.

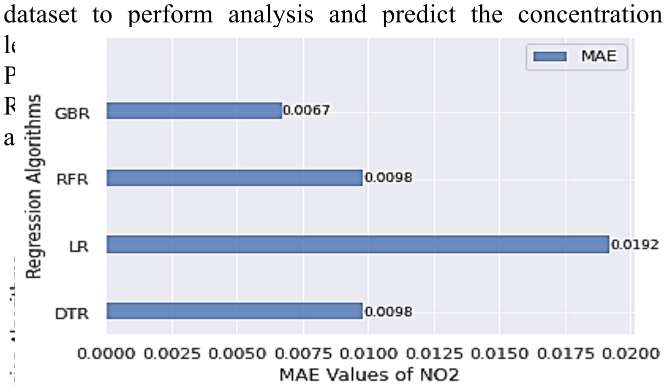
$$RMSE = \sqrt{\frac{1}{n}(\sum_{i=1}^{n}(y_i \; - \; \acute{y}_i\,)^2)}$$

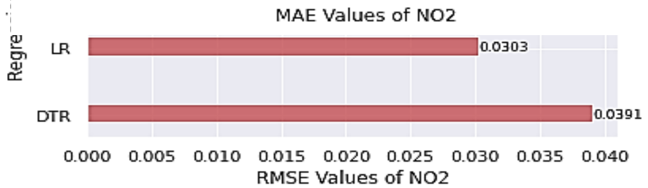
Table 2: Dataset Descriptive Statistics

Table 2 shows the count, mean and standard deviation of each pollutant. Among all pollutants, PM10 recorded the highest of 71.14 and 43.71 and SO2 recorded the lowest of -0.002 and 0.079 mean and standard deviation respectively.

4.1 Results

Different regression techniques were applied to Seoul's dataset to perform analysis and predict the concentration





4. Result and Discussion

In this section, the result of four prominent algorithms in the domain of machine learning called GBR, RFR, LR, DTR has been demonstrated in the form of tables. To make a prediction, RMSE and MAE have been measured and a comparison has been made to get the most accurate algorithm in predicting air pollution related to six pollutants namely SO2, NO2, CO, O3, PM10 and PM2.5.

The predicted values of each pollutant vs. actual results have been indicated by line graph which is considered a good visual technique for estimating the goodness of the regression model at a glance. Time is taken on the x-axis, while predictive values of various regression algorithms have been demonstrated on the y-axis.

4.1.1 Nitrogen Dioxide (NO2)

(a): *MAE for different regression techniques,*

LGURJCSIT Volume 06, Issue 02, Article ID 270, 15 pages
https://doi.org/10.54692/lgurjcsit.2022.0602270

(b): RMSE for different regression techniques

Figure 4. Different regression techniques

Figure 4 demonstrates that GBR has a lower MAE and RMSE value than other algorithms. GBR has MAE 0.0111 while RMSE achieved was 0. 0067 which is lower than MAE and RMSE values of RFR, LR and DTR. Hence**,** it is evident that GBR is better than others algorithms in forecasting the concentration level of pollutants named NO_2.

4.1.2 Sulfur Dioxide (SO_2)

Predictive analysis has been for Seoul's dataset using numerous regression techniques in which each air pollutant was predicted separately. The MAE and RMSE values are shown in Figure 4.2.

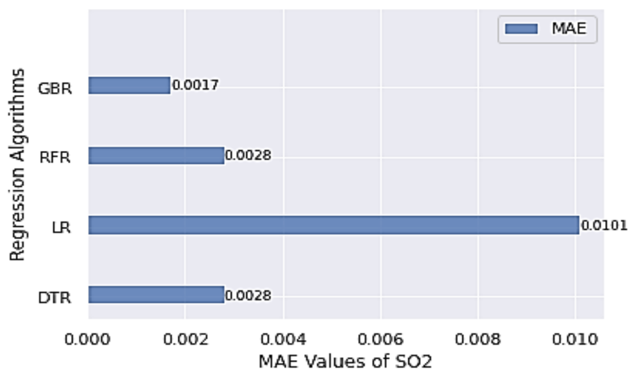


Figure 5 (a): MAE for different regression techniques

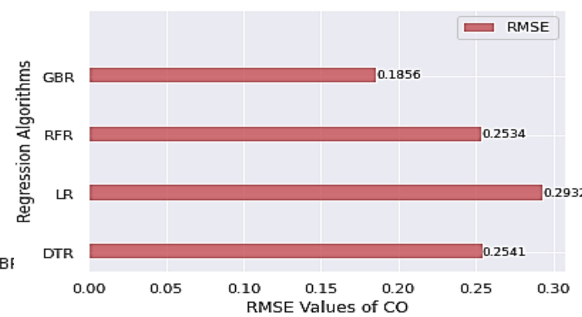
Figure 5 (b): RMSE for different regression techniques

Figure 5 (a) Above results are reflecting that GBR, RFR and DTR have suggested near about the same MAE values of 0.0017,0.0028 and 0.0028 respectively. RMSE values have been computed and displayed in Fig 5 (b). RFR, LR and DTR have expressed almost identical values but GBR has a lower MAE and RMSE value as compared to the remaining algorithms. So, it is apparent that GBR is more promising than others algorithms in anticipating the concentration level of pollutants called SO_2 as GBR has revealed MAE value of 0.0017 and RMSE value of 0.0168 as compared to MAE and RMSE values of other models.

4.1.3 Carbon monoxide (CO)

Predictive Analysis has been conducted for Seoul's dataset using different regression techniques in which each air pollutant was foreseen separately. MAE and RMSE values are exhibited in Figure 4.3.

Figure 6 (a): MAE for different regression techniques



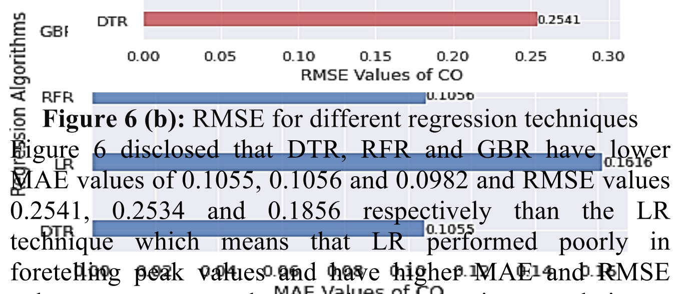


Figure 6 (b): RMSE for different regression techniques

Figure 6 disclosed that DTR, RFR and GBR have lower MAE values of 0.1055, 0.1056 and 0.0982 and RMSE values 0.2541, 0.2534 and 0.1856 respectively than the LR technique which means that LR performed poorly in foretelling peak values and have higher MAE and RMSE values as compared to other regressions techniques. Therefore, it is obvious that GBR is more favourable because it headlined the lowest MAE and RMSE values than others algorithms in predicting the concentration level of pollutants termed CO.

https://doi.org/10.54692/lgurjcsit.2022.0602270

4.1.4 Particular Matter ($PM_{2.5}$)

The four regression techniques were applied on Seoul's dataset to perform analysis and maximum and minimum values have been anticipated and compared. In Figure 4.4(a), LR has expressed the MAE value of 12.15 which is high as compared to other regression techniques that have poor results in predictive analysis to foresee peak values. Comparatively, DTR and RFR performed much better than LR but GBR denoted better output with MAE value of 7.63.

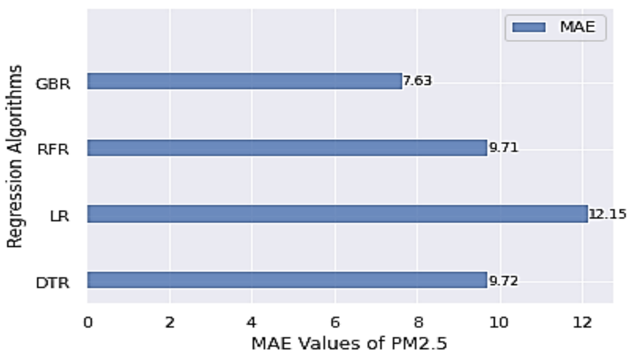
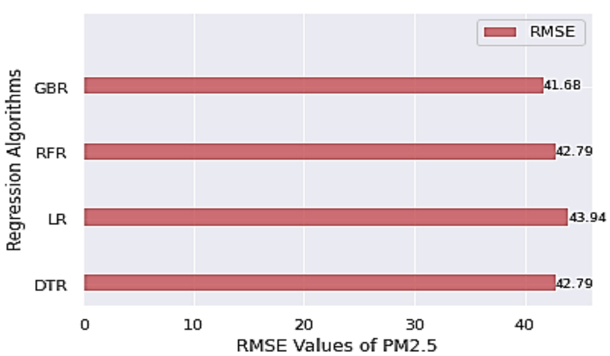
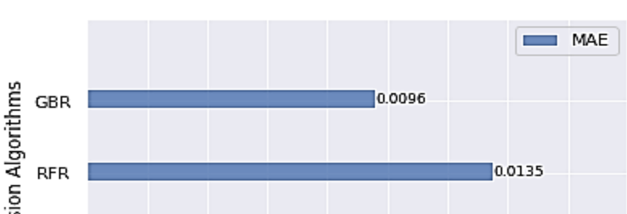


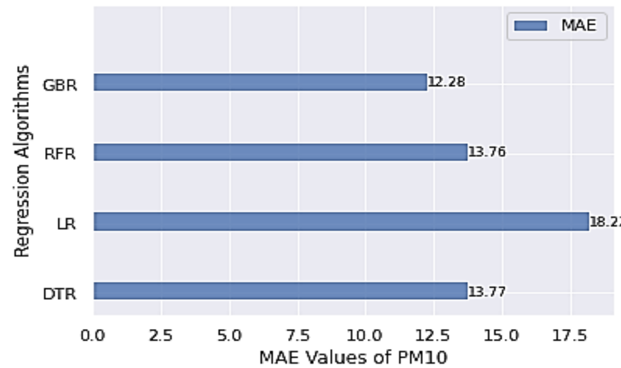
Figure 7 (a): MAE for different regression techniques

Figure 7 (b): RMSE for different regression techniques

Figure 7 (b) denotes that all regression techniques GBR, RFR, LR and DTR have almost identical RMSE values of 41.68, 42.79, 43.94 and 42.79 respectively. LR displayed poor results in foretelling but RFR and DTR remained almost the same. GBR has a lower MAE and RMSE value than other algorithms. It obvious that GBR with RMSE value of 41.68 is more decent than others algorithms in anticipating the concentration level of pollutants named $PM_{2.5}$.





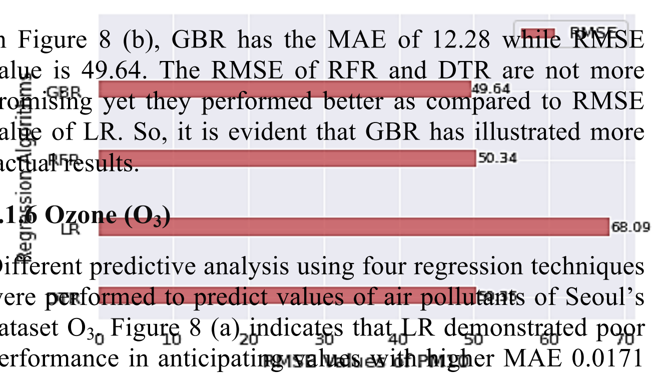


4.1.5 Particular Matter (PM_{10})

The regression analysis was performed utilizing different regression techniques to forecast values for each pollutant of Seoul's dataset. Figure 4.5(a) indicates the MAE values of PM_{10}. Figure 4.5(a) demonstrates that LR has higher MAE values of 18.22. DTR and RFR remain the same with 13.77 and 13.76 respectively. Comparatively, GBR expressed good performance with lower MAE value of 12.28

Figure 8 (a): MAE for different regression techniques

Figure 8 (b): RMSE for different regression techniques



In Figure 8 (b), GBR has the MAE of 12.28 while RMSE value is 49.64. The RMSE of RFR and DTR are not more promising yet they performed better as compared to RMSE value of LR. So, it is evident that GBR has illustrated more factual results.

4.1.6 Ozone (O_3)

Different predictive analysis using four regression techniques were performed to predict values of air pollutants of Seoul's dataset O_3. Figure 8 (a) indicates that LR demonstrated poor performance in anticipating values with higher MAE 0.0171 as compared to other regression techniques. RFR and DTR gave almost similar MAE values of 0.0135 and 0.0134 respectively. So, it is evident that GBR technique gave better results in prediction with a 0.0096 MAE value.

LGURJCSIT Volume 06, Issue 02, Article ID 270, 15 pages
https://doi.org/10.54692/lgurjcsit.2022.0602270

Figure 9 (a): MAE for different regression techniques

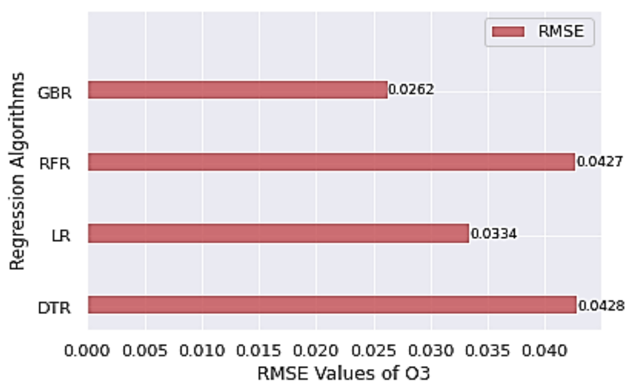


Figure 9 (b): RMSE for different regression techniques

Figure 9 (b) is reflecting that DTR and RFR displayed RMSE values of 0.0427 and 0.0428 respectively was higher compared to the other two regression techniques. Comparatively, GBR and LR have performed much better. GBR has MAE 0.0096 while RMSE value is 0. 0262. So**,** it is clear that GBR is better than other algorithms in predicting the concentration level of O_3 .

4.2 Discussion

In this segment, the findings of the current investigation have been compared with other previous similar inquiries. In addition, the research questions of this research have also been addressed after conducting experiments and exhibitions of results in the form of graphs and tables. Four Algorithms have been run on the dataset consisting of six pollutants namely SO_2, NO_2, CO, O_3, $PM_{2.5}$ and PM_{10}. The results of this study have been compared with another similar study that was conducted on the same dataset termed Seoul by Shen et al. [25]. The result of this investigation is more accurate than the results of a study by [25] because the MAE values of this study are lower than the MAE values of [25] *7.6279, 12.285, 0.00666, 0.00964, 0.001715 and 0.0982* MAE of the current inquiry for *$PM_{2.5}$, PM_{10}, O_3, NO_2, SO_2 and CO* respectively while the Mean Squire Error of [25] was *16.8, 20.72, 0.0134, 0.0129, 0.00241 and 0.387 for $PM_{2.5}$, PM_{10}, O_3, NO_2, SO_2 and CO respectively*. So, it is obvious that the results of this inquiry are more accurate than the investigation of Shen et al [25].

Pollutant	MAE (PFM)	MAE (Current Study)
$PM_{2.5}$	16.8	7.6279
PM_{10}	20.72	12.285
O_3	0.0134	0.00666
NO_2	0.0129	0.00964
SO_2	0.00241	0.001715
CO	0.387	0.0982

Table 3: Comparison of MAE values

It has been found out that "Random Forest" (RF) has been commonly used to predict the concentration level of various pollutants in the different parts of the world such as the investigations of [28][8][11][12][14][16][17][20][22][23][29][31] [35][36][37][39] and the second most commonly used algorithm is "Artificial Neutral Network (ANN)" such as the studies like [9][10] [11] [12] [18] [19] [21] [22] [23][24][27][29][31][36][38] and the third most adopted algorithms is XGBR like in the studies of [8] [11] [13] [15] [16] [23][26][20][40].

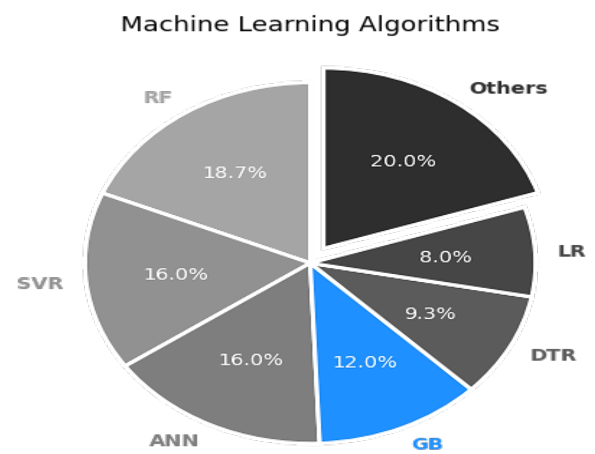


Figure 10. Proportions of machine learning algorithms

Figure 10 reveals that the percentage of use of RF in the earlier studies is 18.7% which is more than any other algorithm. In addition, it is also obvious that the use of ANN and SVM is equal that is 16%. While the third most used algorithm is GB whose percentage is 12%. DTR and LR are at the fourth and fifth numbers respectively. It was inferred that XGB is more promising and feasible in predicting concentration level of NO_2, O_3, SO_2, PM_{10}, $PM_{2.5}$ and CO with the lowest RMSE and MAE values of 0.0111, 0.0262, 0.0168, 49.64, 41.68 and 0.1856 and 0.0067, 0.0096, 0.0017, 12.28, 7.63 and 0.0982 respectively. However, it is noted that while predicting particulate matter, all the algorithms gave higher RMSE and MAE values as compared to RMSE and MAE values of other pollutants and RMSE and MAE values of other pollutants are much lower than the values of particulate matter.

LGURJCSIT Volume 06, Issue 02, Article ID 270, 15 pages
https://doi.org/10.54692/lgurjcsit.2022.0602270

Table 4. Overall performance (MAE) value result on Seoul

Algorithms	NO_2	O_3	SO_2	PM_{10}	$PM_{2.5}$	CO
RFR	0.0098	0.0135	0.0028	13.76	9.72	0.1056
DTR	0.0098	0.0134	0.0028	13.77	9.72	0.1055
GBR	0.0067	0.0096	0.0017	12.28	7.63	0.0982
LR	0.0192	0.0171	0.0102	18.22	12.15	0.1616

Table 4 shows the overall performance through Mean Absolute Error (MAE) values for each model and scenario. It can be seen that GBR has lower MAE values and gave better performance than other models in the prediction of all pollutants concentration levels.

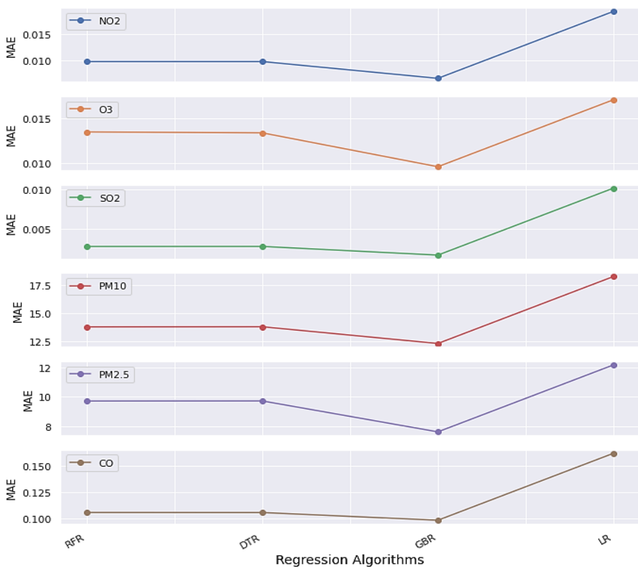


Figure 11: *A comparison of MAE values for each pollution component*

Table 5: Overall performance (RMSE) value result on Seoul pollution dataset

Algorithms	NO_2	O_3	SO_2	PM_{10}	$PM_{2.5}$	CO
RFR	0.0389	0.0427	0.0343	50.34	42.79	0.2534
DTR	0.0391	0.0428	0.0342	50.35	42.79	0.2541
GBR	0.0111	0.0262	0.0168	49.64	41.68	0.1856
LR	0.0303	0.0334	0.0347	64.08	43.94	0.2932

Table 4.4 displays the overall performance through Root Mean Square Error (RMSE) values for each model and scenario. It can be seen that GBR has lower RMSE values and gave better performance than other models in the prediction of all pollutant's concentration levels.

pollution dataset

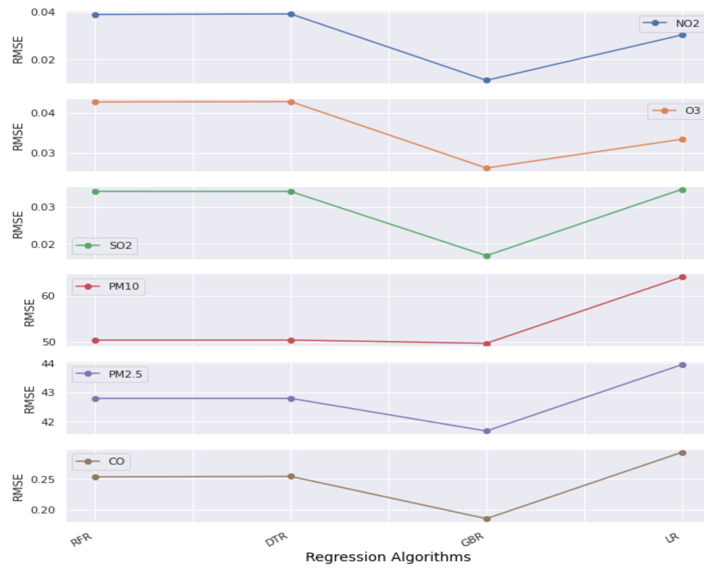


Figure 12: A comparison of RMSE values for each pollution component

The short- and long-term exposure had different but dangerous effects on the health of all the organisms that exist on the earth such as different diseases of the lungs and respiratory system. Short-term effects of nitrogen dioxide include premature cardiovascular disease [49] while the same short-term effect of ozone is debated [50]. Similarly, traffic on the roads and particulate matter also causes the disease named premature cardiovascular disease [51]. Failure of the heart is another effect due to different pollutants such as CO, NO_2 and SO_2 [52].

Furthermore, short-term effects of another other deadly pollutant called particulate matter include coronary syndrome [53][54]. The data of 188 different countries of the world claimed that there is a strong link between stroke and air pollution [55]. It was also identified that hourly increase of pollutants also has fatal effects on living creatures[51][56]. Some studies in the arena of science asserted that pollution caused by road traffic causes hypertension [57][58]. Sudden death, stress, anxiety, tension, inflammation and other psychological effects are also the result of long-term exposure to contaminated air [59][60].

5. Conclusion

It is concluded that air pollution is very toxic to human health. it is pointed out that mostly Random Forest, Gradient Boosting had been used in the world to develop systems to foretell the concentration level of pollutants. In addition, it is also observed that Gradient Boosting and Random Forest offered more accuracy than other algorithms. However, it is evident that the accuracy of a system might change due to changes in parameters and changes in climate in different parts of the world. In the present study, four algorithms termed Random Forest Regression, Linear Regression, Decision Tree Regression and Gradient Boosting Regression were employed on a large dataset of Seoul to proposed an architecture design to predict the level of pollutants accurately. After the experiments, it was inferred that extreme gradient boosting offered more promising and vowing accuracy in predicting the air pollution that is the biggest enemy of healthy living.

6. Future Work

This investigation has been carried out with four algorithms and a dataset having ten parameters in the domain of machine learning. So, it is intended that different deep learning algorithms would be employed with a dataset having more and different parameters.

References

[1]. M. S. Anjum et al., "An Emerged Challenge of Air Pollution and Ever-Increasing Particulate Matter in Pakistan; A Critical Review," J. Hazard. Mater., vol. 402, no. August 2020, p. 123943, 2021, doi: 10.1016/j.jhazmat.2020.123943.

[2]. C. R. Jung, B. F. Hwang, and W. T. Chen, "Incorporating long-term satellite-based aerosol optical depth, localized land use data, and meteorological variables to estimate ground-level PM2.5 concentrations in Taiwan from 2005 to 2015," Environ. Pollut., vol. 237, pp. 1000–1010, 2017, doi: 10.1016/j.envpol.2017.11.016.

[3]. P. W. Soh, K. H. Chen, J. W. Huang, and H. J. Chu, "Spatial-Temporal pattern analysis and prediction of air quality in Taiwan," Ubi-Media 2017 - Proc. 10th Int. Conf. Ubi-Media Comput. Work. with 4th Int. Work. Adv. E-Learning 1st Int. Work. Multimed. IoT Networks, Syst. Appl., pp. 3–8, 2017, doi: 10.1109/UMEDIA.2017.8074094.

[4]. N. Mehrabi, F. Morstatter, N. Saxena, K. Lerman, and A. Galstyan, "A Survey on Bias and Fairness in Machine Learning," ACM Comput. Surv., vol. 54, no. 6, 2021, doi: 10.1145/3457607.

[5]. J. Liu, B. Wang, and L. Xiao, "Non-linear associations between built environment and active travel for working and shopping: An extreme gradient boosting approach," J. Transp. Geogr., vol. 92, no. November 2020, p. 103034, 2021, doi: 10.1016/j.jtrangeo.2021.103034.

[6]. W. Fan, F. Si, S. Ren, C. Yu, Y. Cui, and P. Wang, "Integration of continuous restricted Boltzmann machine and SVR in NOx emissions prediction of a tangential firing boiler," Chemom. Intell. Lab. Syst., vol. 195, 2019, doi: 10.1016/j.chemolab.2019.103870.

[7]. S. Ameer et al., "Comparative Analysis of Machine Learning Techniques for Predicting Air Quality in Smart Cities," IEEE Access, vol. 7, pp. 128325–128338, 2019, doi: 10.1109/ACCESS.2019.2925082.

[8]. T. Enebish, "Predicting ambient PM 2 . 5 concentrations in Ulaanbaatar , Mongolia with machine learning approaches," J. Expo. Sci. Environ. Epidemiol., 2020, doi: 10.1038/s41370-020-0257-8.

[9]. A. Masood and K. Ahmad, "A model for particulate matter (PM2.5) prediction for Delhi based on machine learning approaches," Procedia Comput. Sci., vol. 167, no. 2019, pp. 2101–2110, 2020, doi: 10.1016/j.procs.2020.03.258.

[10]. S. Usmani, "Data Mining & Machine Learning Algorithms for Air Pollutant Prediction," Artif. Comput. Intell., vol. 1, no. 1, pp. 1–7, 2019.

[11]. A. Bozdağ, Y. Dokuz, and Ö. B. Gökçek, "Spatial prediction of PM10 concentration using machine learning algorithms in Ankara, Turkey," Environ. Pollut., vol. 263, 2020, doi: 10.1016/j.envpol.2020.114635.

[12]. J. K. Sethi and M. Mittal, "A new feature selection method based on machine learning technique for air quality dataset," vol. 0510, 2019, doi: 10.1080/09720510.2019.1609726.

[13]. M. Lee et al., "Forecasting Air Quality in Taiwan by Using Machine Learning," Sci. Rep., vol. 10, no. 1, pp. 1–13, 2020, doi: 10.1038/s41598-020-61151-7.

[14]. Doreswamy, Y. Km, I. Gad, K. S. Harishkumar, Y. Km, and I. Gad, "Forecasting Air Air Pollution Pollution Particulate Matter (PM2.5) Using Machine Learning Regression Models," Procedia Comput. Sci., vol. 171, no. 2019, pp. 2057–2066, 2020, doi: 10.1016/j.procs.2020.04.221.

[15]. J. Ma, Z. Yu, Y. Qu, and Y. Cao, "Application of the XGBoost Machine Learning Method in PM 2 . 5 Prediction : A Case Study of Shanghai," pp. 128–138, 2020, doi: 10.4209/aaqr.2019.08.0408.

[16]. M. Z. Joharestani, C. Cao, X. Ni, B. Bashir, and S. Talebiesfandarani, "PM2.5 Prediction Based on Random Forest, XGBoost, and Deep Learning Using Multisource Remote Sensing Data Mehdi," Atmosphere (Basel)., vol. 10, no. 7, p. 373, 2019.

[17]. C. Zhang and D. Yuan, "Fast fine-grained air quality index level prediction using random forest algorithm on cluster computing of spark," Proc. - 2015 IEEE 12th Int. Conf. Ubiquitous Intell. Comput. 2015 IEEE 12th Int. Conf. Adv. Trust. Comput. 2015 IEEE 15th Int. Conf. Scalable Comput. Commun. 20, pp. 929–934, 2016, doi: 10.1109/UIC-ATC-ScalCom-CBDCom-IoP.2015.177.

[18]. H. Karimian et al., "Evaluation of Different Machine Learning Approaches to Forecasting PM 2 . 5 Mass Concentrations," pp. 1400–1410, 2019, doi: 10.4209/aaqr.2018.12.0450.

[19]. M. R. Delavar, A. Gholami, G. R. Shiran, and Y. Rashidi, "A Novel Method for Improving Air Pollution Prediction Based on Machine Learning Approaches : A Case Study Applied to the Capital City of Tehran," ISPRS Int. J. Geo-Inf., 2019, doi: 10.3390/ijgi8020099.

[20]. T. Sethi and R. C. Thakur, "Comparison of Machine Learning Algorithms for Air Pollution Monitoring System," AI IoT□ Based Intell. Autom. Robot., pp. 305–322, 2021, doi: 10.1002/9781119711230.ch19.

[21]. U. Kiftiyani and S. A. Nazhifah, "Deep Learning Models for Air Pollution Forcasting in Seoul South Korea," Int. Conf. Softw. Eng. Comput. Syst. 4th Int. Conf. Comput. Sci. Inf. Manag., pp. 547–551, 2021, doi: 10.1109/ICSECS52883.2021.00106.

[22]. M. Sharma, S. Jain, S. Mittal, and T. H. Sheikh, "Forecasting And Prediction Of Air Pollutants Concentrates Using Machine Learning Techniques : The Case Of India CONCENTRATES USING MACHINE LEARNING TECHNIQUES : THE CASE," 2021, doi: 10.1088/1757-899X/1022/1/012123.

[23]. E. K. Juarez, "A Comparison of Machine Learning Methods to Forecast Tropospheric Ozone Levels in Delhi," pp. 1–26, 2022.

[24]. P. Bhalgat, S. Pitale, and S. Bhoithe, "Air Quality Prediction

using Machine Learning Algorithms," Int. J. Comput. Appl. Technol. Res., vol. 8, no. 9, pp. 367–370, 2019, doi: 10.7753/ijcatr0809.1010.
[25]. J. Shen, D. Valagolam, and S. McCalla, "Prophet forecasting model: A machine learning approach to predict the concentration of air pollutants (PM2.5, PM10, O3, NO2, SO2, CO) in Seoul, South Korea," PeerJ, vol. 8, no. 2, 2020, doi: 10.7717/peerj.9961.
[26]. F. Khan and K. Babar Ali, "AIR POLLUTION FORCASTING SYSTEM USING MACHINE LEARNING," J. Inf. Comput. Sci., vol. 9, no. 10, pp. 322–328, 2019.
[27]. M. F. Kanjo, "INTELIGENT SYSTEM FOR AIR POLLUTION," 2019.
[28]. Rubal and D. Kumar, "ScienceDirect ScienceDirect Evolving Differential evolution method with random forest for Evolving Differential evolution method with random forest for prediction of Air Pollution prediction of Air Pollution," Procedia Comput. Sci., vol. 132, pp. 824–833, 2018, doi: 10.1016/j.procs.2018.05.094.
[29]. A. J. Lepperod, "Air Quality Prediction with Machine Learning," no. June, 2019.
[30]. Z. Fu, H. Lin, B. Huang, and J. Yao, "Research on air quality prediction method in Hangzhou based on machine learning," J. Phys. Conf. Ser., vol. 2010, no. 1, 2021, doi: 10.1088/1742-6596/2010/1/012011.
[31]. M. Sharma, S. Jain, S. Mittal, T. H. Sheikh, and C. Science, "FORECASTING AND PREDICTION OF AIR POLLUTANTS CONCENTRATES USING MACHINE LEARNING TECHNIQUES : THE CASE OF INDIA," 2020.
[32]. M. Castelli, F. M. Clemente, A. Popovič, S. Silva, and L. Vanneschi, "A Machine Learning Approach to Predict Air Quality in California," Complexity, vol. 2020, no. Ml, 2020, doi: 10.1155/2020/8049504.
[33]. M. Asgari, M. Farnaghi, and Z. Ghaemi, "Predictive mapping of urban air pollution using apache spark on a hadoop cluster," ACM Int. Conf. Proceeding Ser., pp. 89–93, 2017, doi: 10.1145/3141128.3141131.
[34]. S. Chen, G. Kan, J. Li, K. Liang, and Y. Hong, "Investigating China's urban air quality using big data, information theory, and machine learning," Polish J. Environ. Stud., vol. 27, no. 2, pp. 565–578, 2018, doi: 10.15244/pjoes/75159.
[35]. S. G. Gocheva-ilieva, A. V Ivanov, and I. E. Livieris, "High Performance Machine Learning Models of Large Scale Air Pollution Data in Urban Area," vol. 20, no. 6, pp. 49–60, 2020, doi: 10.2478/cait-2020-0060.
[36]. I. Bougoudis, K. Demertzis, and L. Iliadis, "HISYCOL a hybrid computational intelligence system for combined machine learning: the case of air pollution modeling in Athens," Neural Comput. Appl., vol. 27, no. 5, pp. 1191–1206, 2016, doi: 10.1007/s00521-015-1927-7.
[37]. S. Masmoudi, H. Elghazel, D. Taieb, O. Yazar, and A. Kallel, "A machine-learning framework for predicting multiple air pollutants' concentrations via multi-target regression and feature selection," Sci. Total Environ., vol. 715, p. 136991, 2020, doi: 10.1016/j.scitotenv.2020.136991.
[38]. H. Peng, A. R. Lima, A. Teakles, J. Jin, A. J. Cannon, and W. W. Hsieh, "Evaluating hourly air quality forecasting in Canada with nonlinear updatable machine learning methods," Air Qual. Atmos. Heal., vol. 10, no. 2, pp. 195–211, 2017, doi: 10.1007/s11869-016-0414-3.
[39]. H. Liu, Q. Li, D. Yu, and Y. Gu, "Air quality index and air pollutant concentration prediction based on machine learning algorithms," Appl. Sci., vol. 9, no. 19, 2019, doi: 10.3390/app9194069.
[40]. J. Ma et al., "Identification of high impact factors of air quality on a national scale using big data and machine learning techniques," J. Clean. Prod., vol. 244, no. xxxx, p. 118955, 2020, doi: 10.1016/j.jclepro.2019.118955.
[41]. L. Breiman, "Random Forests," Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics), vol. 12343 LNCS, pp. 503–515, 2001, doi: 10.1007/978-3-030-62008-0_35.
[42]. A. Bozdağ, Y. Dokuz, and Ö. B. Gökçek, "Spatial prediction of PM10 concentration using machine learning algorithms in Ankara, Turkey," Environ. Pollut., vol. 263, 2020, doi: 10.1016/j.envpol.2020.114635.
[43]. S. Ghosal, S. Sengupta, M. Majumder, and B. Sinha, "Prediction of the number of deaths in India due to SARS-CoV-2 at 5–6 weeks," Diabetes Metab. Syndr. Clin. Res. Rev., vol. 14, no. 4, pp. 311–315, 2020, doi: 10.1016/j.dsx.2020.03.017.
[44]. S. Ray, "A Quick Review of Machine Learning Algorithms," Proc. Int. Conf. Mach. Learn. Big Data, Cloud Parallel Comput. Trends, Prespectives Prospect. Com. 2019, pp. 35–39, 2019, doi: 10.1109/COMITCon.2019.8862451.
[45]. I. L. Cherif and A. Kortebi, "On using eXtreme Gradient Boosting (XGBoost) Machine Learning algorithm for Home Network Traffic Classification," IFIP Wirel. Days, vol. 2019-April, pp. 1–6, 2019, doi: 10.1109/WD.2019.8734193.
[46]. A. Ibrahem Ahmed Osman, A. Najah Ahmed, M. F. Chow, Y. Feng Huang, and A. El-Shafie, "Extreme gradient boosting (Xgboost) model to predict the groundwater levels in Selangor Malaysia," Ain Shams Eng. J., vol. 12, no. 2, pp. 1545–1556, 2021, doi: 10.1016/j.asej.2020.11.011.
[47]. T. Chen and C. Guestrin, "XGBoost: A Scalable Tree Boosting System," J. Assoc. Physicians India, vol. 42, no. 8, p. 665, 2016.
[48]. T. Chai and R. R. Draxler, "Root mean square error (RMSE) or mean absolute error (MAE)? -Arguments against avoiding RMSE in the literature," Geosci. Model Dev., vol. 7, no. 3, pp. 1247–1250, 2014, doi: 10.5194/gmd-7-1247-2014.
[49]. I. C. Mills, R. W. Atkinson, S. Kang, H. Walton, and H. R. Anderson, "Quantitative systematic review of the associations between short-term exposure to nitrogen dioxide and mortality and hospital admissions," BMJ Open, vol. 5, no. 5, 2015, doi: 10.1136/bmjopen-2014-006946.
[50]. M. Jerrett et al., "Long-Term Ozone Exposure and Mortality," N. Engl. J. Med., vol. 360, no. 11, pp. 1085–1095, 2009, doi: 10.1056/nejmoa0803894.
[51]. G. Cesaroni et al., "Long term exposure to ambient air pollution and incidence of acute coronary events: Prospective cohort study and meta-analysis in 11 european cohorts from the escape project," BMJ, vol. 348, no. January, pp. 1–16, 2014, doi: 10.1136/bmj.f7412.
[52]. A. S. V. Shah et al., "Global association of air pollution and heart failure: A systematic review and meta-analysis," Lancet, vol. 382, no. 9897, pp. 1039–1048, 2013, doi: 10.1016/S0140-6736(13)60898-3.
[53]. C. A. Pope, J. B. Muhlestein, H. T. May, D. G. Renlund, J. L. Anderson, and B. D. Horne, "Ischemic heart disease events triggered by short-term exposure to fine particulate air pollution," Circulation, vol. 114, no. 23, pp. 2443–2448, 2006, doi: 10.1161/CIRCULATIONAHA.106.636977.
[54]. T. S. Nawrot, L. Perez, N. Künzli, E. Munters, and B. Nemery,

"Public health importance of triggers of myocardial infarction: A comparative risk assessment," Lancet, vol. 377, no. 9767, pp. 732–740, 2011, doi: 10.1016/S0140-6736(10)62296-9.
[55]. V. L. Feigin et al., "Global burden of stroke and risk factors in 188 countries, during 1990–2013: a systematic analysis for the Global Burden of Disease Study 2013," Lancet Neurol., vol. 15, no. 9, pp. 913–924, 2016, doi: 10.1016/S1474-4422(16)30073-4.
[56]. M. Y. Sade, V. Novack, G. Ifergane, A. Horev, and I. Kloog, "Air pollution and ischemic stroke among young adults," Stroke, vol. 46, no. 12, pp. 3348–3353, 2015, doi: 10.1161/STROKEAHA.115.010992.
[57]. Y. Cai et al., "Associations of Short-Term and Long-Term Exposure to Ambient Air Pollutants With Hypertension: A Systematic Review and Meta-Analysis," Hypertension, vol. 68, no. 1, pp. 62–70, 2016, doi: 10.1161/HYPERTENSIONAHA.116.07218.
[58]. K. B. Fuks et al., "Arterial blood pressure and long-term exposure to traffic-related air pollution: An analysis in the european study of cohorts for air pollution effects (ESCAPE)," Environ. Health Perspect., vol. 122, no. 9, pp. 896–905, 2014, doi: 10.1289/ehp.1307725.
[59]. Y. Bagryantseva et al., "Oxidative damage to biological macromolecules in Prague bus drivers and garagemen: Impact of air pollution and genetic polymorphisms," Toxicol. Lett., vol. 199, no. 1, pp. 60–68, 2010, doi: 10.1016/j.toxlet.2010.08.007.
[60]. L. Jacobs et al., "Traffic air pollution and oxidized LDL," PLoS One, vol. 6, no. 1, 2011, doi: 10.1371/journal.pone.0016200.
[61]. Bukhsh, Madiha, et al. "An Interpretation of Long Short-Term Memory Recurrent Neural Network for Approximating Roots of Polynomials." IEEE Access 10 (2022): 28194-28205.
[62]. Tufail, Hina, M. Usman Ashraf, Khalid Alsubhi, and Hani Moaiteq Aljahdali. "The Effect of Fake Reviews on e-Commerce During and After Covid-19 Pandemic: SKL-Based Fake Reviews Detection." IEEE Access 10 (2022): 25555-25564.
[63]. Mumtaz, Mamoona, Naveed Ahmad, M. Usman Ashraf, Ahmed Alshaflut, Abdullah Alourani, and Hafiz Junaid Anjum. "Modeling Iteration's Perspectives in Software Engineering." IEEE Access 10 (2022): 19333-19347.
[64]. Asif, Muhammad, et al. "A Novel Image Encryption Technique Based on Cyclic Codes over Galois Field." Computational Intelligence and Neuroscience 2022 (2022).
[65]. Mehak, Shakra, et al. "Automated Grading of Breast Cancer Histopathology Images Using Multilayered Autoencoder." CMC-COMPUTERS MATERIALS & CONTINUA 71.2 (2022): 3407-3423.
[66]. Naqvi MR, Iqbal MW, Ashraf MU, Ahmad S, Soliman AT, Khurram S, Shafiq M, Choi JG. Ontology Driven Testing Strategies for IoT Applications. CMC-Computers, Materials & Continua. 2022 Jan 1;70(3):5855-69.
[67]. S. Tariq, N. Ahmad, M. U. Ashraf, A. M. Alghamdi, and A. S. Alfakeeh, "Measuring the Impact of Scope Changes on Project Plan Using EVM," vol. 8, 2020.
[68]. Asif M, Mairaj S, Saeed Z, Ashraf MU, Jambi K, Zulqarnain RM. A Novel Image Encryption Technique Based on Mobius Transformation. Computational Intelligence and Neuroscience. 2021 Dec 17;2021.
[69]. Ashraf, Muhammad Usman. "A Survey on Data Security in Cloud Computing Using Blockchain: Challenges, Existing-State-Of-The-Art Methods, And Future Directions." Lahore Garrison University Research Journal of Computer Science and Information Technology 5, no. 3 (2021): 15-30.
[70]. Ashraf MU, Rehman M, Zahid Q, Naqvi MH, Ilyas I. A Survey on Emotion Detection from Text in Social Media Platforms. Lahore Garrison University Research Journal of Computer Science and Information Technology. 2021 Jun 21;5(2):48-61.
[71]. Shinan, Khlood, et al. "Machine learning-based botnet detection in software-defined network: a systematic review." Symmetry 13.5 (2021): 866.
[72]. Hannan, Abdul, et al. "A decentralized hybrid computing consumer authentication framework for a reliable drone delivery as a service." Plos one 16.4 (2021): e0250737.
[73]. Fayyaz, Saqib, et al. "Solution of combined economic emission dispatch problem using improved and chaotic population-based polar bear optimization algorithm." IEEE Access 9 (2021): 56152-56167.
[74]. Hirra I, Ahmad M, Hussain A, Ashraf MU, Saeed IA, Qadri SF, Alghamdi AM, Alfakeeh AS. Breast cancer classification from histopathological images using patch-based deep learning modeling. IEEE Access. 2021 Feb 2;9:24273-87.
[75]. Ashraf MU, Eassa FA, Osterweil LJ, Albeshri AA, Algarni A, Ilyas I. AAP4All: An Adaptive Auto Parallelization of Serial Code for HPC Systems. INTELLIGENT AUTOMATION AND SOFT COMPUTING. 2021 Jan 1;30(2):615-39.
[76]. Hafeez T, Umar Saeed SM, Arsalan A, Anwar SM, Ashraf MU, Alsubhi K. EEG in game user analysis: A framework for expertise classification during gameplay. Plos one. 2021 Jun 18;16(6):e0246913.
[77]. Siddiqui N, Yousaf F, Murtaza F, Ehatisham-ul-Haq M, Ashraf MU, Alghamdi AM, Alfakeeh AS. A highly nonlinear substitution-box (S-box) design using action of modular group on a projective line over a finite field. Plos one. 2020 Nov 12;15(11):e0241890.
[78]. Ashraf, Muhammad Usman, et al. "Detection and tracking contagion using IoT-edge technologies: Confronting COVID-19 pandemic." 2020 international conference on electrical, communication, and computer engineering (ICECCE). IEEE, 2020.
[79]. Alsubhi, Khalid, et al. "MEACC: an energy-efficient framework for smart devices using cloud computing systems." Frontiers of Information Technology & Electronic Engineering 21.6 (2020): 917-930.
[80]. Riaz S, Ashraf MU, Siddiq A. A Comparative Study of Big Data Tools and Deployment PIatforms. In2020 International Conference on Engineering and Emerging Technologies (ICEET) 2020 Feb 22 (pp. 1-6). IEEE.
[81]. Ashraf MU, Eassa FA, Ahmad A, Algarni A. Empirical investigation: performance and power-consumption based dual-level model for exascale computing systems. IET Software. 2020 Jul 27;14(4):319-27.
[82]. Ashraf, Muhammad Usman, et al. "IDP: A Privacy Provisioning Framework for TIP Attributes in Trusted Third Party-based Location-based Services Systems." , International Journal of Advanced Computer Science and Applications (IJACSA) 11.7 (2020): 604-617.
[83]. Manzoor, Anam, et al. "Inferring Emotion Tags from Object Images Using Convolutional Neural Network." Applied Sciences 10.15 (2020): 5333.
[84]. Alsubhi, Khalid, et al. "A Tool for Translating sequential source code to parallel code written in C++ and OpenACC." 2019 IEEE/ACS 16th International Conference on Computer

Systems and Applications (AICCSA). IEEE, 2019.
[85]. Ashraf MU, Naeem M, Javed A, Ilyas I. H2E: A Privacy Provisioning Framework for Collaborative Filtering Recommender System. International Journal of Modern Education and Computer Science. 2019 Sep 1;11(9):1.
[86]. Ashraf MU, Ilyas I, Younas F. A Roadmap: Towards Security Challenges, Prevention Mechanisms for Fog Computing. In2019 International Conference on Electrical, Communication, and Computer Engineering (ICECCE) 2019 Jul 24 (pp. 1-9). IEEE.
[87]. Ashraf MU, Qayyum R, Ejaz H. "State-of-the-art Challenges: Privacy Provisioning in TPP Location Based Services Systems.", International Journal of Advanced Research in Computer Science (IJARCS). 2019 Apr 20;10(2):68-75.
[88]. Ashraf MU, Arshad A, Aslam R. Improving Performance In Hpc System Under Power Consumptions Limitations. International Journal of Advanced Research in Computer Science. 2019 Mar;10(2).
[89]. Javed, Rushba, et al. "Prediction and monitoring agents using weblogs for improved disaster recovery in cloud." Int. J. Inf. Technol. Comput. Sci.(IJITCS) 11.4 (2019): 9-17.
[90]. Ali, Muhammad, et al. "Prediction of Churning Behavior of Customers in Telecom Sector Using Supervised Learning Techniques." 2018 International Conference on Computer, Control, Electrical, and Electronics Engineering (ICCCEEE). IEEE, 2018.
[91]. Ashraf MU, Eassa FA, Albeshri AA, Algarni A. Performance and power efficient massive parallel computational model for HPC heterogeneous exascale systems. IEEE Access. 2018 Apr 9;6:23095-107.
[92]. Ashraf MU, Eassa FA, Albeshri AA, Algarni A. Toward exascale computing systems: An energy efficient massive parallel computational model. International Journal of Advanced Computer Science and Applications. 2018 Jan;9(2).
[93]. Ashraf MU, Arif S, Basit A, Khan MS. Provisioning quality of service for multimedia applications in cloud computing. Int. J. Inf. Technol. Comput. Sci.(IJITCS). 2018;10(5):40-7.
[94]. Ashraf MU, Eassa FA, Albeshri AA. Efficient Execution of Smart City's Assets Through a Massive Parallel Computational Model. InInternational Conference on Smart Cities, Infrastructure, Technologies and Applications 2017 Nov 27 (pp. 44-51). Springer, Cham.
[95]. Alrahhal, Mohamad Shady, et al. "AES-route server model for location based services in road networks." International Journal Of Advanced Computer Science And Applications 8.8 (2017): 361-368.
[96]. Ashraf MU, Eassa FA, Albeshri AA. High performance 2-D Laplace equation solver through massive hybrid parallelism. In2017 8th International Conference on Information Technology (ICIT) 2017 May 17 (pp. 594-598). IEEE.
[97]. Mumtaz, Mamoona, et al. "Iteration Causes, Impact, and Timing in Software Development Lifecycle: SLR." IEEE Access (2022).
[98]. Ahmad, Mubashir, et al. "Efficient Liver Segmentation from Computed Tomography Images Using Deep Learning." Computational Intelligence and Neuroscience 2022 (2022).
[99]. Tufail, Hina, et al. "The Effect of Fake Reviews on e-Commerce During and After Covid-19 Pandemic: SKL-Based Fake Reviews Detection." IEEE Access 10 (2022): 25555-25564.
[100]. Ashraf, Muhammad Usman. "Measuring the Impact of Factors Affecting Game Development in Distributed Software Development." Lahore Garrison University Research Journal of Computer Science and Information Technology 5.4 (2021): 50-61.
[101]. Khayam, Khuram Nawaz, et al. "Local-Tetra-Patterns for Face Recognition Encoded on Spatial Pyramid Matching." CMC-COMPUTERS MATERIALS & CONTINUA 70.3 (2022): 5039-5058.
[102].