Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 171 (2020) 2057–2066

Procedia Computer Science

www.elsevier.com/locate/procedia

Third International Conference on Computing and Network Communications (CoCoNet'19)

Forecasting Air Pollution Particulate Matter ($PM_{2.5}$) Using Machine Learning Regression Models

Doreswamy, Harishkumar K S[1], Yogesh KM, Ibrahim Gad

Department of Computer Science, Mangalore University, Mangalagangothri-574199, Karnataka, India

Abstract

From the past few decades, it has been observed that the urbanization and industrialization are expanding in the developed nations and are confronting the overwhelming air contamination issue. The citizens and governments have experienced and expressed the increasingly concerned regarding the impact of air pollution affecting human health and proposed sustainable development for overriding air pollution issues across the worldwide. The outcome of modern industrialization contains the liquid droplets, solid particles and gas molecules and is spreading in the atmospheric air. The heavy concentration of particulate matter of size PM_{10} and $PM_{2.5}$ is seriously caused adverse health effect. Through the determination of particulate matter concentration in atmospheric air for the betterment of human being well in primary importance. In this paper machine learning predictive models for forecasting particulate matter concentration in atmospheric air are investigated on Taiwan Air Quality Monitoring data sets, which were obtained from 2012 to 2017. These models were compared with the existing traditional models and perform better in predictive performance. The performance of these models was evaluated with statistical measures: Root Mean Square Error (RMSE), Mean Absolute Error (MAE), Mean Square Error (MSE), and Coefficient of Determination (R^2).

Keywords: TAQMN; Air Pollution; Prediction; Forecasting; Particulate matter ($PM_{2.5}$); Gradient Boosting regression.

1. Introduction

Air is one of the major components for all living organisms on the earth. From the last 50 years, still pollution is increasing because of urbanization, industrialization, automobiles, power plants, chemical activities, and some of the other natural activities such as agricultural burning, volcanic eruptions, and wildfires. All these activities cause the pollution growth, particularly particulate matter (PM) is one of the significant reason for air pollution [1]. Recently, air pollution turned into a serious issue on the earth and furthermore it is one of the significant reasons for death. The World Health Organization (WHO) estimated that around 6.9 millions of deaths owing to air contamination throughout

* Corresponding author Email: kskharish88@gmail.com Tel.:8884949395

1877-0509 © 2020 The Authors. Published by Elsevier B.V.

10.1016/j.procs.2020.04.221

the world. The recent updates of the Global Burden of Disease (GBD) study ranked the particulate matter (PM) as the 4^{th} leading cause of death out of 85 risk factors, being responsible for over 5 million deaths in 2017 [2].

A study of throughout the world, around 5000 cities have the foremost danger zone, they found that the most pollution cities are the main in Taiwan such as Pingtung station Newport city. Numerous individuals are dying on each year because of air pollution in Taiwan. Thus Taiwan has the most elevated death proportion from incessant illnesses and asthma in the world. In Newport city, contaminated air gets to be harmful to numerous human lives and particularly fifty percent of the kids are unhealthy influenced of pollution [3]. There are a strong relationship between the particulate matter and other pollutant factors such as $PM_{2.5}$, SO_2, NO_x, CO, PM_{10}, O_3, etc. Which gives the level of air pollution and the particulate matter level in air pollution. $PM_{2.5}$ particle size is less than 2.5 micrograms within the across which has been connected to numerous unfavorable well being impacts, counting cardiovascular and respiratory morbidity. Thus, $PM_{2.5}$ plays a pivotal role to health. Expectation and developing an early caution framework to supply air quality report towards the citizen have become up a self-evident and imperative need[4].

One of the main aim of smart cities is to act the basis of the data from sensors. However, some times sensors may be failures and errors by obtaining the data. To resolve these problems predictive model for forecasting the air quality in the smart cities are seems promising. The main goal of this paper is to analyze the different machine learning techniques for forecasting the particulate matter ($PM_{2.5}$). Thus the event of sensors failure can be predicted by this model on time, with the least possible errors in the amount of particulate matter ($PM_{2.5}$) in the atmosphere to create an alert by reaching the specific thresholds values are reached.

In this paper, The proposed machine learning models are used for $PM_{2.5}$ concentration prediction. We collected the data, Taiwan Air Quality Monitoring Network (TAQMN) from Taiwan country to train the model. TAQMN is containing air pollution data as well as meteorological data. The experiments are conducted on related data from Pingtung station, Newport city, Taiwan country. The rest of the paper is organized as follows: Section 2 we review the related work on air quality prediction. Section 3 introduces study area and methodology in details. Section 4 introduces models for predicting $PM_{2.5}$ concentration. Section 5 represents the performance criteria. Section 6 shows the results of the prediction of air pollution. Finally, Section 7 presents conclusion remarks.

2. Literature Review

Smart cities have pulled in impressive consideration within the setting of urban advancement approaches. The Web and broadband networking advances are seen as enablers of e-services and are getting to be progressively vital for urban improvement. Cities progressively accept a key part as drivers of development in areas such as health, consideration, environment and business [5].

Accordingly, there are various considers identified with action control cities [6], and this point is identified inside the urban communities. In the same area and related to actively control is the subject of discussing quality checking. Numerous of the works within the region of discussing contamination in smart cities centre focus on the checking of parameters considered as dangers.

A big information analytical-based approach [7] uses Ozone, CO, NO_2 and SO_2 levels together with information from smart cities, activity, time, observation, etc., assist in urban arranging choice making. In any case, these works dont propose any method for forecasting contamination for the following days or recognizing related variables. With an ever-increasing air pollution proportions, it is vital to execute effective discuss quality observing models, getting from the information collected by pollution sensors, that offer assistance to anticipate the concentration of artisans and give an evaluation of air contamination in each region. Consequently, air quality evaluation and forecast have become an important research region. In association with works inside the composition that do think about the conjecture of air defilement, there's an unmistakable bigger piece of use of Artificial Neural Networks (ANN) contrasted with different models, for example, multiple linear regressor (MLR) [8]. Being this the common slant, it is critical to watch that ANNs present a few shortcoming for this theme, as distinguished [9]. An expository strategy for show determination and they take after a long-running process to get the foremost precise demonstration. At long, it is worth saying that modern approaches that combine machine learning methods with the utilize of air pollution information to get it and make strides forecasts on air contamination [10].

Past thinks about utilizing AOD to gauge ground $PM_{2.5}$ concentrations were conducted in North American, Europe, and territory China. There are couple of considers performed in districts with complex geology. Has utilized

Multi-Point Execution of Air Redress Multi-Angle Implementation of Atmospheric Correction (MAIAC) recoveries of AOD (1-km spatial determination) to assess ground PM_{10} and $PM_{2.5}$ amidst 2003 to 2013 in Israel. They found that the coefficient of judgments (R^2) of cross validation (CV) were 0.79 and 0.72 for PM_{10} and $PM_{2.5}$, separately, illustrating that the AOD-based demonstrate can be connected in a little range with the strenuous territory and parched climate conditions. Moderate Resolution Imaging Spectroradiometer (MODIS) AOD to gauge ground-level PM_{10} concentrations amid 2000 to 2009 in Lombardy, North Italy. They detailed a tall execution of AOD-based estimation show (R^2 = 0.787, year to year variety 0.738 to 0.818) and declared that AOD-based estimation show could be a solid strategy to assess PM_{10} concentration in a region with complex geographic and climate designs [11]. It would be important to investigate the suitable application of satellite-retrieved AOD to the estimation of ground-level $PM_{2.5}$ concentrations over tropical/subtropical islands that have heterogeneous arrive utilize sorts, complex geology, and sticky climate conditions, such as Taiwan [12]. Due to soaking height alter and rough territories, the scenes and climate conditions in Taiwan display tall spatial heterogeneity, and mostly cloudy or completely cloudy sky happens habitually. These all contribute to the challenge of building up AOD-$PM_{2.5}$ estimation show. Inspected three a long time off today AOD and surface $PM_{2.5}$ information at one location in northern Taiwan. They found the straight relapse between AOD and $PM_{2.5}$ can be essentially moved forward when the blended layer statures decided from the near-by ground-based pollution estimations are joined into the relapse. The advance made strides,comes by isolating the relapse models for distinctive administrations of vaporized vertical dispersion decided by pollution termination profiles. Utilized 4-year fawning AOD information and $PM_{2.5}$ concentrations at eight stations in Taiwan to get the AOD-$PM_{2.5}$ demonstrate, and utilized the evaluated $PM_{2.5}$ levels to explore the impact of $PM_{2.5}$ on the event of respiratory maladies in 2008 [13].

3. Study Area and Methodology

3.1. Data set Description

Newport is one of the major city in Pingtung station in Taiwan country that is situated between the latitude of 23°28′45″N and longitude of 120°26′56″E as shown in Figure 1. Air pollution is becoming a very critical problem in Taiwan. The main sources of air pollution are industrialization and particulate matter ($PM_{2.5}$) which is the most important pollutant [14]. The upper limits of the annual average particulate matter allowed in Taiwan weather stations are Lunbei (132 gm3), Taixi (100gm3), Annan (101gm3), Tainan (104 gm3), Douliu (150gm3), Xingang (126gm3), Puzi (108gm3), Newport (161gm3) are the cities of the highest level of particulate matter. We have collected the data from Taiwan for the years 2012 to 2017, Taiwan Air Quality Monitoring Network (TAQMN) data set available for 76 stations in different locations [15]. All of the parameters can be isolated into two groups of particular data and chronological data. Chronological data refers to that information, which quickly alters in a brief length of time. Particular parameters are parameters that change in area and time which cause changes within the concentration of air pollution. Figure 2 outlines the division of the utilized information within the two groups of particular and chronological data [16].

4. Models

The parameters affecting the air pollution incorporate topographic segments (x, y, z) because of the tremendous zone of Taiwan and huge tallness changes meteorological parts because of their immediate or round about impacts on poison outflow week-time segments, as in different long stretches of the week, the traffic in Taiwan is unique and in like manner the measure of toxin focus will be extraordinary [17]. For instance, on the ends of the week, Taiwan encounters less traffic thickness and the long stretch of year part additionally changes because of residue, and temperature reversal prompts changes in the degree of the toxins. Then in this examination, the air pollution estimations of the two closest air pollution station neighbors are utilized as a compelling parameter in the expectation of the air pollution.

The pertinent information to the centralization of PM_{10} and $PM_{2.5}$ contamination were getting from TAQMN during 2012 to 2017. The grouping of these toxins is gathered once a day and conveyed by the organization [18]. Taiwan region has 76 stations for estimating air contamination where it is oversaw and constrained by Taiwan air

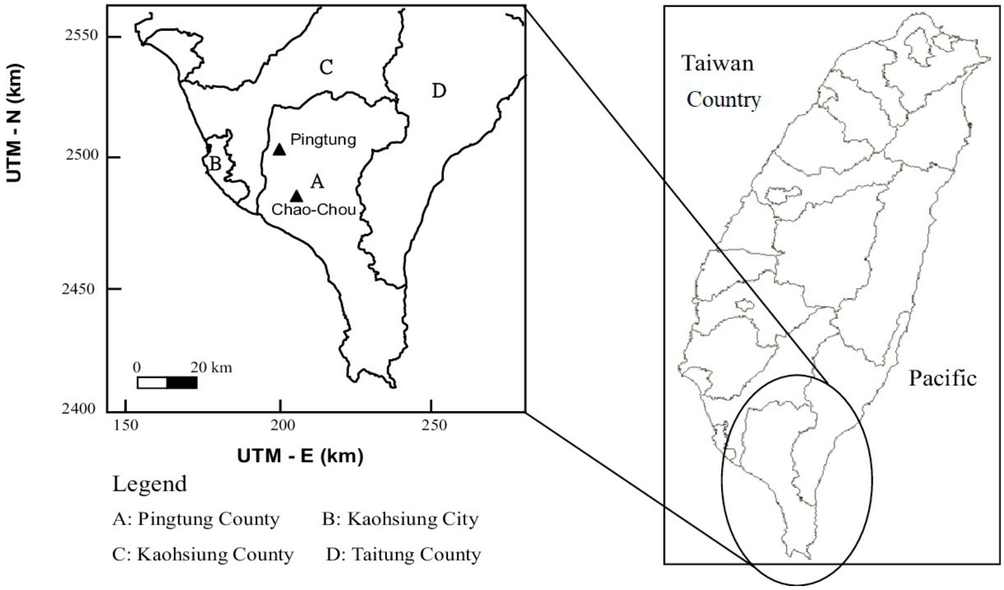

Fig. 1: Location of the study region.

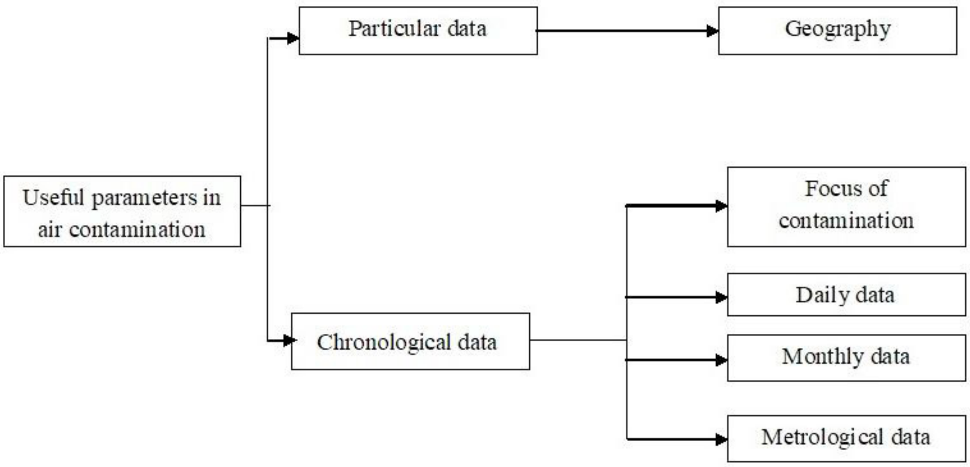

Fig. 2: The pollution data in this research.

quality monitoring network. The experiments needs day by day information which are accessible for every one of the stations, the information identified with 76 stations has been utilized.

To forecast the air pollution, we utilized machine learning strategies, such as linear regression, random forest regressor, gradient boosting regressor, k neighbors regressor, MLP regressor, and Decision Tree regressor CART [19]. Figure 3 shows the proposed architecture of the prediction models for air pollution. Moreover, Figure 4 represents the general steps for each model used.

The data collected from the TAQMN for the six years information was incomplete. To make up for this information fill, this examination has used Fourier arrangement and spline multinomial approaches.These two methods used to fill the missing values in TAQMN data set those are shown in Figure 3. Having filtered the meteorological information, estimations of these parameters should utilize this data in anticipating the air pollution.

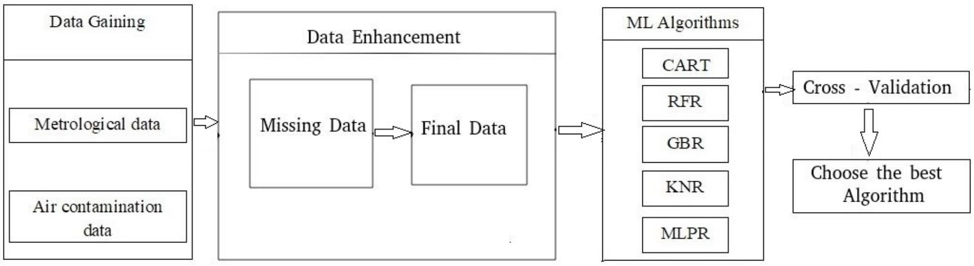

Fig. 3: The proposed architecture of prediction model for air pollution.

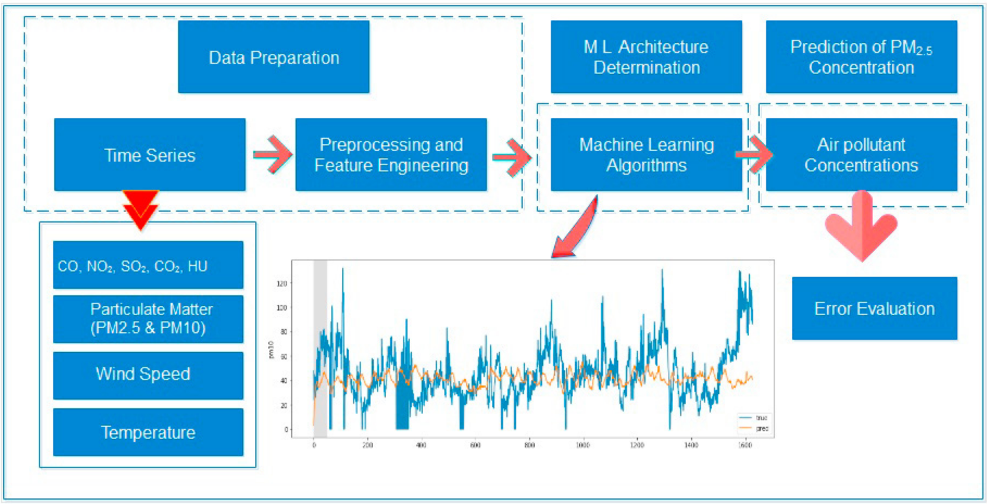

Fig. 4: The general steps for air pollution.

After the preprocessing and enhancement of data, the filtered meteorological data and air pollution data are used to forecast the $PM_{2.5}$ level. In order to perform prediction, a set of machine learning algorithms used such as linear regression, random forest, gradient boosting, K-neighbors, MLP, decision tree CART. Finally, the cross-validation are used to select the best model, that can be used for forecasting [20].

4.1. Random Forest

Random forest algorithm is the most popular ensemble learning model that can used for classification and regression problem. Moreover, random forest creates the forest with a number of trees, and the algorithm is more robust when the number of trees is large[21]. The advantages of random forest algorithm are:

1) Can use for both classification and regression task.
2) Handle missing values in data set.
3) It can used for categorical values.

To perform prediction by random forest:

1) Randomly create a decision tree by the test features and use the rules to predict the target.
2) Calculate the votes for each predicted target.
3) The final prediction is calculated by high voted predicted target.

4.2. Gradient Boosting Regression

Recently, one of the most widely used ensemble algorithms is gradient boosting (XG Boost) and it is can be used for regression and classification problems. Moreover, XG Boost is designed for more speed and performance based on gradient boosted decision trees.

The features of XG Boost are:

1) It can be used to create the trees in parallel.

2) In the training step, it is possible to use a cluster of computers to run XG Boost, simply this refers to distributed computing.

3) It is used optimize the hardware by cache optimization of data.

$$D_{t+1}(i) = \frac{D_t(i)\exp\left(-\alpha_t y_i h_t(x_i)\right)}{Z_t} \tag{1}$$

where, $h(i)$ will be predicted by the learner, y is the actual response (+ 1 or -1), and α is the kind of learning rate [22].

4.3. Decision Tree Regression

Decision tree (DT) algorithm is a well known machine learning model that falls under the category of supervised learning. Moreover, it can be used for both classification and predictions tasks. The basic idea of DT is use the tree representation to solve the proposed problem in which each leaf node corresponds to a specific class label and columns are represented on the internal nodes of the created tree [23]. The main problem to apply DT is to select the attribute for the root node in each level. To over come this problem (attribute selection) there are two attribute selection measures namely information gain and gini index.

$$Information\ Gain(S,A) = Entropy(S) - \sum_{v \in Values(A)} \frac{|S_v|}{|S|} Entropy(S_v) \tag{2}$$

$$Gain(S,A) = Entropy(S) - \sum_{v \in Values(A)} \frac{|S_v|}{|S|} Entropy(S_v) \tag{3}$$

where S is the set of instances, A is an attribute, S_v is subset of S, and values(A) is the set of all possible values of A.

4.4. MLP Regression

MLP is stand for a multilayer perception, which is a famous class of Artificial Neural Network (ANN). Moreover, MLP is consists of multiple layers of perceptrons or at least three layers of nodes namely input layer, hidden layer, and output layer. The nodes in each layer is called neurons and each neuron has an activation function such as Sigmoid, Softmax ,...., and ReLU [24].

5. Performance Criteria

Some of the statistical evaluations are used to evaluate the model performance such as Mean Absolute Error (MAE), Root Mean Square Error (RMSE), Mean Square Error (MSE), and coefficient of determination (R^2). The criteria formulas are shown in below:

$$MAE = \frac{\sum_{i=1}^{m} |x_i - \hat{x}_i|}{m} \tag{4}$$

$$RMSE = \sqrt{\frac{\sum_{i=1}^{m} (x_i - \hat{x}_i)^2}{m}} \tag{5}$$

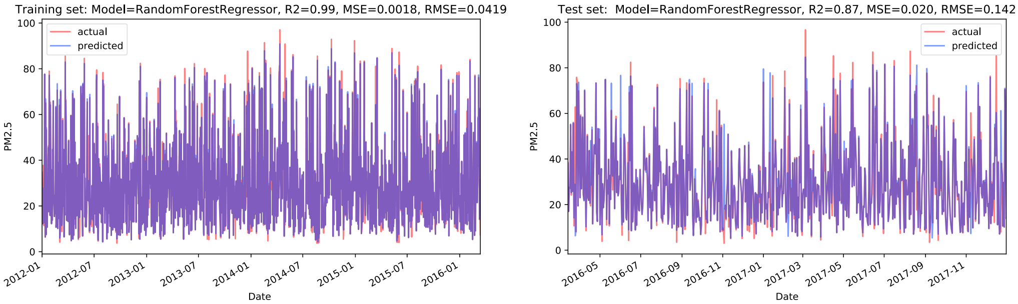

Fig. 5: prediction model for Random Forest regression.

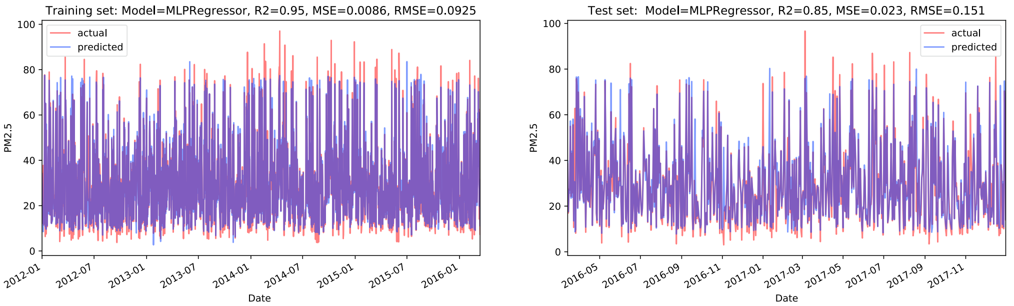

Fig. 6: prediction model for MLP Regression.

$$MSE = \frac{1}{m}\sum_{j=1}^{m}(x_i - \hat{x}_i)^2 \tag{6}$$

where, m is the number of observations, $\hat{x}_i$ is the predicted value, and x_i is the actual value.

$$R^2 = \left[\frac{1}{M}\frac{\sum_{j=1}^{M}\left[(Y_j - \bar{Y})\left(X_j - \bar{X}\right)\right]}{\sigma_y \sigma_x}\right]^2 \tag{7}$$

Where, M is the number of observations, σ_x is the standard deviation of the observation X, σ_y is the standard deviation of Y, X_j is the observed values, $\bar{X}$ is the mean of the observed values, Y_j is the calculated values, and $\bar{Y}$ is the mean of the calculated values [25].

6. Results and Discussion

The considered TAQMN data set in this work is stationary time-series data from January 2012 to December 2017 besides, its included the major cities of Taiwan country namely Annan, Chiayi, Good, Mcmug, and Newport. The major pollutant city is Newport that is selected and focuses only on particulate matter $PM_{2.5}$ because its the most harmful effect of pollutant in the air pollution. We collected the data in the form of hourly then it's converted into daily as well as monthly. Likewise, the proposed model gradient boosting regression was examined with including models like linear regression, lasso regression, ridge regression, random forest regression, K-neighbors regression, MLP regression, and decision tree regression. The experiments are done by using different models, and for each model the cross-validation and performance criteria is used to measure the accuracy of the model. For instance, the determination coefficient explains the relation between the actual values and predicted values.

Table 1: The best result of the diagnostics train and test values for different machine learning models for forecasting of the $PM_{2.5}$

	Train				Test			
Models	R^2	**MSE**	**RMSE**	**MAE**	R^2	**MSE**	**RMSE**	**MAE**
Linear Regression	0.7927	0.0331	0.1819	0.1186	0.7207	0.0446	0.2111	0.1364
Lasso	0.6262	0.0597	0.2443	0.1888	0.5652	0.0665	0.2579	0.1918
Ridge	0.7927	0.0331	0.1819	0.1186	0.1104	0.1361	0.3690	0.1646
Random Forest Regressor	0.9872	0.0020	0.0451	0.0125	0.8699	0.0199	0.1411	0.0372
Gradient Boosting Regressor	**0.9983**	**0.0002**	**0.0161**	**0.0101**	**0.8891**	**0.0169**	**0.1302**	**0.0380**
K Neighbors Regressor	1	0	0	0	0.6755	0.0496	0.2228	0.1491
MLP Regressor	0.9471	0.0084	0.0918	0.0388	0.8587	0.0216	0.1470	0.0562
Decision Tree Regressor CART	0.9996	6.2811	0.0079	0.0023	0.7992	0.0307	0.1753	0.0421

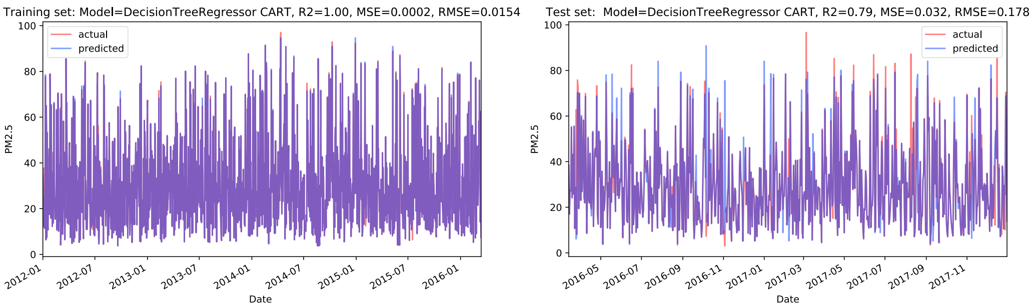

Fig. 7: prediction model for Decision Tree Regression.

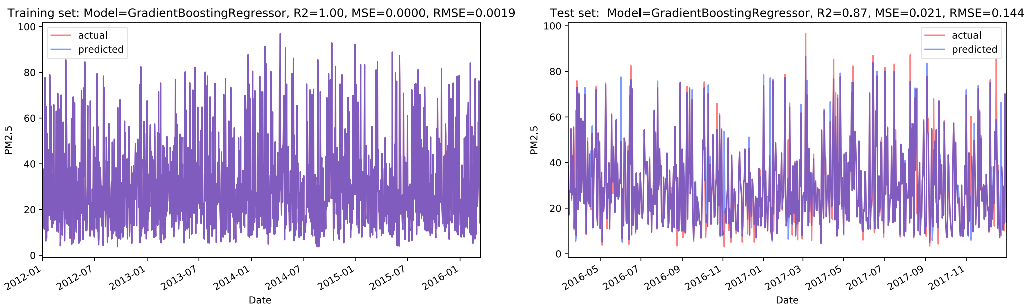

Fig. 8: prediction model for Gradient Boosting Regression.

Figure 5 shows the results of prediction using random forest regression algorithm, left part indicate the training data and right part for testing data where R^2 is 0.99 and 0.87 respectively. Moreover, the red color line indicates the actual values and blue color line indicates the prediction of the particulate matter $PM_{2.5}$. Similarly, Figure 6 shows the results of prediction for MLP regression algorithms, R^2 is 0.95 and 0.85 for training data and testing data respectively. Figure 7 shows decision tree algorithm training data and testing data R^2 are 1.00 and 0.79 respectively. Figure 8 shows the results of prediction using gradient boosting regression algorithm, R^2 the training data and testing data are 1.00 and 0.89 respectively. Moreover, coefficient of determination values are shown in Table 1.

Table 1 show the results for each algorithm, gradient boosting regression performs better compared with the rest of algorithms. Moreover, Table 2 and Figure 9 presents the comparison between the best model from Table 1 and the related work such as Bayesian-Down-scaling model [26], LSTM, and CART-CM-LF-AN.

Table 2: The comparison results for proposed with model others models for forecasting of the $PM_{2.5}$

Models	R^2	RMSE
Bayesian Down-scaling model MCMC algorithms	0.7062	0.6895
LSTM(L)	0.7485	0.4236
CART-CM-LF-AN	0.7909	0.3258
Gradient Booting regression	**0.9336**	**0.1302**

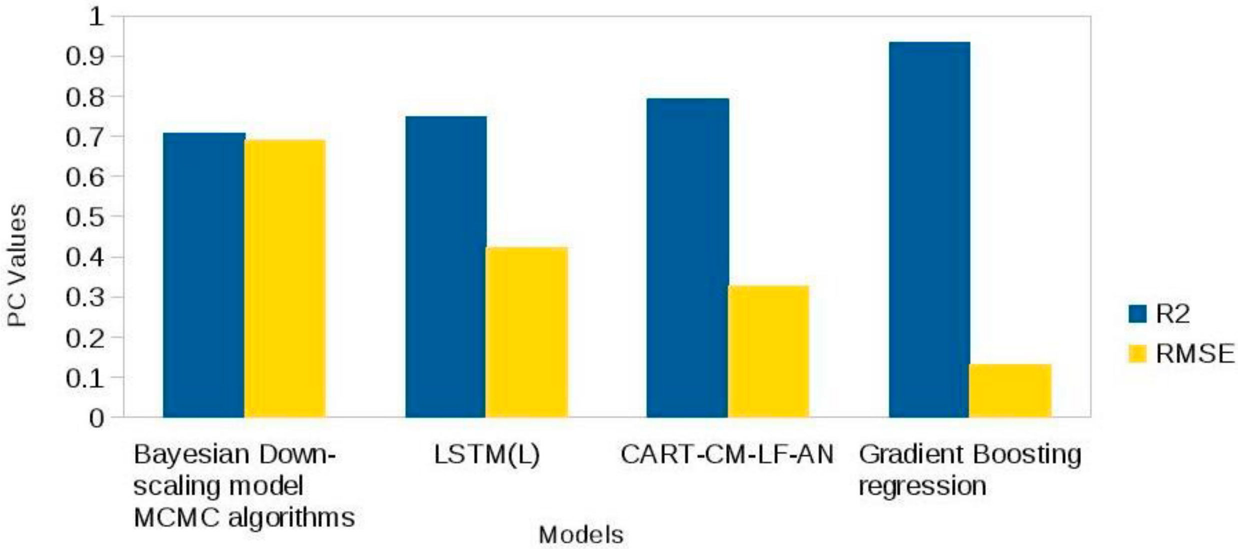

Fig. 9: The comparison results for proposed with model others models for forecasting of the $PM_{2.5}$

7. Conclusions

In this work, the proposed machine learning models to analyze the air pollution on TAQMN data in Taiwan is presented. Particularly, there are 76 air pollution stations are recorded in TAQMN data from 2012 to 2017. The particulate matter $PM_{2.5}$ prediction are done using machine learning models based on the statistical calculations of metrics such as MAE, MSE, RMSE, and R^2. The results show that the proposed models values are perform better compares to the previous models and also it show that the actual values and predicted values are very close to each other. Finally, we conclude that gradient boosting regressor model is better for forecasting air pollution on the TAQMN data.

References

[1] Jung, Chau-Ren, Bing-Fang Hwang, and Wei-Ting 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." *Environmental Pollution* **237** (1): 1000-1010.

[2] Soh, Ping-Wei, Kai-Hsiang Chen, Jen-Wei Huang, and Hone-Jay Chu. "Spatial-Temporal pattern analysis and prediction of air quality in Taiwan." *10th International Conference on Ubi-media Computing and Workshops (Ubi-Media)IEEE* (2), pp. 1-6.

[3] Mehdipour, Vahid, David S. Stevenson, Mahsa Memarianfard, and Parveen Sihag. "Comparing different methods for statistical modeling of particulate matter in Tehran, Iran." *Air Quality, Atmosphere and Health* **11**, no.(10)(3): 1155-1165.

[4] Delavar, Mahmoud Reza, Amin Gholami, Gholam Reza Shiran, Yousef Rashidi, Gholam Reza Nakhaeizadeh, Kurt Fedra, and Smaeil Hatefi Afshar. "A Novel Method for Improving Air Pollution Prediction Based on Machine Learning Approaches: A Case Study Applied to the Capital City of Tehran." *ISPRS International Journal of Geo-Information* **8**, no. 2 (4): 99.

[5] Wang, Wei, Feiyue Mao, Lin Du, Zengxin Pan, Wei Gong, and Shenghui Fang. "Deriving hourly PM2. 5 concentrations from himawari-8 aods over beijingtianjinhebei in China." *Remote Sensing* **9**, no. 8 (5): 858.

[6] Pan, Bingyue. "Application of XGBoost algorithm in hourly PM2. 5 concentration prediction." *In IOP Conference Series: Earth and Environmental Science,IOP Publishing* **vol. 113**, no. 1 (6), p. 012127.

[7] Sun, Xiaotong, Wei Xu, and Hongxun Jiang. "Spatial-temporal prediction of air quality based on recurrent neural networks." *In Proceedings of the 52nd Hawaii International Conference on System Sciences* **9** (7).

[8] Sharma, Nidhi, Shweta Taneja, Vaishali Sagar, and Arshita Bhatt. "Forecasting air pollution load in Delhi using data analysis tools." *Procedia computer science* **132** (8): 1077-1085.

[9] Martnez-Espaa, Raquel, Andrs Bueno-Crespo, Isabel Maria Timon-Perez, Jess Soto, Andrs Muoz, and Jos M. Cecilia. "Air-Pollution Prediction in Smart Cities through Machine Learning Methods: A Case of Study in Murcia, Spain." *J. UCS 24* **132**, no. 3:(9) 261-276.

[10] Wang, Yikai, Xuefei Hu, Howard Chang, Lance Waller, Jessica Belle, and Yang Liu. "A bayesian downscaler model to estimate daily PM2. 5 levels in the conterminous US."*International journal of environmental research and public health* **15**, no. 9 (10): 1999.

[11] Soh, Ping-Wei, Jia-Wei Chang, and Jen-Wei Huang. "Adaptive deep learning-based air quality prediction model using the most relevant spatial-temporal relations."*IEEE Access* **6** (11):38186-38199.

[12] Liu, Hui, Haiping Wu, Xinwei Lv, Zhiren Ren, Min Liu, Yanfei Li, and Huipeng Shi. "An intelligent hybrid model for air pollutant concentrations forecasting: Case of Beijing in China."*Sustainable Cities and Society 47* **1** (12): p.101471.

[13] KamiÅska, Joanna A. "A random forest partition model for predicting NO2 concentrations from traffic flow and meteorological conditions." *Science of The Total Environment 47* **651** (13): p. 475-483.

[14] Park, Seohui, Minso Shin, Jungho Im, Chang-Keun Song, Myungje Choi, Jhoon Kim, Seungun Lee et al. "Estimation of ground-level particulate matter concentrations through the synergistic use of satellite observations and process-based models over South Korea."*Atmospheric Chemistry and Physics* **19**, no. 2(14): p.1097-1113.

[15] Kang, Ganganjot Kaur, Jerry Zeyu Gao, Sen Chiao, Shengqiang Lu, and Gang Xie. "Air quality prediction: Big data and machine learning approaches." *International Journal of Environmental Science and Development* **9**, no. 1 (15): p.8-16.

[16] Lyu, Baolei, Yuzhong Zhang, and Yongtao Hu. "Improving PM2. 5 air quality model forecasts in China using a bias-correction framework."*Atmosphere* **8**, no. 8 (16): p.147.

[17] Rybarczyk, Yves, and Rasa Zalakeviciute. "Machine learning approaches for outdoor air quality modelling: A systematic review." *Applied Sciences* **8**, no. 12 (17): p.2570.

[18] Environmental Protection Administration. AQI value and impact on health. Available from: https://taqm.epa.gov.tw/taqm/en/

[19] Karpatne, Anuj, Imme Ebert-Uphoff, Sai Ravela, Hassan Ali Babaie, and Vipin Kumar. "Machine learning for the geosciences: Challenges and opportunities." *IEEE Transactions on Knowledge and Data Engineering* (18).

[20] Grange, Stuart K., David C. Carslaw, Alastair C. Lewis, Eirini Boleti, and Christoph Hueglin. "Random forest meteorological normalisation models for Swiss PM 10 trend analysis." *Atmospheric Chemistry and Physics* **18**, no. 9 (19): p.6223-6239.

[21] Zhao, Fang, and Weide Li. "A Combined Model Based on Feature Selection and WOA for PM2. 5 Concentration Forecasting."*Atmosphere* **10**, no. 4 (20): p.223.

[22] Isakov, Vlad, Saravanan Arunachalam, Stuart Batterman, Sarah Bereznicki, Janet Burke, Kathie Dionisio, Val Garcia et al. "Air quality modeling in support of the near-road exposures and effects of urban air pollutants study (NEXUS)."*International journal of environmental research and public health* **11**, no. 9 (21): p.8777-8793.

[23] Zhu, Dixian, Changjie Cai, Tianbao Yang, and Xun Zhou. "A machine learning approach for air quality prediction: Model regularization and optimization."*Big data and cognitive computing* **2**, no. 1 (22): p.5.

[24] Suleiman, A., M. R. Tight, and A. D. Quinn. "Applying machine learning methods in managing urban concentrations of traffic-related particulate matter (PM10 and PM2. 5)." *Atmospheric Pollution Research* **10**, no. 1 (23): p.134-144.

[25] Xie, Jingjing, Xiaoxue Wang, Yu Liu, and Yun Bai. "Autoencoder-based deep belief regression network for air particulate matter concentration forecasting." *Journal of Intelligent and Fuzzy Systems* **34**, no. 6 (24): p.3475-3486.

[26] Yao, Fei, Jiansheng Wu, Weifeng Li, and Jian Peng. "Estimating Daily PM2. 5 Concentrations in Beijing Using 750-M VIIRS IP AOD Retrievals and a Nested Spatiotemporal Statistical Model."*Remote Sensing* **11**, no. 7 (25): p.841.