Air pollution has had negative effects on human health and has interfered with social functions; partilces with diameters less than 2.5 𝝁 m (PM2.5) have especially been the primary pollutants in many cities in the USA. Among air pollutants, PM2.5 is onethe most harmful pollutants and can easily cross the human defense barrier, entering the lungs, causing human disease and even death, because of its small size particle and potential for long-term exposure (Wu et al., 2018; Chen et al., 2019c; Wei et al., 2019a). The PM2.5 observations were from environmental monitoring stations, however, the quantity of available PM2.5 data presented regional differences due to the uneven station distribution. He et al. (2016) conducted research that indicates the PM2.5 pollution index was positively correlated with the emergency admission rate of female acute myocardial infarction, and with the increased incidence of diabetes and hypertension.

Recently, due to an increase in the application of machine learning models to various fields in order to increase the accuracy of predictions, machine learning has also been used to predict particle concentrations (Kuremoto et al., 2014; Ong et al., 2016; Gui et al., 2020). The regression and boosting models, and deep learning-based methods to make predictions display remarkable performance in time-series data processing to make predictions (Hochreiter and Schmidhuber, 1997). The estimation using traditional statistical methods requires a large amount of historical data to construct the relationship between explanatory variables and target variables (Breiman, 2001b).

Earlier studies used limited number of statistical models, but in our study we used nearly six machine learning models to find the best accuracy of predictions. In addition to this, our research paper takes a novel approach in PM2.5 concentration research, by exploring concentrations over USA as opposed to China, where many existing PM2.5 studies have already been conducted. .

Datasets:

2.1 Ground PM2.5 measurements:

Daily PM2.5 observational data was collected from January 2015 to December 2021 from the openaq air quality datasets (<https://openaq.org/>). The datasets are available from nearly 1081 stations around USA. The PM2.5 concentrations of ground sites were taken as the dependent variable of the model. In this paper, the daily PM2.5 concentration data (Mg/m3) of 1081 ground monitoring stations were sorted in to monthly and seasonal data from January 2015 to December 2021, The data integrity exceeded 97%. The datasets were calibrated and quality-controlled according to national standards. Figure 1 shows the ground-level monitoring sites coverage over the United States, and these sites collected 7 years of daily continuous observations. From the Figure, we can see that PM2.5 monitoring sites are greater in number in the Eastern part than in the western part of USA. We observed small data gaps and we applied linear interpolation for filling the gaps of PM2.5 datasets. However, stations are sparsely located, therefore ground level PM2.5 monitoring sites face difficulties in meeting the data requirements (Lin et al., 2015). As expected, the PM2.5 concentrations were much lower at remote sites compare to the urban ares, mainly due to the absence of anthropogenic sources.

This study aims to achieve the best statistical comparison of nine machine learning models; Linear Regression, Kneighbors Regressor, Logistic Regression, Gradient Boosting Regressor, Ada Boost Regressor, Decision Tree Regressor, XG Boost, Support Vector Regressor, Random Forest, Support Vector Machines, and LSTM for estimating the PM2.5 concentrations over the specified period. The datasets are split into 80% and 20% as training and testing datasets, respectively. . The training datasets are used to build the model, and the testing dataset is used to verify the model performance of the trained model.

K Nearest Neighbors (K-NN): K-NN technique is one of the earliest ML model (reference). The K-NN model categorizes each unknown instance in the training set by a majority among its nearest neighbors. Its performance is also crucially dependent on the Euclidean distance metric used to define the most immediate neighbors. After determining the Euclidean distance between the data, the database samples sorted in ascending order from the least distance (maximal similarity) to maximum distance minimum similarity) [Wu et al. 2008].

Random Forest (RF): RF is a machine learning algorithm and was proposed by Breiman (2001a);it integrates multiple trees through the idea of ensemble learning, utilizes classification and regression tree (CART) as learning algorithms of decision trees. The RF is a set of decision trees, where the structure of each one and the space of the variables is divided into smaller subspaces so that the data in each region is as uniform as possible [Hastie et al., 2005 and Breiman, 2001].It uses the bootstrap resampling technique to randomly (with replacement) extract k samples from the original training set to generate new training samples. RF uses multiple base classifiers to obtain higher accuracy classification results by voting or averaging.

XGBoost: is a highly efficient and optimized distributed gradient boosting algorithm. XGBoost supports a range of different predictive modeling problems, such as classification and regression. It is trained by minimizing the loss of an objective function against a dataset, and the loss function is a critical hyperparamater which is tied directly to the type of problem being solved. Regular gradient boosting, stochastic gradient boosting, and regulararized gradient boosting are the three main forms of gradient boosting. For efficiency, the system features include parallelization, distributed computing, out-of-core computing, cache optimization, and optimization of data structures to achieve the best global minimum and run time.

Long Short Term Memory (LSTM): LSTM is well suited for prediction based on time-series data, with better performance, to learn long-term dependency, and it deals with exploding and vanishing gradient problems [Alahi et al., 2016, Kong et al., 2017]. LSTM is superior to traditional ML methods in processing large input data, relatively fast computational speed, and is a type of Recurrent Neural Network (RNN) [Rumelhart et al., 1986], that has been proposed to predict future outputs using past inputs. LSTM is great at processing time-series data because the PM2.5 concentrations are time-dependent, and it can predict future air pollution concentrations by learning features contained in past air pollution concentration time-series data.

Decision Tree (DT): Decision Trees are one of the most commonly used machine learning models in classification and regression problems. DT is a simple model that has branches, nodes, and leaves. To split a node into two or more sub-nodes DT uses mean squared error (MSE). It is a tree structure with three types of nodes. The root node is the initial node, which may get split into further nodes of the branched tree that finally leads to a terminal node (leaf node) that represents the prediction or final outcome of the model. The interior nodes and branches represent features of a data set and decision rules respectively. The final prediction is the average of the value of the dependent variable in that particular leaf node.

Gradient Boosting Regression: The type of boosting that combines simple models called weak learners into a single composite model. Gradient boosting involves optimizing the loss function and a weak learner makes predictions. Generally, the gradient descent procedure is used to minimize a set of parameters, such as coefficients in a regression equation or weights in a neural network. After estimating loss or error, the weights are updated to minimize that error. Gradient Boosting algorithms minimize the bias error of the model. The Gradient Boosting algorithm predicts the target variable using regressor with a Mean Square Error (MSE), cost function (for regression problems) or predicts, the target variable with a classifier using a Log Loss cost function (for classification problems)..

Support Vector Regression (SVR): The SVR model is widely applied to time series prediction problems. It is a novel forecasting approach, which is trained independently based on the same training data with different targets. The SVR there are several functions subjected to linear or non-linear kernel functions. The linear function is used for the linear regression model, the model and, evaluates results with metrics such as Root Mean Square Error (RMSE) and Mean Absolute Error (MAE) to estimate the performance of the model.

AdaBoost Regressor: AdaBoost (Adaptive Boosting) is a popular technique, as it combines multiple weak classifiers to build one strong classifier. The boosting approach is a class of ensembles of ML algorithms and is described by Schapire (1990). Generally, the boosting approach requires a large amount of training data, which is not possible for many cases, and one way of mitigating this issue is by using AdaBoost (Freund and Schapire, 1997). The main difference of AdaBoosting from most of the other boosting approaches is in computing loss functions using relative error rather than absolute error. AdaBoost regressor fits the data set and adjusts the weights according to the error rate of the current prediction, and reduces the bias as well as the variance for supervised learning.

Linear Regression:

This plot shows the USA monthly anomalies and quantiles for four years using daily PM2.5 values. The monthly anomalies are in percent form, so subtract 100 to set the average value to zero. In addition, we estimated the anomaly to be positive or negative. Using anomalies we estimated the minimum, maximum values, the 25%, 75% quantiles, and the interquartile range for each month of the entire time period, and the resultant plot is shown in Figure 2. During 2018, in USA, the highest levels were observed at the inland locations and it declined nearly 20% in the year 2019. In the inland areas, are primarliy influenced by the secondary particles secondary particle formation resulting from the oxidation of gaseous precursors (NOx, SOx,and NH3) (South Coast Air Quality Management District, 2017). PM2.5 concentrations shows drastic changes during before and pandamic years. Before pandamic years the PM2.5 concentrations are are high in spring and summer months and the concentration are high in end of summer (August) and Early fall (September) dusing summer years. The monthly PM2.5 concentrations are greatest in 2018 when compared to other years. The positive anomalies are observed on a higher frequency in August 2018 whereas negative anomalies are observed more in September 2018. This indicates that before COVID-19 the PM2.5 concentrations were a little higher than in other years throught the USA. PM2.5 values were also higher in the Eastern USA than in the Western USA (Figure not shown).

Evaluation Parameters:

For model evaluation, The error between the estimated and true values was evaluated using several evaluation indexes, the statistical metrics are selected for comparing the performance of the models, namely, the coefficient of determination (R2), Error-values between computed and observed data are evaluated by Mean Square Error (MSE), mean absolute error (MAE), PSR, NSE, and PBIAS as follows:

The nine machine learning models can describe daily variations of observed and estimated values of PM25 concentrations as shown in Figure 3, in which the blue curve represents the observed PM25 concentrations, while the red curve represents the estimated PM25 concentrations. The PM25 concentrations in autumn and winter are less accurate because air pollution is more severe than that in spring and summer. The SVM and RF models give desirable results observed using the entire PM25 concentrations.

California (Figure 4) and New York (Figure 5) states scatter plots of the observed vs estimated daily PM25 concentrations during the period of observations using different machine learning models. The performance and statistical metrics are estimated for each model and are shown in Table 1. PM2.5 estimations are lower and higher than observations with high and low PM25 concentration scenarios, indicating that estimation accuracy will decline in extreme cases in both states. Zhan et al. (2017) also found similar behavior using PM2.5 concentration in some parts of China. This may be due to the model caused by the a smaller amount of training data, especially during extreme PM25 concentrations. Overall, the performance of LSTM is reasonable, with California’s R2, RMSE, and MAE values of 0.85, 17.77 mg/m3, and 10.66 mg/m3, respectively. New York’s R2, RMSE, and MAE values are 0.81, 20.77 mg/m3, and 10.67 mg/m3, respectively. Comparing California’s to New York’s results, we observe that the PM2.5 concentration values and biases are slightly higher. Overall, the average error values are slightly higher in the Eastern states than in the Western states. Each state’s R2, RMSE, MAE, and bias values are estimated for each model and we observed LSTM and RF models produce better estimates than the other models, and these results are tabulated in Table 2. On average, the R2 of the RF models is 10% higher than the that of the LSTM model. The biases are 15% higher in the Eastern states than in the Western states of the USA.

However, the PM25 estimations in the autumn and winter are less accurate because air pollution is more severe than that present in the spring and summer. Among the seven machine learning models, only the SVM and RF models give desirable results in the mildest air pollution cases. The GAM model performs the worst among all models, which can neither reflect the variations of PM25 concentrations significantly nor estimate the PM25 concentrations accurately.

Wu, X.; Kumar, V.; Quinlan, J. R.; Ghosh, J.; Yang, Q.; Motoda, H.; McLachlan, G. J.; Ng, A.; Liu, B,; Philip, S. Y.; Top 10 algorithms in data mining. Knowl. Inf. Syst. 2008, 14, 1-37.

Hastie, T,; Tibshirani, R.; Friedman, J,; Franklin, J. The elements of statistical learning: Data mining, inference and prediction. Math. Intell. 2005, 27, 83-85.

Breiman, L. Random Forests, Mach. Learn. 2001, 45, 5-32.

A. Alahi, K. Goel, V. Ramanathan, A. Robicquest, L. Fei-Fei, and S. Savarese, “Social LSTM: Human trajectory prediction in crowded spaces”, in Proc. IEEE Conf. Comput.Vis. Pattern Recognit., Jun, 2016, pp.961-971.

D.E. Rumelhart, G. E. Hinton, and R. J. Williams, “learning representations by back-propagating errors, “Nature, Vol. 323, no.6088, pp. 533-536, 1986.

W. Kong, Z. Y. Dong, Y. Jia, D. J. Hill, Y. Xu, and Y. Zhang. “Short-term residential load forecasting based on LSTM recurrent neural network”, IEEE Trans. Smart Grid, vol. 10, no.1, pp. 841-851, Jan. 2017.

Schapire, R. (1990). The strength of weak learnability. Machine Learning, 5 (2), 197-227.

Freund, Y., Schpire, R (1997). A decision-theoretic generalisation of on-line learning and an application of boosting. J. Computer and System Sciences, 55 (1), 119-139.

https://www.r-bloggers.com/2019/07/visualize-monthly-precipitation-anomalies-2/