# PM2.5 Prediction using Machine Learning Hybrid Model for Smart Health

J. Angelin Jebamalar , A. Sasi Kumar

***Abstract:*** ***Air Pollution is one of the current serious issue attributable to people's health causing cardiopulmonary deaths, lung cancer and several respiratory problems. Air is polluted by numerous air pollutants, among which Particulate Matter (PM2.5) is considered harmful consists of suspended particles with a diameter less than 2.5 micrometers.This paper aims to acquire PM2.5 data through IoT devices,store it in Cloud and propose an improved hybrid model that predicts the PM2.5 concentration in the air. Finally through forecasting system we alert the public in case of an undesired condition. The experimental result shows that our proposed hybrid model achieve better performance than other regression models.***

***Keywords: IoT,Cloud, Air pollution, PM2.5, Machine Learning, Prediction, Ensemble, Regression algorithms***

## I. INTRODUCTION

Air pollution is a serious environmental issue leading to global warming and having a greater impact on human health causing premature death, cancer, respiratory illnesses or heart disease. The Air Quality Index (AQI) is an indicator to describe the air quality level based on the concentration of several pollutants in the atmosphere, commonly PM2.5, PM10, carbon monoxide, sulphur dioxide, nitrogen dioxide and ozone.

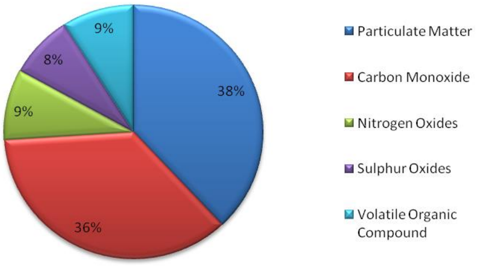

**Fig. 1. Major air pollutants**

Among the air pollutants PM2.5 is most dangerous fine particles with diameters that are generally 2.5 micrometers and smaller in diameter that can penetrate easily into the bloodstream and causes serious health hazards. A recent research shows that the air particles penetrated their way from the lungs to the placenta and may reach foetus directly by the mother.The major contributor to these fine particles in air are produced from industries,motor vehicles,burning fossil fuels. Internet of Things (IoT) is a collection of “smart devices” capable of sensing and connect with their surroundings to acquire the data. The huge amount of data captured from these devices introduces challenges associated with the storage and processing capabilities of the data.An efficient solution for the managing these challenges is Cloud Computing.This paper acquires the air pollutant data using sensor and relay on cloud for storage and processing.

In predicting PM2.5 concentration, machine learning regression algorithms play a major by extracting data and finds the hidden information and helpful in predictive analysis. In this paper we predict the air pollutant PM2.5 using a hybrid model which is a combination of decision tree and light GDM machine learning regression technique.We have also presented the comparative analysis of this hybrid model with other regression techniques, based on the two metrics MAE and RMSE. The Mean Absolute Error(MAE) is the average absolute differences between actual and prediction and the Root Mean Squared Error(RMSE) is the average squared differences between actual and predicted values.

## II. LITERATURE REVIEW

At present, many machine learning techniques have been proposed for solving air pollution prediction problems based on simple regression models.Authors in [3] have performed to estimate PM2.5 using random forest model and two other traditional regression models, the random forest shows the high accuracy in predicting the PM2.5 concentration.

In [1] authors made a comparative study of machine learning techniques to predict the quality of air using Apache spark with multiple data sets and concluded that the random forest was a best technique in prediction but it actually work well for small size dataset and performs well only on classification problems. In [6] author proposed new model based on LSTM(Long Short-Term Memory) to forecasting PM2.5 based on the historical data and in [18] author achieves in predicting PM2.5,NO2,SO2 air pollutants concentrations with ARIMA(Auto Regressive Integrated Moving Average, Simple and Exponential Weighted Moving Average , KF algorithm and obtained the optimal result but the data handling and processing time was not discussed.

Revised Manuscript Received on October 30, 2019.

* Correspondence Author

**J. Angelin Jebamalar***, Ph.D Research Scholor, Department of Information Technology, School of Computing Sciences, Vels Institute of Science, Technology and Advanced Studies (VISTAS), Pallavaram, Chennai, India. E-mail : jebamalar1@gmail.com

**Dr. A. Sasi Kumar**, Professor, Department of Information Technology, School of Computing Sciences, Vels Institute of Science, Technology and Advanced Studies, Pallavaram, Chennai, India. E-mail : askmca@yahoo.com

*Retrieval Number: A1187109119/2019©BEIESP*
*DOI: 10.35940/ijeat.A1187.109119*
*Journal Website: www.ijeat.org*

*Published By:*
*Blue Eyes Intelligence Engineering*
*& Sciences Publication*

Q. Zhou, H. Jiang [20] proposed a model to forecast air quality using a hybrid model of multiple neural networks and shown that the random forests model gave higher predictive accuracy than the other two traditional regression model for a smaller dataset consisting of few hours .

Many other previous research works has shown XGboost algorithms works well in air quality predictions.Also many works using deep learning techniques but the data was too small.

This paper, proposed a new hybrid model based on the LightGBM boosting ensemble technique and decision tree algorithm that is suitable for the prediction of air quality features and metrological features.

## III. PROPOSED ARCHITECTURE

The proposed architecture is shown in the figure 3 consists of four phases:

- Data Sensing:The PM2.5 data captured using the particle sensor SDS011 with Raspberry Pi. The collected data can be stored local or cloud.
- Data Storage:In this phase,the collected data is stored in cloud and retrieved for predictive analysis.
- Predictive Phase:In the predictive phase,PM2.5 concentration is predicted using the new hybrid.

Model and the performance is compared with other regression models then an alert can be sent to the public or hospitals to manage resources for respiratory illness.

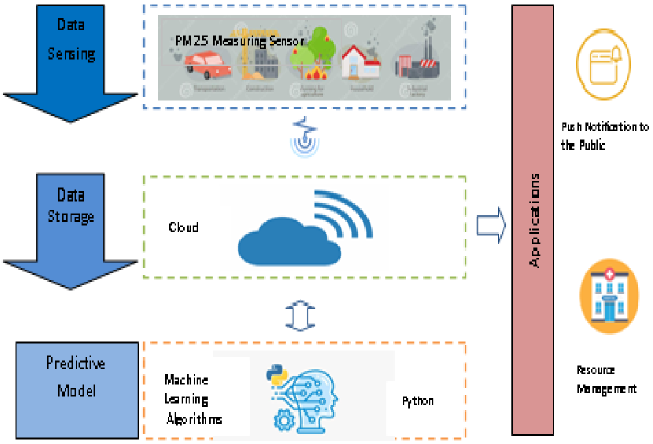

**Fig.2 Architecture Diagram**

## IV. METHODOLOGY

### A. Machine Learning

Machine Learning is a notion that enables the machine to learn from examples and experience of the given data. The two popular types of machine learning methods are classification and regression.

**Classification:** Classification is a supervised machine learning algorithm in which the computer program gains from the input information it receives and then utilizes this figuring to group new perception. Classification algorithms are used when the desired output is a discrete label. Classification technique can be performed on structured or unstructured data which categorize the data into given number of classes. The primary objective of a classification issue is to define the class and to which the new information will fall under.

**Regression:** Regression is also a supervised learning algorithm useful for predicting outputs that are continuous. Regression algorithms forecast the output values based on input attribute from the data fed in the system. The go-to methodology creates model based on the features of training data and helps in predicting values for new data.

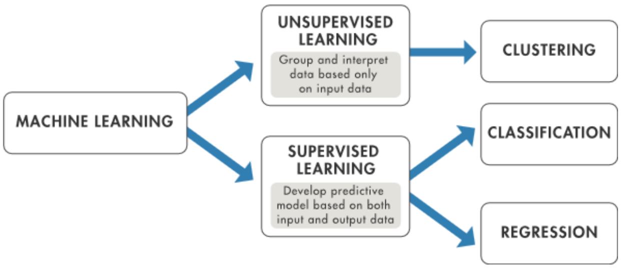
MACHINE LEARNING
UNSUPERVISED LEARNING
Group and interpret data based only on input data
SUPERVISED LEARNING
Develop predictive model based on both input and output data
CLUSTERING
CLASSIFICATION
REGRESSION

**Fig. 3. .Machine Learning Techniques**

In this paper, we predict PM2.5 using a hybrid model and few other regression algorithms and compare its performance.

- Linear Regression
- Lasso Regression
- Support Vector Regression
- Neural Network
- Random Forest
- Decision Tree
- XGBoost

*1) Linear regression*

Linear regression is a linear approach using a best fit straight line equation **y=a+b*x** to model the relationship between two variables dependent and independent variables.This line equation can be used to predict the target variable value(y) based on predictor variable(x).

*2) Lasso regression*

Least Absolute Selection Shrinkage Operator(LASSO) is a type of linear regression that attain the predictors by summing up the regression coefficients and hence said to be shrinking the data values to the mean value.Thus minimize prediction error of regression coefficient.

*3) SVR*

Support vector regression is a linear model that fits the error within certain threshold i.e.,based on margin-based loss function .Thereby minimize the error and maximize the margin.Thus it helps in high prediction accuracy.

*4) Neural network regression*

A neural network is a supervised algorithm used in machine learning to make predictions based on existing data. Neural network input layer take inputs based on existing information.

*Published By:*
*Blue Eyes Intelligence Engineering*
*& Sciences Publication*

Hidden Layer of network layer use backpropagation strategy to optimize the input variable weights that improve the prediction.The data from input and hidden layers yield the prediction of the output layer.

*5) Random forest*

The additive random forest model combines decisions from a sequence of basic models to make predictions.It uses bagging ensembling method that train individual models in parallel way.

*6) Decision tree*

Decision tree regression trains a model in the structure of a tree to predict the data.It searches for every distinct values for your predictors and chooses to the split the target variable.

*7) XGBoost*

e**X**treme **G**radient **B**oosting is a bagging ensemble based machine learning algorithm that train the individual models in a sequential way from the previous model.

The performance of each algorithm is measured using the following metrics:

- MEAN ABSOLUTE ERROR is the average absolute differences between actual and predicted value.
- ROOT MEAN SQUARED ERROR is the average of squared differences between actual and predicted values.

### B. Hybrid Model

The proposed hybrid model is a combination of decision tree and light gradient boosting model. Hybrid model is very fast with high-performance, as it uses the boosting ensembling technique. Boosting is a sequential ensemble technique that tries to correct the errors from the previous model. The light boosted decision tree hybrid model splits the tree leafwise with the best fit whereas other boosting algorithms split the tree levelwise. Because of its high speed and lower memory, Light boosted decision tree algorithm can handle large size of data and hence gives high accuracy of results.

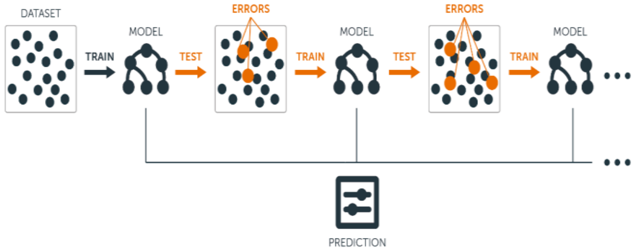

**Fig. 4. Implementation Model**

<u>Algorithm: Light Boosted decision tree</u>

```
Input: Training dataset $(x,y)i^N$, Loss function L, Number of
         iterations M
Output: Final model $f_M(x)$
1. Initialize $f_0(x)$ be constant.
2. for m =1 to M do
   Compute the weighted error residual $e_{im} = \Sigma(y_i - y_i^p)^2$
where $y^i$ is the target value and $y_i^p$ the prediction for i=1 to N
3. Correct the previous predictions by adding the predicted
   residuals by traversing the decision tree leafwise.
4. Fit the decision tree to the error residual.
5. Repeat steps 2-4 until the  a sum of residuals become
   constant
```

### C. Implementation model

Data captured by sensors goes through preprocessing steps for removing of irrelevant features and to handle missing values using PCA.

Principal Component Anlysis (PCA) decreases the amount of factors by capturing the highest variance in the data in to a new coordinate system with main components ‘axes. The dataset divided into training to fit the model and testing data to test it. It is processed using a hybrid model which combines the ensembled boosting technique in decision tree model that helps in predicting the PM2.5 in air. Using a cloud-based forecasting system we alert the public in case of an undesired condition in the PM2.5 level.

Implementation was carried using Python Programming Language on Windows XP operating system. Pre-processing steps carried out using Pandas. Machine learning algorithms were implemented using scikit learn library and and evaluated using sklearn metrics.Visualization of data was done using plotly library,The code for all these was written on Jupyter Notebook.

### D. Dataset

The predictive feature PM2.5 dataset was captured using SDS011 sensor with Raspberry pi in Ayanavaram,Chennai city from Feb 2017 to Jan 2019.The PM2.5 prediction uses several features PM10,humidity,temperature ,pressure, wind direction, wind speed ,date and time and merged with the meteorological data.

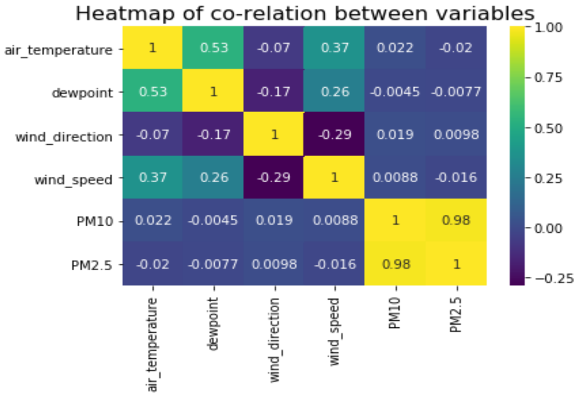

**Fig 4.Heatmap of interrelated features**

## V. RESULTS AND DISCUSSIONS

To forecast the performances in predicting the PM2.5 concentration, the models was trained using the historical data. The experimental results show that our proposed model’s performance is precise than other models.Also our proposed hybrid model, the light boosted decision tree resulted in lower root mean square error (*RMSE*) and mean absolute error(*MAE*) values which makes it suitable for PM2.5 prediction .The following figure shows the metrics scores of each models.

*Retrieval Number: A1187109119/2019©BEIESP*
*DOI: 10.35940/ijeat.A1187.109119*
*Journal Website: www.ijeat.org*

*Published By:*
*Blue Eyes Intelligence Engineering*
*& Sciences Publication*

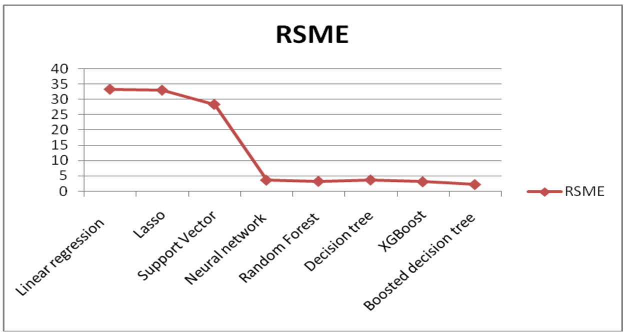

**Fig 5. Root Square Mean Error**

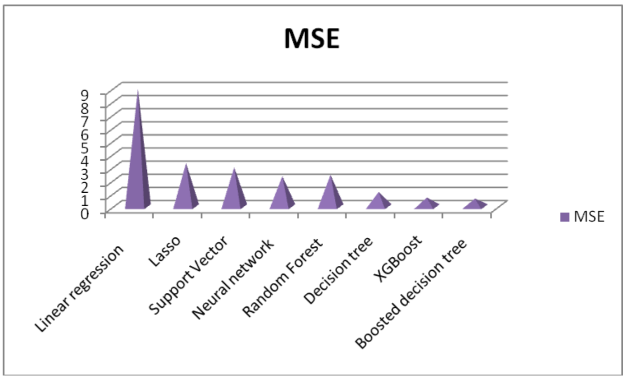

**Fig 5. Mean Square Error**

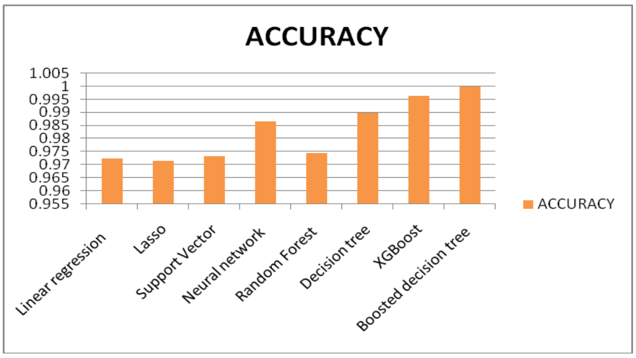

**Fig 5. Accuracy**

## VI. CONCLUSION

This paper proposed a Hybrid light boosted decision tree model process the air sensor data to foretell the PM2.5 concentration. The approach is a combination of decision tree and light GDM,where the light GDM is based on ensemble boosting technique that corrects the errors of previous model sequentially.

Also the computation time of our hybrid model is very fast taking less storage memory.Thus our hybrid model experimental results shows that it outperforms compared to the other models.Forecasting the PM2.5 level helps the public and the hospital to take precautionary actions.

## FUTURE WORK

In future, we plan to explore the performance of the techniques using other ensembling methods and to study other factors that affects the air pollution.Future work will also focus on forecasting image-based air quality using convolutional deep learning neural network.

## REFERENCES

1. Saba Ameer, Munam Ali Shah, Abid Khan,"Comparative analysis of machine learning techniques for predicting air quality in smart cities",IEEE 2019.
2. Ping Wei Soh, Jia Wei Chang "Adaptive Deep Learning-Based Air Quality Prediction Model Using the Most Relevant Spatial-Temporal Relations"IEEE,May 2018.
3. Gongbo Chen, Shanshan,"A machine learning method to estimate PM2.5 concentrations across China with remote sensing, meteorological and land use information",Elsevier,2018, Science of the Total Environment,pp.52-60.
4. Chen,Zhang,"Estimating PM2.5concentrations based on non-linear exposure-lag-response associations with aerosol optical depth and meteorological measures. Atmos. Environ",2018,pp.30–37.
5. Yi Lin, Long ,"Air quality forecasting based on cloud model granulation",Springer,2018.
6. Sachit Mahajan,Liu," Improving the accuracy and efficiency of PM2.5 Forecast Service Using Cluster-Based Hybrid Neural Network Model ",IEEE,2017.
7. K. A. Delic, "On resilience of iot systems: The internet of things (ubiquity symposium)",Ubiquity, vol.1,February, 2016.
8. Y.Xing,Xu,Shi,Y.-X. Lian, "The impact of PM2.5 on the human respiratory system" ,Journal of Thoracic Disease, vol. 8,pp. 69–74, January 2016.
9. Qi Meng, Guolin Ke, Taifeng Wang, Wei Chen, Qiwei Ye, Zhi-Ming Ma, Tieyan Liu. A communication-efficient parallel algorithm for decision tree. In Advances in Neural Information Processing Systems, pp. 1271–1279, 2016.
10. Tianqi Chen, Carlos Guestrin, "XGBoost: A Scalable Tree Boosting System", ACM, KDD , August 13-17, 2016.
11. G. Ke et al.,"LightGBM: A highly efficient gradient boosting decision tree" in Proc. Adv. Neural Inf. Process. Syst., 2017, pp. 3149_3157.
12. A. C. Cosma and R. Simha, ``Machine learning method for real-time non-invasive prediction of individual thermal preference in transient conditions," Building Environ., vol. 148, pp. 372-383, Jan. 2019.
13. Million premature deaths annually linked to air pollution." [Online]. Available: https://www.who.int/phe/eNews_63.pdf
14. Asgari, Marjan, Mahdi ,Zeinab Ghaemi"Predictive mapping of urban air pollution using Apache Spark on a Hadoop cluster." International Conference on Cloud and Big Data Computing, pp. 89-93. ACM,2017.
15. D. Zhu, C. Cai, T. Yang, and X. Zhou, "A Machine Learning Approach for Air Quality Prediction: Model Regularization and Optimization"
16. Q. Zhou, H. Jiang, J.Wang, and J. Zhou, ``A hybrid model for PM2:5 forecasting based on ensemble empirical mode decomposition and a general regression neural network," Sci. Total Environ., vol. 496, pp. 264_274,Oct. 2014.
17. Y.F. Xing, Y.H. Xu, M.H. Shi, and Y.X. Lian. The impact of PM2.5 on the human respiratory system. In Journal of Thoracic Disease, vol.8, no. 1, pp. 6974, January 2016.
18. Xiaozheng Lai, Ting Yang , ZetaoWang ;Peng Chen 2," IoT Implementation of Kalman Filter to Improve Accuracy of Air Quality Monitoring and Prediction"Applied Sxiences,May 2019.
19. Prettenhofer, Peter, and Gilles Louppe. "Gradient boosted regression trees in scikit-learn." ,2014.
20. Decision Trees scikit learn0:20:1documentation:Scikit learn:org; 2018: [Online]:Available: https: scikit-learn.org/stable/modules/tree.html.
21. Liang, X., S. Li, S. Zhang, H. Huang, and S. X. Chen, "PM2.5 data reliability, consistency, and air quality assessment in five Chinese cities", J. Geophys.Res. Atmo,2016.
22. Chai, Tianfeng & Draxler, R."Root mean square error (RMSE) or mean absolute error (MAE)Arguments against avoiding RMSE in the literature. Geoscienti_c Model Development",pp.1247-1250,2014.
23. P. Jiang, Q. Dong, and P. Li, "A novel hybrid strategy for PM2.5 concentration analysis and prediction," Journal of Environmental Management, vol. 196, pp. 443–457, 2017.
24. Yuan Y, Liu S, Castro R and Pan X. PM2.5 monitoring and mitigation in the cities of China. Environ Sci Technol. 2012.

25. D. Wang, S. Wei, H. Luo, C. Yue, “A novel hybrid model for air quality index forecasting based on two-phase decomposition technique and modified extreme learning machine," Sci.Environ.,vol. 580, pp.719-733, Feb. 2017.
26. G. Ke et al.”Light GBM: A highly efficient gradient boosting decision tree," in Proc. Adv. Neural Inf. Process. Syst.,pp. 3149-3157,2017.