Application of Data Mining Techniques on Air Pollution of Dhaka City

Al-Sadman Chowdhury, Md. Shihab Uddin, Md Rashad Tanjim, Fariha Noor, Rashedur M. Rahman

Department of Electrical and Computer Engineering, North South University, Dhaka, Bangladesh

ABSTRACT

In recent times, the air quality level of Dhaka city has been termed as hazardous. The weather of Dhaka city has gone through some drastic changes because of extreme air pollution. In this paper, we have applied several machine learning models that include deep learning such as Long Short-Term Memory (LSTM) and proposed different techniques to forecast the air quality level of Dhaka city. Furthermore, we demonstrate the applicability of machine learning and deep learning models in the classification and prediction of the Air Quality Index (AQI) based on some pre-determined range. The novelty of this approach is that we have considered daily temperature as a parameter for air pollution prediction. We conduct an extensive evaluation of these models and show that different machine learning models can classify the AQI of different places of Dhaka city. LSTM models can also forecast hourly and daily AQI with optimal performance.

Keywords-component; Data Mining; Air Quality Index; Machine Learning; LSTM.

INTRODUCTION

Air Pollution occurs when the level of air pollutants exceeds a certain limit. In our paper, we have used machine learning models to classify AQI level of different places of Dhaka city and we have used deep learning approaches using time series modeling to show in what way the air quality has decreased over the years.

- For the machine learning part, we have used decision tree, random forest, SVM, Kstar, Ensemble selection, Multi-Layer Perception and bagging models.
- For the deep learning part, we only have used LSTM in two scenarios. One is for hourly prediction and the other is for daily prediction.

AQI	Air Pollution Level	Health Instructions	Cautionary Statement
0-50	Good	No health implications.	Normal Outdoor activity for everyone.
51-100	Moderate	Acceptable air quality. However, it can be harmful for hypersensitive people.	Caution for Hypersensitive people.
101-150	Unhealthy For sensitive groups	People with sensitive health condition can be affected to a large extent	Caution for children, elders and hypersensitive people
151-200	Unhealthy	Normal People may feel a bit uncomfortable while breathing while sensitive group can have a heart disease.	Sensitive people should avoid outdoor activities and general people should reduce outdoor activities
201-300	Very Unhealthy	Normal people will have a slight effect while sensitive people will be affected significantly.	People should remain indoor unless it's an emergency.
300+	Hazardous	Healthy people can have a respiratory problem. Elders and the sick will be affected the most. Healthy people should also remain at home	Everyone should remain indoor and avoid physical exertion specially for the sensitive people

AQI Categorization and its Implication

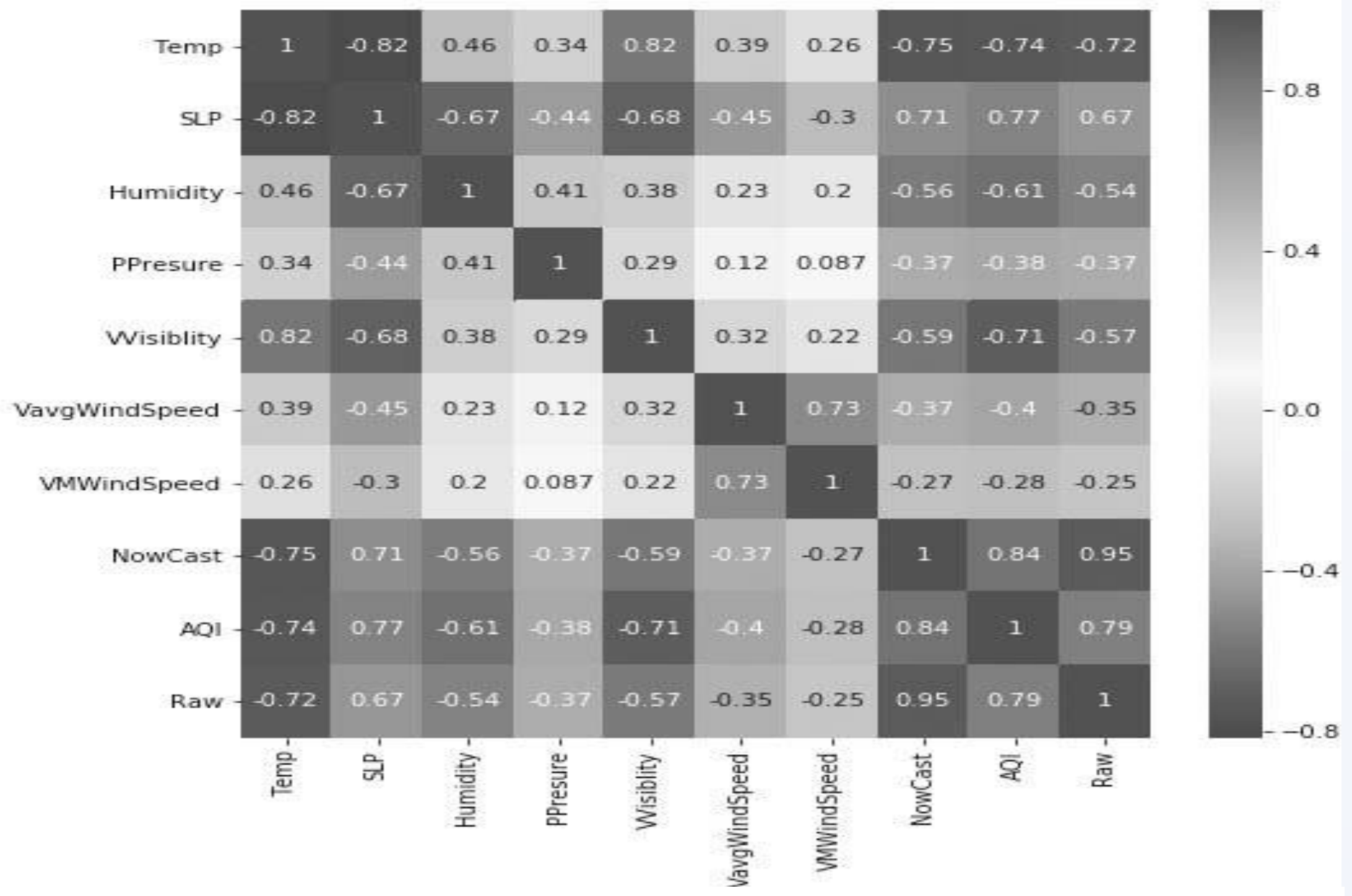
METHODOLOGY

Datasets:

For our paper, we have used two datasets in our analysis. 1. Weather Dataset. 2. Air pollution Dataset

Preprocessing:

1. Conversion 2. Replacement 3. Normalization of the Attributes 4. Feature Selection 5. Correlation Matrix



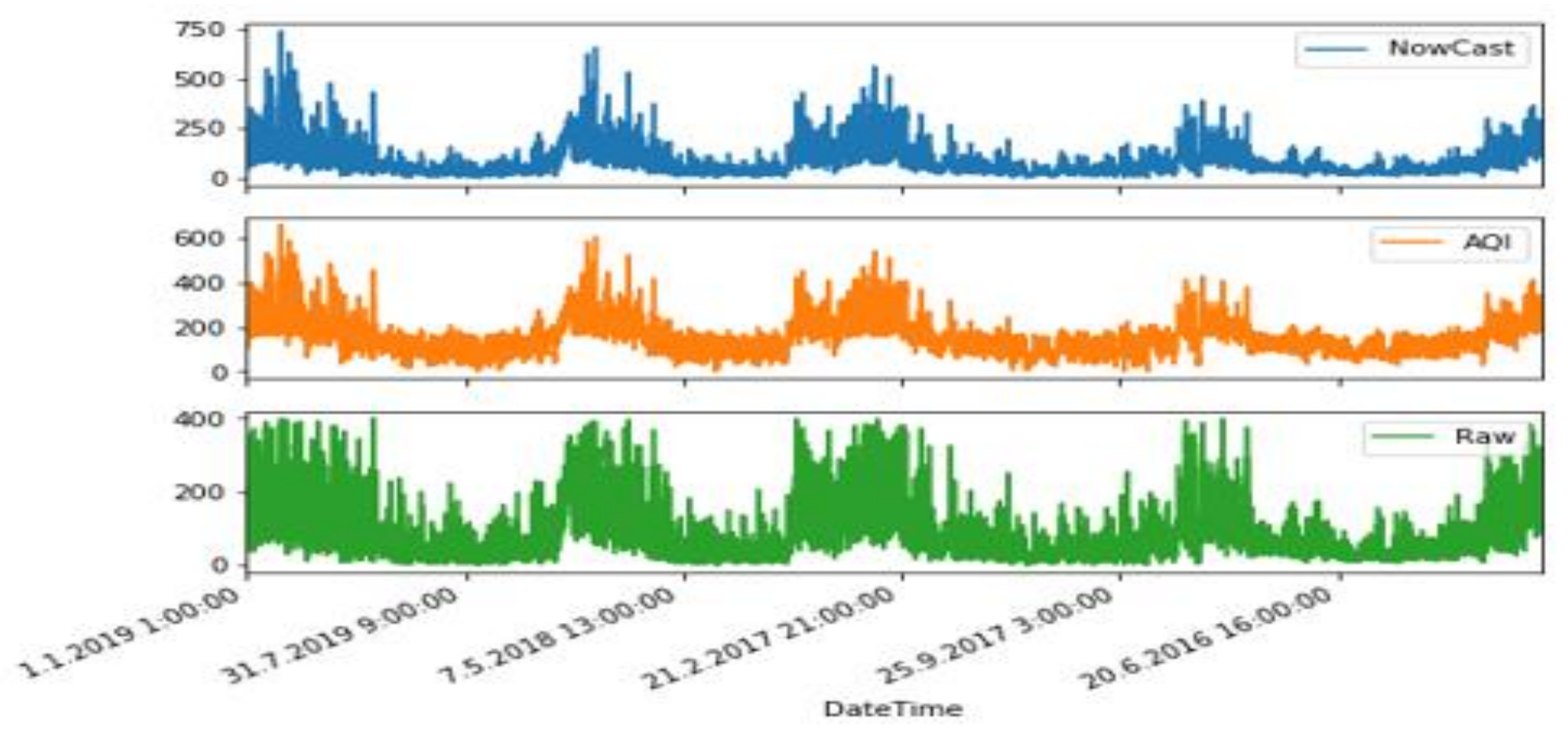
	Temp	SLP	Humidity	PPresure	Wisiblity	VavgWindSpeed	VMWindSpeed	NowCast	AQI	Raw
Temp	1	-0.82	0.46	0.34	0.82	0.39	0.26	-0.75	-0.74	-0.72
SLP	-0.82	1	-0.67	-0.44	-0.68	-0.45	-0.3	0.71	0.77	0.67
Humidity	0.46	-0.67	1	0.41	0.38	0.23	0.2	-0.56	-0.61	-0.54
PPresure	0.34	-0.44	0.41	1	0.29	0.12	0.087	-0.37	-0.38	-0.37
Wisiblity	0.82	-0.68	0.38	0.29	1	0.32	0.22	-0.59	-0.71	-0.57
VavgWindSpeed	0.39	-0.45	0.23	0.12	0.32	1	0.73	-0.37	-0.4	-0.35
VMWindSpeed	0.26	-0.3	0.2	0.087	0.22	0.73	1	-0.27	-0.28	-0.25
NowCast	-0.75	0.71	-0.56	-0.37	-0.59	-0.37	-0.27	1	0.84	0.95
AQI	-0.74	0.77	-0.61	-0.38	-0.71	-0.4	-0.28	0.84	1	0.79
Raw	-0.72	0.67	-0.54	-0.37	-0.57	-0.35	-0.25	0.95	0.79	1

Classifier	F-measure (Avg.)	ROC Area (Avg.)	Accuracy	Build Time (sec)
Decision Tree	0.913	0.965	91.48%	0.02
Random Forest	0.933	0.993	93.37%	0.86
SVM	0.779	0.837	77.43%	0.41
Kstar	0.883	0.976	88.47%	0.01
Bagging	0.900	0.989	90.52%	0.22
Ensemble Selection	0.890	0.982	89.42%	1.75
Multilayer Perceptron	0.854	0.956	85.79%	2.22

Comparisons of Different Classifiers with Accuracy, F- Measure, ROC Area and Model Build Time

Hazardous	Very Unhealthy	Sensitive	Unhealthy	Moderate Good	Good
24	12	1	0	0	0
4	222	0	3	0	2
2	6	384	6	5	1
0	5	12	237	0	3
9	0	3	0	298	2
0	1	0	0	8	26

Confusion Matrix of Random Forest

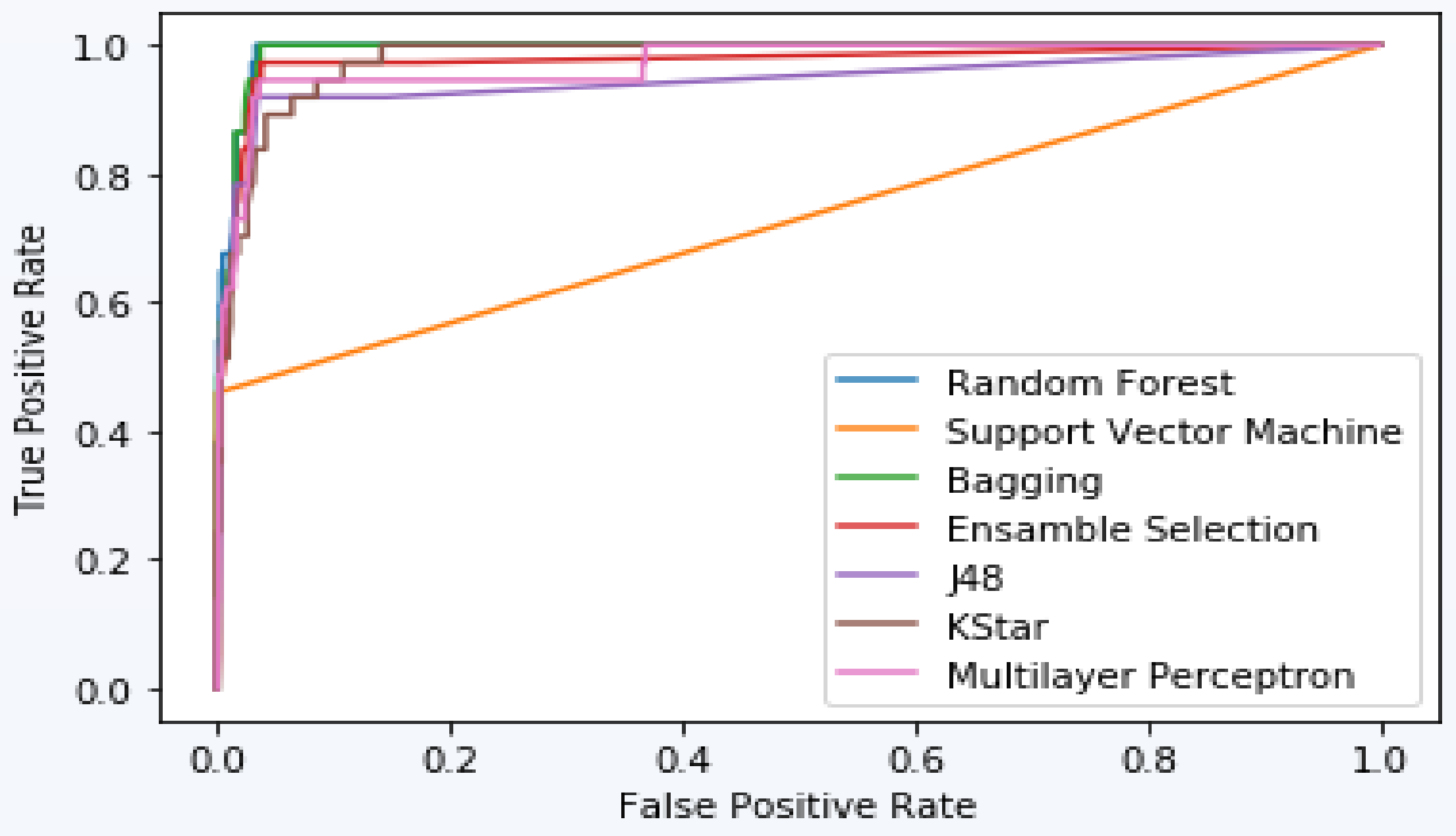


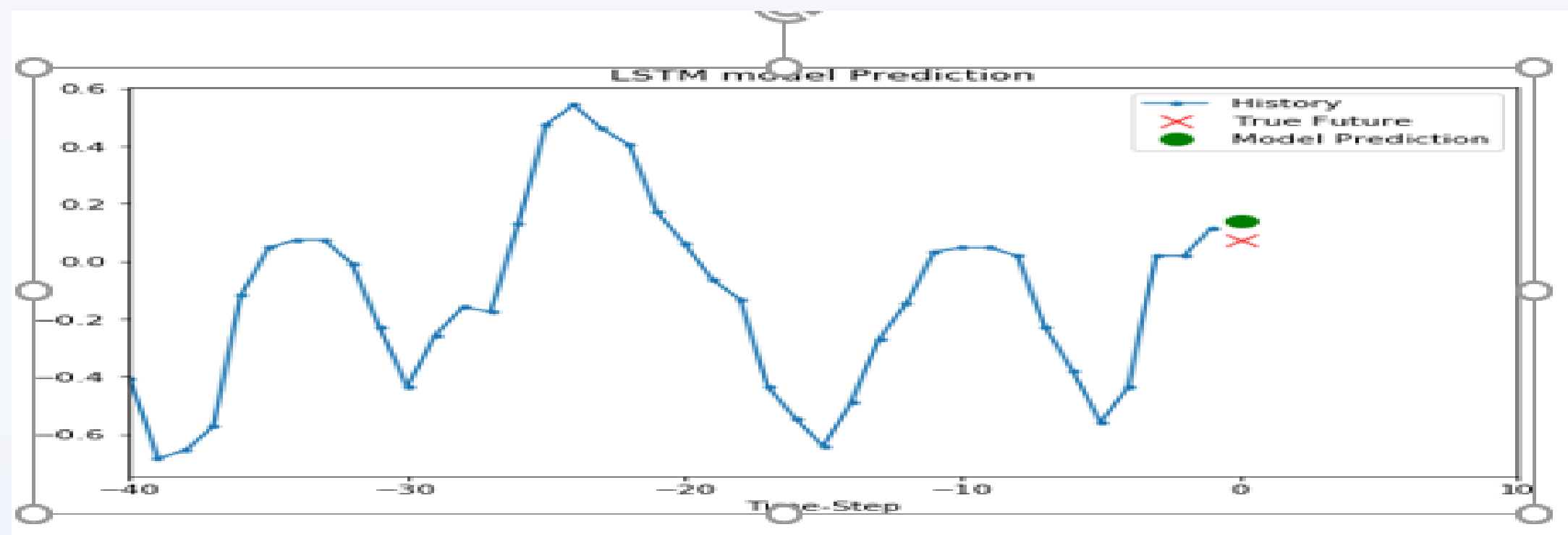
Time series analysis of AQI

Moreover, we have used LSTM to identify time series pattern.

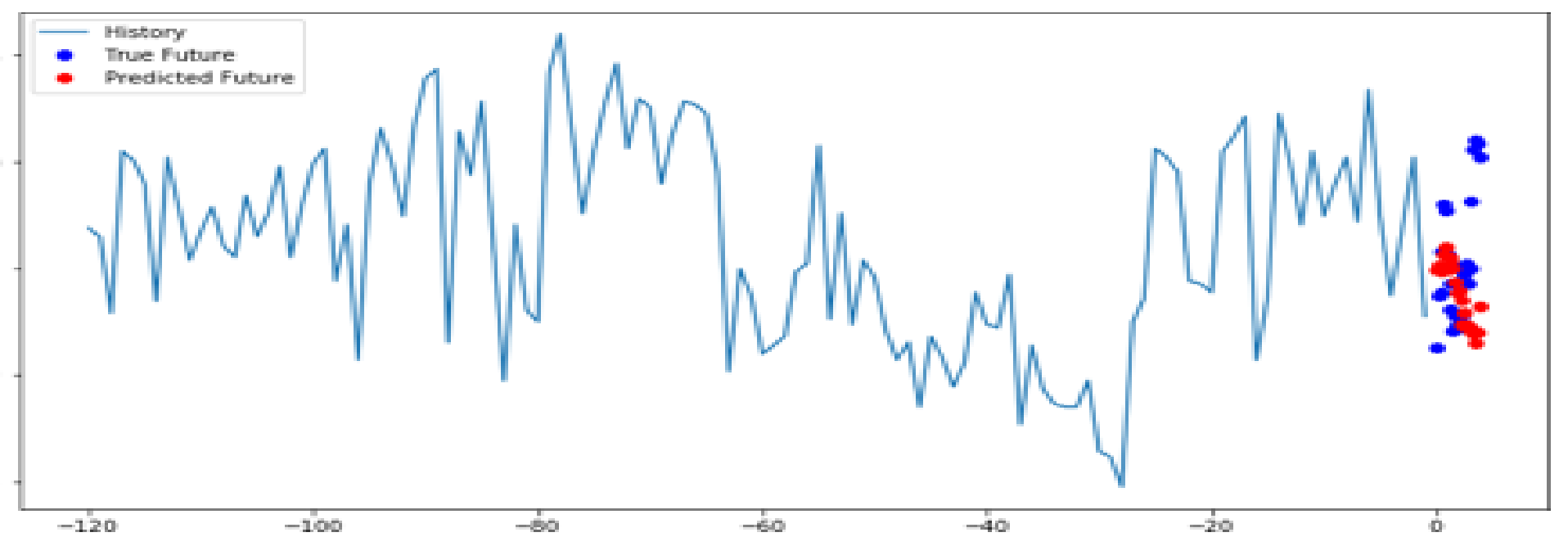
RESULT ANALYSIS

A receiver operating characteristics curve shows the best model. Among all classifiers the Random Forest gives the best result, so we have selected it for further analysis.

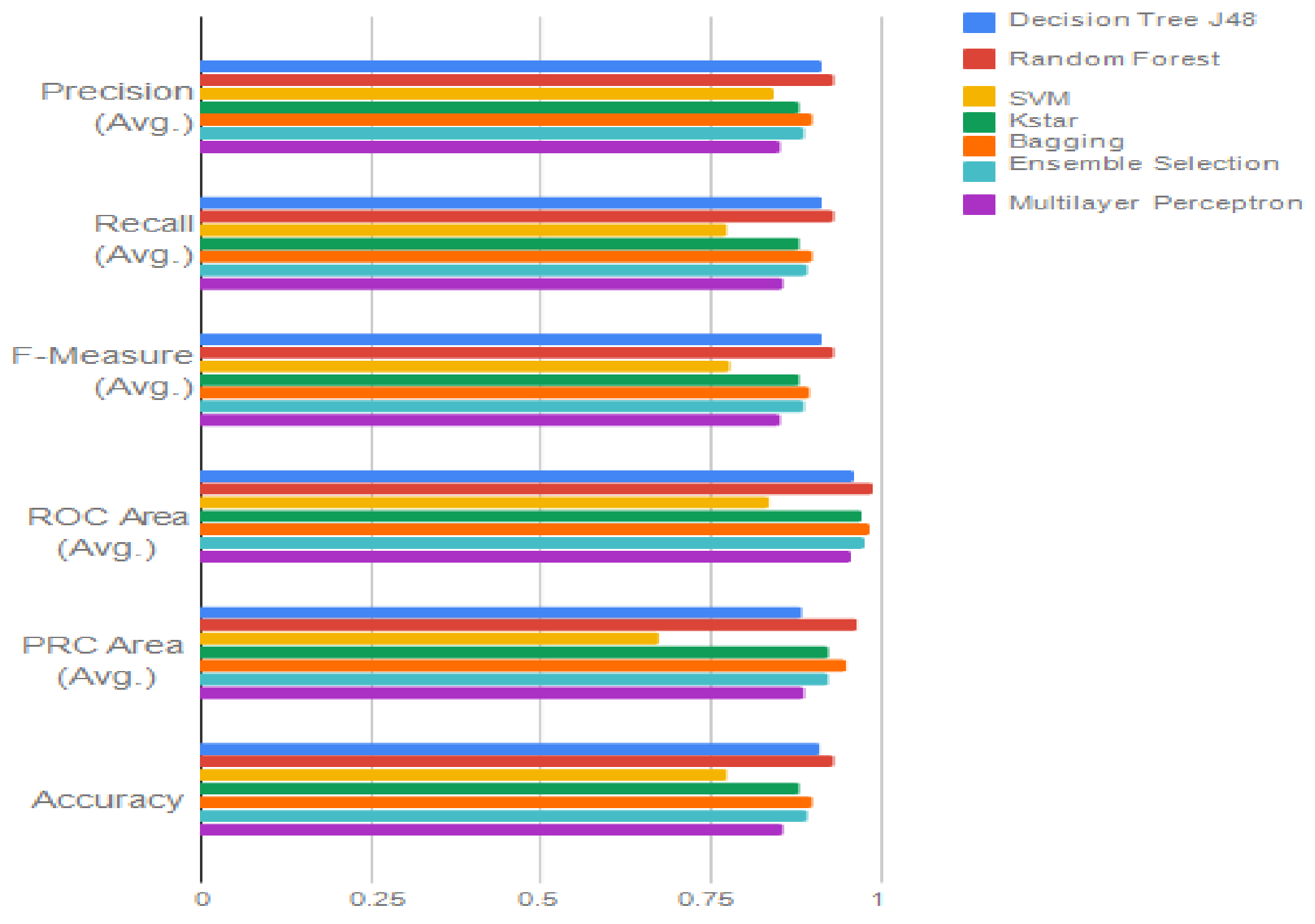




Day wise air pollution prediction using LSTM



Hourly air pollution prediction using LSTM



Model	MSE (Train)	MSE (Valid)
LSTM (Univariate Hourly Forecast)	0.059	0.026
LSTM (Multivariate Daily Forecast)	0.390	0.340

Different Accuracy Metrics

The last table describes the mean square error of our LSTM models.

CONCLUSION & FUTURE WORK

- We have applied different types of machine learning algorithms to predict the levels of pollutants in air based on previous air pollution and weather data.
- Random forest classifier gives the best accuracy of 93.37%.
- LSTM shows the pattern of increasing and decreasing of AQI based on different days and seasons.
- Analyzing this attributes we can more accurately predict the Air Quality Index and daily pollution rate.
- More instance of data and attributes might gives a better analysis of our study on highly polluted city like Dhaka.
- In our future work, we want to use satellite images of different areas of Dhaka city, as well as include a few other major cities of Bangladesh to predict air pollution.
- We want to find patterns and relation how greenery and air pollution is related to each other.
- We will further explore our data-set to find interesting patterns such as the AQI level during the holidays.

REFERENCES

1. Khan, S. Rahman, A. Haque, A. Chen, M. Hammond, S. Djordjevic´, and D. Butler, "Flood damage assessment for Dhaka city, bangladesh," *Flood Risk Management: Science, Policy and Practice: Closing the Gap*, p. 138, 11 2012.

2. "The World Bank." http://data.worldbank.org/indicator/SP.DYN.LE00. FE.IN. Accessed: 2019-12-11.

3. S. Hossain, "Rapid urban growth and poverty in Dhaka city," *Bangladesh e-Journal of Sociology*, vol. 5, 02 2008.

4. M. Rahman and A. Al-Muyeed, "Urban air pollution: a bangladesh perspective," 01 2005.

5. L. Miller and X. Xu, "Ambient pm2.5 human health effects—findings in china and research directions," *Atmosphere*, vol. 9, p. 424, 10 2018.

6. Y. Lin, J. Zou, W. Yang, and C. Q. Li, "A review of recent advances in research on pm2.5 in china," *International Journal of Environmental Research and Public Health*, vol. 15, p. 438, 03 2018.

7. P. K. Hopke, D. D. Cohen, B. A. Begum, S. K. Biswas, B. Ni, G. G. Pandit, M. Santoso, Y.-S. Chung, P. Davy, A. Markwitz, *et al.*, "Urban air quality in the Asian region," *Science of the Total Environment*, vol. 404, no. 1, pp. 103–112, 2008.

8. A. Kurt, B. Gulbagci, F. Karaca, and O. Alagha, "An online air pollution forecasting system using neural networks," *Environment international*, vol. 34, pp. 592–8, 08 2008.

9. P. Raj, "Prediction and optimization of air pollution-a review paper," *International Journal for Research in Applied Science and Engineering Technology*, vol. 7, pp. 3896–3904, 05 2019.

10. G. Kaur, J. Gao, S. Chiao, S. Lu, and G. Xie, "Air quality prediction: Big data and machine learning approaches," *International Journal of Environmental Science and Development*, vol. 9, pp. 8–16, 01 2018.

11. T. Chiwewe and J. Ditsela, "Machine learning based estimation of ozone using spatio-temporal data from air quality monitoring stations," 07 2016.

12. R. Yu, Y. Yang, L. Yang, and G. Han, "Raq–a random forest approach for predicting air quality in urban sensing systems," *Sensors*, vol. 16, p. 86, 01 2016.M. Delavar, A. Gholami, G. Shiran, Y. Rashidi, G. Nakhaeizadeh,

13. Dhaka Weather Datatset, 2016 to 2019, Tutiempo Network, S.L., 2019. [Online]. Available: https://en.tutiempo.net/climate/ws-419230.html. Accessed: 2019-12-11

14. Air Pollution Datatset, 2016 to 2019, National Oceanic and Atmospheric Administration., USA. [Online]. Available: https://www.airnow.gov/index.cfm?action=airnow.global_summary#Bangladesh$Dhaka. Accessed: 2019-12

15. R. Rashu, S. T. Jishan, N. Haq, and M. Rahman, "Implementation of optimum binning, ensemble learning and re-sampling techniques to predict student's performance," *International Journal of Knowledge Engineering and Soft Data Paradigms*, vol. 5, p. 1, 01 2015.

16. Rashad Tanjim , Application of Data Mining Techniques on Air Pollution of Dhaka City, (2020), GitHub repository, https://github.com/RashadTanjim/Application_of_Data_Mining_Techniques_on_Air_Pollution_of_Dhaka_City.git