International Journal of Engineering and Advanced Technology (IJEAT)
ISSN: 2249-8958 (Online), Volume-8 Issue-6S3, September 2019

Air Pollution Prediction using Machine Learning Algorithms

Hanan Aljuaid, Norah Alwabel

Abstract*: Air pollution has a serious impact on human health. It occurs because of natural and man-made factors. The major contribution of this research is that it provides a comparison between different methodologies and techniques of mathematical and machine learning models. The process began with integrating data from different sources at different time interval. The preprocessing phase resulted in two different datasets: one-hour and five-minute datasets. Next, we established a forecasting model for particulate matter $PM_{2.5}$, which is one of the most prevalent air pollutants and its concentration affects air quality. Additionally, we completed a multivariate analysis to predict the $PM_{2.5}$ value and check the effects of other air pollutants, traffic, and weather. The algorithms used are support vector regression, k-nearest neighbors and decision tree models. The results showed that for the one-hour data set, of the three algorithms, support vector regression has the least root-mean-square error (RMSE) and also lowest value in mean absolute error (MAE). Alternatively, for the five-minute dataset, we found that the auto-regression model showed the least RMSE and MAE; however, this model only predicts short-term $PM_{2.5}$.

***Index Terms*: Air, Pollution, Machine, Learning, Prediction.**

I. INTRODUCTION

A. The problem: what is air pollution?

Air pollution occurs when air mixes with other components called air pollutants. Pollutants include sulfur dioxide, nitrogen dioxide, carbon monoxide, and particulate matter $PM_{2.5}$ [1]. Outdoor air can be polluted with natural or man-made sources [2]. PM is a mixture of small particles and liquid droplets that are mainly from industry or motor vehicle exhaust dust. The size of particulate is measured in micrometers; for example, PM with a particulate size of 2.5 micrometers is written as $PM_{2.5}$.

Sulfur dioxide is produced when sulphur fuels are burned during industrial processes, while nitrogen dioxide is usually formed when fuel is burned at high temperature [1]. A common source of sulfur dioxide and nitrogen dioxide is motor vehicle exhaust. Carbon monoxide is colourless and is formed when the carbon fuel does not completely burn [1].

Manuscript published on 30 September 2019.
* Correspondence Author (s)
Hanan Aljuaid, Princess Nourah bint Abdulrahman University, Riyadh, Saudi Arabia.
Norah Alwabel, 2Princess Nourah bint Abdulrahman University Riyadh, Saudi Arabia.

B. Air quality measurement

Governments have adopted regulations and standards to control and monitor air quality. These standards focus on the concentration of specific air pollutants allowed in the air. A useful way to track air pollution against national standards is a measure called the air quality index [3]. European cities, for example, use a standard called the common air quality index (CAQI), which is based on the $PM_{2.5}$ value and classifies air according to the following levels: very high, high, medium, low and very low [4], [5]. Table I shows air pollution levels based on $PM_{2.5}$ value for one-hour measurement [4], [5].

Table I. Common air quality index

Level	Rang (per hour)
Very low	0-15
Low	15-30
Medium	30-55
High	55-110
Very high	More than 110

The United State use another standard and includes five major air pollutants to calculate air quality index (AQI). AQI classifies pollution levels as follows: good, moderate, and unhealthy for sensitive groups, unhealthy, very unhealthy and hazardous [3], [6]. Table II illustrates the AQI value and the associated effect level [3], [6].

Table II. US pollution level and AQI value

Level	AQI value
Good	0-50
Moderate	51-100
Unhealthy for sensitive groups	101-150
Unhealthy	151-200
Very Unhealthy	201-300
Hazardous	301-500

C. Health effects

Air pollution can cause disruptions in human enzyme levels. In 2008, Farhat and other researchers measured enzymes levels in Olympic athletes: before, during and after the Beijing Olympics. The results showed that the level of enzymes changed according to air pollution [7] In 2017, Taylan's research found that high levels of ozone are a major cause of air pollution [8].

Exposure to air pollution can result in a wide range of short- and long-term health effects that depend on the pollutant, concentration of that pollutant, the length of the exposure, and the affected individual's demographic and health condition.

The effects can range from minor symptoms like irritation of eye, nose and throat to more severe conditions like heart and lung disease or cancer [9], [10]. A study done in United States, for example, illustrated that mitigation in the exposure of $PM_{2.5}$ can improve life expectancy [11], [12].

A scientific question exists: How can air pollution be predicted using support vector regression, decision tree, K-nearest neighbors and auto-regression, and which one minimises the error in prediction the $PM_{2.5}$ value?

In this paper, we will discuss how machine learning algorithms can predict air pollution from data gathered via Internet of Things (IoT) devices.

II. Literature Review

Taylan (2017) conducted a study in Jeddah, where he used back-propagation neural networks to predict the ozone concentration in the air. The network had three layers: hidden, input and output layers. It used the nitrogen level, air pressure and weather data to predict the pollution levels in the air. The input was classified according to its level (it converted the values to levels): very low, low, normal, high and very high. This model provided equations to predict the level of air pollution, and the results showed a negative correlation between air pressure and temperature [8].

Research done by Blagojevi, Sucurovic and Papic (2018) to predict air pollution in the Moravica district. They sampled air at altitudes from 1.5m to 10m, far from the source of pollution. Measuring points were based on location, population density, and landscape and weather conditions. Soot levels were measured using nitrogen oxide and sulphur dioxide levels, while the PM concentration was measured by analysing soluble and insoluble particles in the air [13].

The researchers developed a back-propagation network with three layers: the hidden, input and output layers. Input layers contained input data, including the year, municipality name and measuring site as additional attributes. The hidden layer contained neurons for receiving the input and processing the output, while the additional attribute for this layer was the air pollution level [13].

The finding of this study showed a mean square error (MSE) of 0.0635. The model used to test the algorithm was 72.73% out of 82%. Right cases displayed were 82%. From the model, the deviation was 9.27, which represented 9.2%. The conclusion was that the neurons developed were efficient. After the successful training of the algorithm, new values for sulphur oxide, nitrogen oxide, particulate matter and soot (20.24, 34, 54.3 and 190, respectively) were put into the system. The chance of obtaining one of these data values was 97.4%. If the algorithm gave a result of 1, then the air was not polluted [13].

Yang, Deng and Wang (2018) used a support vector regression model to predict the concentration of $PM_{2.5}$ in the air. They conducted this study in Beijing and data were collected between March 2014 and April 2014. The inputs used to predict $PM_{2.5}$ which is a continues variable, were temperature, humidity and wind direction and force. The authors used a support vector regression algorithm, and they clustered the data into sub-parts based on location. Thus, the prediction of hourly $PM_{2.5}$ was based on space-time support vector regression. Fig. 1 shows the details of the framework used by the researchers [14].

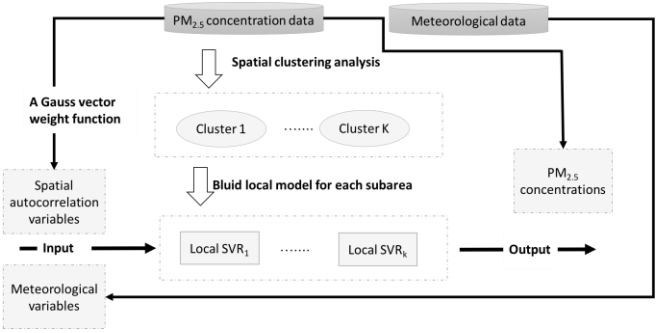

Fig. 1. Framework used by researchers to predict hourly $PM_{2.5}$

Bougoudis, Demertzis and Iliadis (2016) developed a hybrid machine learning technique called HISYCOL, which consists of three steps. The first step is to segment the dataset into clusters. The next step is to run a linear regression for each segment in addition to a random forest algorithm. The final step involves obtaining the root mean square error (RMSE) and using each segment as input vectors for a Mamdani fuzzy inference system [15]. A study done on the short-term prediction of the concentration of $PM_{2.5}$ used time series analysis divided into two parts. The first part identified factors affecting $PM_{2.5}$ using multivariate statistical analysis and a back-propagation neural network to identify machine learning correlation. The attributes used included physical data, such as temperature, wind speed and average rainfall, in addition to social media data about Beijing. The second part of the research created a short-term prediction of the $PM_{2.5}$ value using the auto-regressive integrated moving average. Fig. 2 shows framework details used by the researchers [16]. The findings indicated that wind speed, carbon dioxide, nitrogen dioxide, PM_{10} and social media data were highly correlated with $PM_{2.5}$. Additionally, the study showed that when predicting $PM_{2.5}$ between August 2014 and September 2014, the RMSE was 6.76 [16].

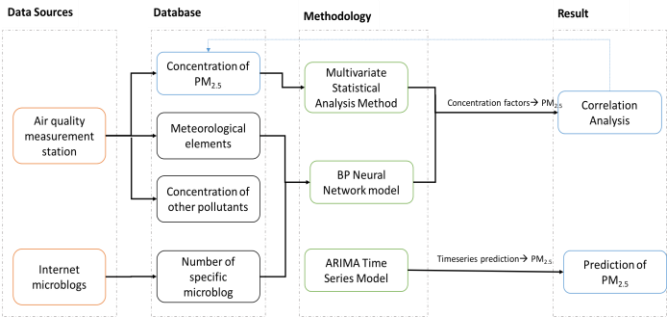

Fig. 2. Framework used to study the relevance of and predict air pollution

III. Data Set: CityPulse

CityPulse is a public dataset that was collected from sensors and transformed using data modelling. CityPulse data processing creates information for analysis and joins different datasets. CityPulse includes around 8 million records and consists of pollution data, traffic data, weather, cultural event, library event and parking data [17], [18]. The CityPulse database includes around 8 million records and consists of traffic data, pollution data, weather, cultural event, library event and parking data.

The aim of this study was to focus on data related to Aarhus, Denmark. The Citypulse team collect data from 449 Internet of Thing (IoT) devices to formalise air and traffic data sets [17], [18].

We completed a preliminary analysis on the air and traffic data for a sample device. Air data consisted of the following measurement data: ozone, particulate matter, carbon monoxide, sulfur dioxide and nitrogen dioxide, in addition to location data: longitude, latitude and time stamp for the recording.

The preliminary analysis for air data showed that the time stamp is monotonic, which means that the difference between each recording is equal; observation were recorded each five minutes between August 2014 and October 2014 with no missing data. All the measurement data fall between 15 and 215, and all measurement units are micrograms per cubic meter (mg/m^3). The distribution of data is reflected in Fig. 3 while Fig. 4 depicts the $PM_{2.5}$ value throughout the experiment.

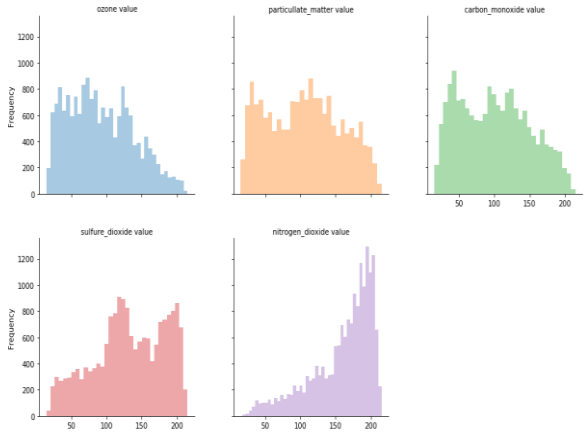

Fig. 3. Distribution of air attributes values

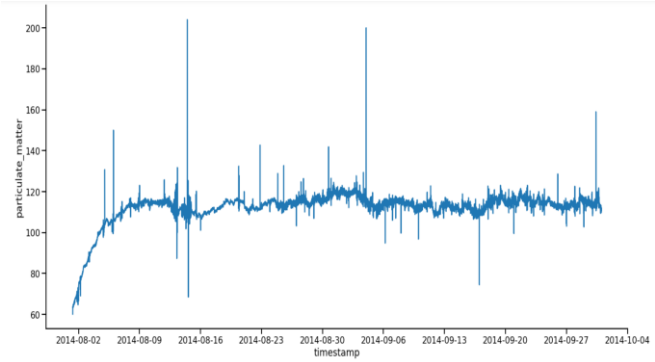

Fig. 4. $PM_{2.5}$ value from Aug 2014 to Oct 2014

After finalizing the preliminary analysis, the air and traffic data grouped by each hour based on assumption the mean of recording within one hour is representing the actual recording (5 minutes). Grouping data per hour were used by some researched previously to study the hour concentration of $PM_{2.5}$ value [19]. Moreover, other study mine data produced from IoT devices reduce the data set from 400 million to 20 million [20]. This demonstrates that data reduction is paramount to eliminating noise and producing clean data or if we need a specific value, as in our case.

IV. METHODOLOGY

The methodology used in this research includes four algorithms applied on two different datasets, one-hour and five-minute datasets. Then, we used evaluation metrics to compare these algorithms using both datasets.

The predictor used in this study $PM_{2.5}$ which is a continues variable. The process included multivariate and univariate algorithms. Multivariate algorithms include support vector regression, decision tree and K-nearest neighbor. The input used by these algorithms are carbon_monoxide, ozone, nitrogen_dioxide, sulfure_dioxide and DURATION_IN_SEC and the output is $PM_{2.5}$.

On other hand, univariate analysis was done using auto-regression. This allowed us to study the value of past results of $PM_{2.5}$ and predict the future values based on past values. Fig. 5 illustrates the whole process.

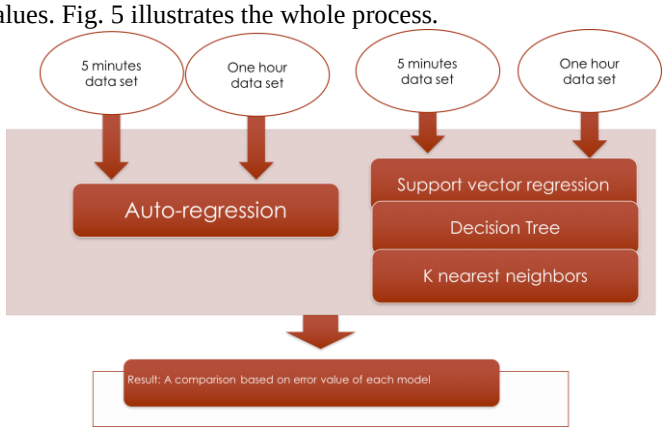

Fig. 5. A detailed view of our process for this experiment.

The measurement used to evaluate each algorithm was based on the error of predicted value. We used root mean squared error (RMSE) and mean absolute error (MAE).

RMSE is the summation of the squared of error for each case (difference between the predicted and real value). Then, it is divided by the number of cases, resulting the mean squared error and take the root [21] and the formula described in Eq.1 where y represents the real value, $\hat{y}$ represents the prediction made by the model and n represents the number of cases

$$RMSE = \sqrt{\frac{\sum (y - \hat{y})^2}{n}} \quad (1)$$

MAE is calculated by summing the absolute values of the errors and then dividing that by the number of cases [21] It is defined by Eq. 2.

$$MAE = \frac{\sum \left| \left(y - \hat{y} \right) \right|}{n} \quad (2)$$

Although that Willmott and Matsuura(2005) claimed that RMSE cannot be used as measure of average error made by the model and MAE can be used to evaluate the performance of model [22], but chai and Draxter(2014) dispute that and present RMSE as a valid measure of model performance[23]

V. RESULTS

The result of the pre-processing phase is that we have two datasets: one-hour with around 643 thousand recordings and a five-minute dataset with around 7 million recordings. This section provides a comparison of the results of the used algorithms (support vector regression, decision tree, K nearest neighbors and auto-regression). These results are based on training data for the first seven weeks and test data for the final week.

A. One-hour dataset results

RSME values, as demonstrated in Fig. 5, indicated that support vector regression has the lowest value: 1.9. Predictions from the auto-regression model have an RMSE value of 2.3 and the predictions using k-nearest neighbors have a value of 3.4. However, the decision tree model give greatest value: 5.3, which is an RMSE increase of 178.95% from the support vector regression model.

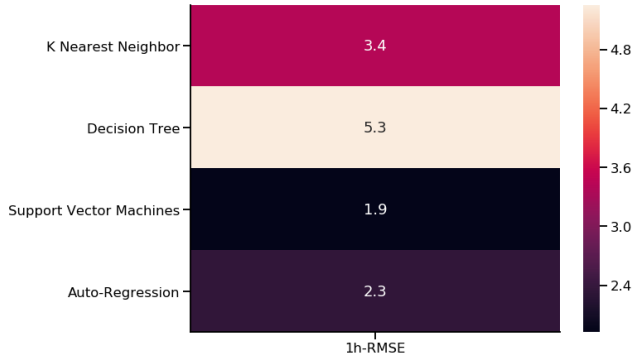

Fig. 6. RMSE of one-hour models

Similarly, the support vector regression model has the lowest MAE value (1.6), as shown in Fig. 7 The other models (auto-regression, k-nearest neighbors and decision tree) had the following results for MAE: values of 1.9, 2.7 and 4.3, respectively.

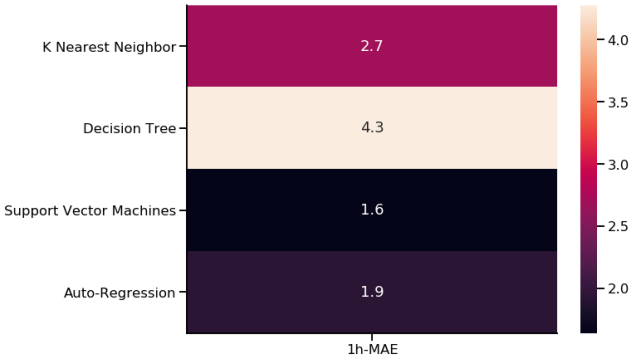

Fig. 7. MAE of one-hour models

B. Five-minute dataset results

These RMSE results revealed that the auto-regression model performed best with a value of 2.4, while the k-nearest neighbors model showed a value of 3.9. Interestingly, we found that decision tree and support vector regression models showed identical values of 4.3 (see Fig. 8).

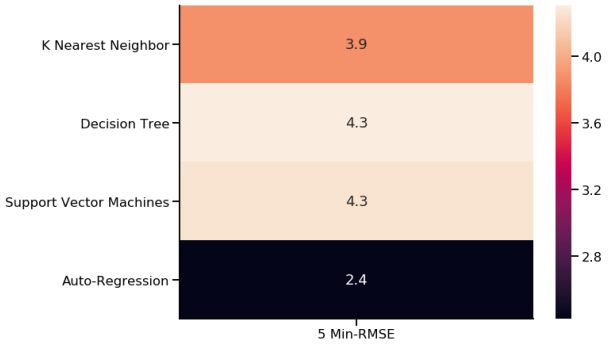

Fig. 8. RMSE results for five-minute models

Nevertheless, we discovered that the auto-regression model gave an MAE value of 2, while the k-nearest neighbors model showed a value of 2.3 (see Fig. 9). Moreover, the MAE value for support vector regression is the highest value at 3, and decision tree is performed only slightly better at 2.8.

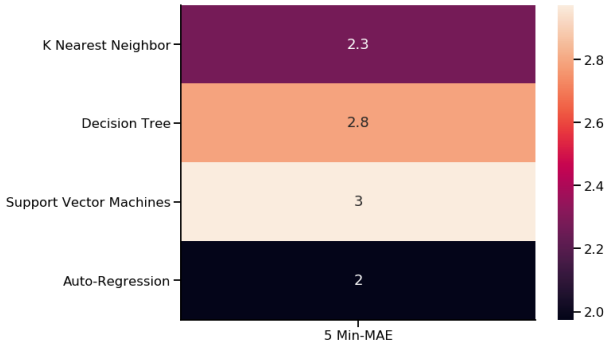

Fig. 9. MAE results for five-minute models

VI. CONCLUSION AND DISCUSSION

In general, we found that the one-hour dataset had better results than the five-minute dataset, regardless of the algorithm used. Surprisingly, we discovered that support vector regression provided the best results for the five-minute dataset in terms of RMSE and MAE and the worst for the one-hour dataset. The main differences between these two datasets are the number of records and weather data that were one-hour data sets but not integrated with the five-minute data sets.

As we saw previously, there is debate between different researchers as to whether RMSE is a good measure for the error made by models or not. However, our results showed that RMSE and MAE sort all examined models the same with difference in values; overall, MAE resulted in lower values than RMSE for each model.

Finally, we found that auto-regression gave the best results for five-minute data and second best model for one-hour datasets.

REFERENCES

1. K. K. Lee, M. R. Miller, and A. S. Shah, "Air pollution and stroke," *Journal of stroke*, vol. 20, no. 1, p. 2, 2018.

2. Y. Lin, L. Zhao, H. Li, and Y. Sun, "Air quality forecasting based on cloud model granulation," *EURASIP Journal on Wireless Communications and Networking*, vol. 2018, no. 1, p. 106, 2018.
3. J. Bachmann, "Will the circle be unbroken: a history of the us national ambient air quality standards," *Journal of the Air & Waste Management Association*, vol. 57, no. 6, pp. 652–697, 2007.
4. S. van den Elshout, K. Léger, and H. Heich, "Caqi common air quality index—update with pm2. 5 and sensitivity analysis," *Science of the Total Environment*, vol. 488,pp. 461–468, 2014.
5. "Air quality index (aqi) basics," 8 2016.
6. Air Quality in Europe 8 2012.
7. Z. Farhat, R. W. Browne, M. R. Bonner, L. Tian, F. Deng, M. Swanson, and L. Mu, "How do glutathione antioxidant enzymes and total antioxidant status respond to air pollution exposure?," *Environment international*, vol. 112, pp. 287–293, 2018.
8. O. Taylan, "Modelling and analysis of ozone concentration by artificial intelligent techniques for estimating air quality," *Atmospheric Environment*, vol. 150, pp. 356–365, 2017.
9. J. Lelieveld, J. S. Evans, M. Fnais, D. Giannadaki, and A. Pozzer, "The contribution of outdoor air pollution sources to premature mortality on a global scale," *Nature*, vol. 525, no. 7569, p. 367, 2015.
10. Daiber, S. Steven, A. Weber, V. V. Shuvaev, V. R. Muzykantov, I. Laher, H. Li, S. Lamas, and T. Münzel, "Targeting vascular (endothelial) dysfunction," *British journal of pharmacology*, vol. 174, no. 12, pp. 1591–1619, 2017.
11. C. A. Pope III, M. Ezzati, and D. W. Dockery, "Fineparticulate air pollution and life expectancy in the united states," *New England Journal of Medicine*, vol. 360,no. 4, pp. 376–386, 2009.
12. T. Münzel, M. Sørensen, T. Gori, F. P. Schmidt, X. Rao, F. R. Brook, L. C. Chen, R. D. Brook, and S. Rajagopalan, "Environmental stressors and cardiometabolic disease: Part ii–mechanistic insights," *European heart journal*, vol. 38, no. 8, pp. 557–564, 2016.
13. M. BLAGOJEVI, M. I. PAPI, MILOand VUJI, and M. UROVI, "Artificial neural network model for predicting air pollution. case study of the moravica district, serbia," *Environment Protection Engineering*, vol. 44, no. 1, 2018.
14. W. Yang, M. Deng, F. Xu, and H. Wang, "Prediction of hourly pm2. 5 using a space-time support vector regression model," *Atmospheric Environment*, vol. 181, pp. 12–19, 2018.
15. 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 Computing and Applications*, vol. 27, no. 5, pp. 1191–1206, 2016.
16. X. Ni, H. Huang, and W. Du, "Relevance analysis and short-term prediction of pm2. 5 concentrations in Beijing based on multi-source data," *Atmospheric Environment*, vol. 150, pp. 146–161, 2017.
17. K. R. Malik, Y. Sam, M. Hussain, and A. Abuarqoub, "A methodology for real-time data sustainability in smart city: Towards inferencing and analytics for big-data," *Sustainable Cities and Society*, vol. 39, pp. 548–556, 2018.
18. M. I. Ali, F. Gao, and A. Mileo, "Citybench: A configurable benchmark to evaluate rsp engines using smart city datasets," in *International Semantic Web Conference*, pp. 374–389, Springer, 2015.
19. A. T. DeGaetano and O. M. Doherty, "Temporal, spatial and meteorological variations in hourly pm2. 5 concentration extremes in new york city," *Atmospheric Environment*, vol. 38, no. 11, pp. 1547–1558, 2004.
20. A. Yassine, S. Singh, and A. Alamri, "Mining human activity patterns from smart home big data for health care applications," *IEEE Access*, vol. 5, pp. 13131–13141, 2017.
21. J. S. Armstrong and F. Collopy, "Error measures for generalizing about forecasting methods: Empirical comparisons," *International journal of forecasting*, vol. 8, no. 1, pp. 69–80, 1992.
22. C. J. Willmott and K. Matsuura, "Advantages of the mean absolute error (mae) over the root mean square error (rmse) in assessing average model performance," *Climate research*, vol. 30, no. 1, pp. 79–82, 2005.
23. T. Chai and R. R. Draxler, "Root mean square error (rmse) or mean absolute error (mae)?–arguments against avoiding rmse in the literature," *Geoscientific model development*, vol. 7, no. 3, pp. 1247–1250, 2014.

AUTHORS PROFILE

Hanan Aljuaid is an Assistance Professor in the Department of Computer Science at Princess Norah bint Abdurrahman University PNU University in Riyadh, where she has been since 2010. From 2003 to 2010 she worked at Taif University in Saudia Arabia.From 2004 to 2006 she served as Department Chair at Taif University. From 2014 to 2016 she served as Vice Dean of Educational Affairs, and From 2016 to 2018 she served as Vice-Dean of Electronic Learning, Deanship of Electronic and Distance Learning, PNU University.. From 2014 to 2016 she served as scientific community chair of International Conference on Cloud Computing (ICCC15), PNU University. She received a B.S. from KAU University in 2002, and an M.S. from the UTM University. He received his Ph.D. in Computer Science from the UTM University in 2014. Her research interests span both computer vision and NLP. Much of his work has been on improving the understanding, design, and performance of pattern recognition, mainly through the application of data mining. She has given numerous invited talks and tutorials.

Norah Alwabel is a master student in Master of Science in computing data analytics major in Princess Norah bint Abdurrahman University in Riyadh with collaboration with Duplin City University. Norah Alwabel currently working at King Abdullah bin Abdulaziz university hospital at Princess Norah bint Abdurrahman University since 2014. She received a B.S from King Saud University in 2012 at Riyadh, in information and technology degree. On the same year she joined Saudi Arabia ministry of health and work in information and technology department. She was responsible for patient information data base and medical devices integration of all hospitals under ministry of health.