Prediction of the Level of Air Pollution During Wildfires Using Machine Learning Classification Methods

Syed Mohammed Khalid, Raini Hassan
Department of Computer Science, Kulliyyah of ICT, International Islamic University
smkhalid1998@gmail.com, hrai@iium.edu.my

Abstract— The recent increase of forest fires due to agricultural field burning in the South East Asian region has led to haze episodes in Malaysia which contributed to the increasing number of hospital visits for treatments related to respiratory diseases. With the increase of air pollution, it becomes a necessity to attempt at investigating and predicting the air pollution levels, which would in turn which would lead to proper strategies so untimely effects to human health can be kept at a minimum. The Air Pollutant Index (API) is used to identify and classify the ambient air quality status. However, the lack of ground air quality monitors which compute the API generally leads to unreliable warning information. Recent studies indicate that data retrieved from remote sensing satellites is now an emerging alternative for air quality prediction at the ground level. Hence, this research aims to use satellite-based data to predict the air quality of East Malaysian cities with the help of different Machine Learning classification algorithms. Aerosol optical data, meteorological data and fire data were collected from different satellite sources. Two algorithms were selected and implemented, and they are Random Forest and Gradient Boosting. When trained and validated, both algorithms performed reasonably well with an accuracy 0.89 and 0.85 respectively, for the city of Kuching, Sarawak, Malaysia.

Keywords— air quality monitoring systems, air pollutant index, aerosol optical depth, machine learning classification algorithms, gradient boosting, random forest, sustainable development goals.

I. Introduction

Air pollution has caused major problems to the environment in Malaysia. One of the key sources of Air pollution in Malaysia in trans-boundary pollution from neighbouring countries [1]. Forest burring and deforestation activities occurring in Sumatra and Kalimantan, Indonesia can be recognized as the contributing factors to high intensity combustions that lead to transboundary haze in Malaysia [2]. As a consequence of the repeated haze episodes, the Malaysian government established the Malaysian Air Quality Guidelines, the Air Pollution Index, and the Haze Action Plan to help improve air quality. Henceforth air quality monitoring became a part of the initial strategy to combat the effects of air pollution in Malaysia. The Air Pollutant Index is used to help in identifying and classifying the air quality in Malaysia based on the possible health implications to the public, if the API value raises then the level of danger or the status of API is raised to "Unhealthy" or "Extremely Unhealthy", The API values provide a substantial method to perform the determination and evaluation of the changes in pollution levels. In this project the aim is to be able to predict exposure of a population to pollution or the level of danger with regards to API, through the help of satellite-based data. This will be done with the help of the many different classification algorithms that have come to existence over the years. Generally, the API is measured by Air Quality Monitoring Systems which compute the exposure of a population to air pollution, once it's computed it's made public as to whether it's "Good" or "Hazardous", however many regions which do not have any Air Quality Monitoring Systems (AQMS) in their respective locations may not get information about the health hazards in their region. During the most recent episode being the 2019 Kalimantan fires, which saw more than 857,756 hectares (2.12 million acres) of land burned [3], residents of Miri, Sarawak raised their concerns concerning the lack of air quality monitoring in the division, as their town must depend on API readings from the nearby Samarahan air quality monitoring station [4]. The use of satellite-based data with help of Machine Learning could potentially be a means for these regions to know the level of air pollution where they live.

Since there isn't a cost-efficient alternative to AQMS, the possibility of being able to use satellite-based data comes as a relief. It would be cost efficient in the sense that satellites already exist and are continuously monitoring every region of our planet and sending back data to earth, most of which are available freely in the public domain. The proper use of this data with the combination Machine Learning algorithms can potentially save the cost of using AQMS and will ensure every region knows the levels and dangers of pollution in the respective region. The 2030 Agenda for Sustainable

Development which in 2015 was approved by all the member states of the United Nations, which included a blueprint encouraging peace and prosperity for people and the planet, today and in the future. Efforts to combat air pollution will contribute to one of the goals and targets of the United Nations SDG (Sustainable Development Goals) which would be SDG 3 (good health and well-being), SDG target 7.2 on access to clean energy in the home and SDG target 11.6 on air quality in cities [5].

II. Related work

In the following section, a brief background of the related work is presented. Over the years there has been some significant work by researchers in relation to predicting air pollution levels in many regions. Below we discuss some of the reviewed papers where some focus on prediction of air pollution levels during wildfires while others focus on air pollution in general circumstances.

Currently, three different approaches are used to predict air pollution levels: chemical transport, statistical models, and Machine Learning. Chemical transport with the combination of Atmospheric Dispersion Modelling is one of the methods used, an example of it is WRF-Chem model. These models are used to predict atmospheric pollution; however, they lack in accuracy in some cases. Statistical models are generally based on single variable linear regression and they show a decent correlation between different parameters and the level of air pollution [6].

Unlike the statistical methods, using Machine Learning method to predict the air pollution level can make it more effective as one can use several parameters in a single model. Some of the most popular Machine Learning algorithms used to forecast air pollution from different parameters are Artificial Neural Network based algorithms. Many recent studies show that the Machine Learning approach in most cases outperforms the other two methods for forecasting air pollution. This is the main reason why it has recently been increasingly used to predict air quality [7].

If we were to look at the problem in recent times, researchers such as Owusu-Akyaw, Li and Moran [8] conducted their study as recently as last year and, in their study, they targeted the temporal pollution levels in Californian counties during the 2018 California wildfires. Their study was more focused on how wildfires directly contribute to the increase and decrease in pollution levels. Their main predictor variables were distance from the fire, direction from the centre of fire and size of fires. In their preliminary models they used two different algorithms, RANSAC and Linear Regression, the RANSAC model performed slightly better with a MSE of 38.14, they hence selected that algorithm. They later achieved better results with the addition of more predictor variables. However, their study only depended on fire and wind pattern data, there was no use of other factors with regards to spatiotemporal environment.

Wu, Winer and Delfino [7] have looked at predicting PM (particulate matter) concentrations levels at zip code levels during the wildfires in Southern California. The one issue they were trying to address was that the PM concentrations were only recorded on every 3rd and 6th day of a week, hence it became necessary to perform spatial interpolation for the missing data. The methods used for spatial interpolation were inverse distance weighting (IDW), kriging or cokriging methods for the non-fire periods. They noticed Kriging was not particularly better than IDW in most cases, this could be due to the limited number of monitoring stations. IDW and Kriging did not work during the fires as the fire and smoke creates highly heterogeneous pollution surfaces.

Reid et al. [9] must be one of the closest to what we are trying to achieve, in their work they focused on Machine Learning techniques to predict particulate matter (PM 2.5) for a California wildfire. Their work focuses on trying to find the optimal algorithm among Generalized Boosting Model (GBM), Random Forest, Bagged Trees, Elastic Net Regression, Multivariate Adaptive Regression Splines, Lasso Regression, Support vector machines, Gaussian processes and Generalized linear model using a 10-fold cross validation. Among the data used for their work, there was fire data, meteorological data and other spatiotemporal variables. In conclusion the best performing model for them was the with 29 predictor variables the Generalized boosting model which achieved a CV-R2 value of 0.803.

A study conducted by Sukitpaneenit and Oanh [10] proposed to understand how satellite data can be used to monitor carbon monoxide (CO) and particulate matter (PM) in Northern Thailand when forest fires occur, the authors acknowledge that forest fires are a one of the significant causes of air pollution. The target variables being the CO and PM concentrations were obtained from monitoring stations across the northern region of Thailand, and the main predictor variables were data obtained from the Moderate Resolution Imaging Spectroradiometer Satellite (MODIS), which provided the Measurement of Pollution in the Troposphere (MOPPIT), Aerosol Optical Depth (AOD) and MODIS fire hotspots data. Their results showed that correlations between the ground-monitored CO and PM, respectively, with satellite monitoring data were quite resonable in comparison with earlier studies conducted for other regions around the world. AOD and PM10 were generally better correlated (R=0.50–0.73) than MOPITT CO and ground monitored CO which correlated at (R=0.36–0.71).

Like Sukitpaneenit and Oanh, Kanabkaew [11] focused on the Thailand region. He indicates that the constant reoccurrence of forest fires Chiangmai and northern Thailand is a matter of concern, he acknowledges that lack of ground monitoring systems may lead to unreliability for warning information, hence he suggests that satellite remote sensing is could potentially be a good way of predicting air quality at the ground level. His study is focused on coming up with a satellite model to predict PM concentrations using satellite data. AOD data was collected from MODIS- Terra platform and ground level air quality were retrieved from ground stations. Two models were implemented using the data, the first model being single linear regression and the second one being multiple linear regression. The second model gave a slightly better performance with R2of 0.77 and 0.71, respectively for PM2.5 and PM10. In order to check the legitimacy of the model, the obtained regression equation was then used with the smog data over Chiangmai in March 2007. The model exhibited a decent performance with an R2 of 0.74.

The next few papers reviewed were studies focused on the Malaysian air pollution levels. Ng and Awang [12] in their paper used Multiple linear regression (MLR) and regression with time series error (RSTE) models in order to predict the concentrations of PM10 in Peninsular Malaysia. The predictor variables used by them were hourly temperature, humidity, wind speed and direction. If any missing values occurred, they used linear interpolation to perform the imputation. The validation of the model was done with six different parameters, namely root mean squared error (RMSE), mean absolute percentage error (MAPE), maximum absolute percentage error (maxAPE), fractional bias (FB) percent bias (PBIAS) and mean absolute error (MAE). The evaluation results suggested that MLR and RTSE are similar in their predictive ability. However, the MLR was identified as the best performing model with low values of RMSE, MAPE and maxAPE. At the same time RTSE was also identified as a good model when compared to the two MLR models in terms of MAE, FB and PBIAS. The authors concluded that the RTSE model did not particularly show any superiority over the simple MLR model.

Siew and Chin [13], they started off by highlighting the danger of haze and how much it has affected Malaysia over the years. The objective of their project was to use time series models to forecast the API in Shah Alam, Selangor. The data they used in their study consisted of 70 monthly observations of API (from March 1998 to December 2003). The time series models that they considered for selection were the Integrated Autoregressive Moving Average (ARIMA) and the Integrated Long Memory Model (ARFIMA) models. Through model evaluation it came to their attention that the integrated ARFIMA model is a better performing model as it had a very low MAPE value. However, they noticed that the actual value of May 2003 falls outside the 95% forecast interval, this could be possibly due to emissions from mobile sources.

Azid and Juhair [14] in their research focused on recognizing the pattern of Malaysian air quality based on the data obtained from the Malaysian Department of Environment (DOE). The collected data consisted of 8 air quality parameters from ten monitoring stations in Malaysia for 7 years (2005–2011). They made use of Principal component analysis (PCA) to identify the different sources of pollution in the locations in study. Both PCA and artificial neural networks (ANN) were modelled in order to determine their ability to predict the air pollutant index (API). The PCA-ANN models performed slightly better in determining the API with less variables, with R2 and root mean square error (RMSE) values of 0.618 and 10.017, respectively.

From the literature review conducted above, it can be concluded that most of the research has been done in other regions, very little research has been done on the Malaysian region. Even the few papers which worked on the Malaysian region did not use the combination of fire data, meteorological data and AOD data. In this research, the aim is to entirely use satellite-based data to evaluate performances of Machine Learning models in predicting the exposure of pollution to different regions before, during and after wildfires. One might also notice that all the previous studies have relied on regression algorithms for similar objectives. In this project the aim is to try to convert this into a classification problem in order to achieve better precision and accuracy and this will be done with help through data binning. Often the question arises as to how to pick the classes while performing the binning? This is made easier with the classes already being specified and standardized (Table I) by the Malaysian Ambient Air Quality Standard (MAAQS) [15].

TABLE I
HEALTH CLASSIFICATIONS USED BY THE MAAQS

API	Air Pollution Level
0 -50	Good
51 - 100	Moderate
101 - 200	Unhealthy
201 – 300	Very Unhealthy
301 - 500	Hazardous
500+	Emergency

III. METHODS AND MATERIALS

In this section, we will be discussing the different methods used in our research to address the problems at hand. The data collection process and the data sources, the modelling process and the evaluation will all be presented. Figure 1 highlights all the important tasks in the project workflow.

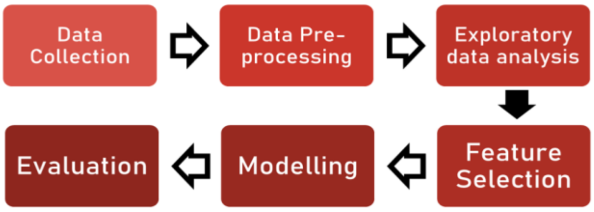

Fig. 1 Project Workflow

A. *Data Collection*

The process of data collection was initiated by the collection of daily ground-based monitoring data for Air Quality Index (AQI). In Malaysia the AQI is referred to as the Air Pollution Index (API), which is a simple way to quantify the air quality, it is calculated from numerous different sets of air pollution data. To determine the API for a given day, the sub-index values of 6 different air pollutants which consist of sulfur dioxide (SO2), particulate matter with the size of less than 10 micron, (PM10), nitrogen dioxide (NO2), and ground level ozone (O3), carbon monoxide (CO), and particulate matter with the size of less than 2.5 micron (PM2.5) concentrations. The maximum sub-index of the above-mentioned pollutants is then used as the API for the given day. The API data this research was retrieved from the Air Quality Historical Data Platform (https://aqicn.org/data-platform/) which obtained its data from the Department of Environment, Ministry of Environment and Water. Since the research's focus is on the east Malaysia region data was collected from stations located in this region. 7 different stations were selected in an arbitrary manner (Table 2).

TABLE II
LIST OF STATIONS

No.	Stations
1.	Kapit, Sarawak
2.	Kota Kinabalu, Sabah
3.	Kuching, Sarawak
4.	Samarahan, Sarawak
5.	Sibu, Sarawak
6.	Sri Aman, Sarawak
7.	Tawau, Sabah

The Aerosol optical depth (AOD) measurements were next retrieved from the Suomi National Polar-orbiting Partnership (SNPP) Visible Infrared Imaging Radiometer Suite (VIIRS) satellite, this satellite provides daily deep blue aerosol product which in turn provides satellite-derived measurements of Aerosol Optical Depth (AOD) and their properties over land and ocean as gridded aggregates, on a daily basis, globally. is provided in a 1° x 1° horizontal resolution grid. Each data field, in most cases, represents the arithmetic mean of all the cells whose latitude and longitude coordinates positions them within each grid element's bounding limits. This data was in netCDF format and contained 44 Science Data Set (SDS) layers, hence the dataset required for the research needed to be extracted and converted into a csv format file.

The most important set of data needed for the research being the fire data, was retrieved through the help of the Visible Infrared Imaging Radiometer Suite (VIIRS) satellite where a specific layer displays the active fire detections and thermal anomalies, which include volcanoes, and gas flares. The fire layer can be deemed useful in order to study the spatial and temporal distribution of fire and to locate the source of air pollution that may possibly have human health impacts.

Lastly the meteorological data for the selected region was retrieved from satellite data belonging to ERA5-Land, which is a reanalysis dataset providing a reliable view of the change in the land variables over past few decades at an improved resolution. All the above-mentioned datasets were downloaded for a specific time period of three months from August 2019 to October 2019. The selection of the time period was on the basis that the haze episode in Malaysia occurred around the selected time period. They were also specific to the selected ground stations as in that case it would correspond with the air quality. Table III displays all the collected datasets and the variables.

B. *Data Preprocessing*

The process of data preprocessing began with API dataset, however the API dataset only contained two columns, one being date and other the AQI and most of data was clean and did not contain any missing values, hence very limited treatment was performed to this dataset. Since our problem is a classification problem the target cannot be a continuous variable hence the API went through the data binning process which is a way to group several continuous values into a smaller number of "bins" resulting in a categorical variable. The data binning process was done in accordance with the Malaysian Ambient Air Quality Standard (MAAQS). Table 2 shows the different classes to which the continuous values can be binned into, the newly created variable will be called "Air Quality Level" and it will act as the target variable for the future selected models.

The next dataset, which was the Aerosol Optical Thickness, when it was collected the data was in NetCDF format. Generally, the NetCDF format is popular for storing multi-dimensional data The task was to see what types of variables are inside the dataset and accordingly extract the data. When looking at the literature that has been reviewed, there was no mention of how they dealt with this kind of file,

TABLE III
COLLECTED DATA

Dataset	Feature	Description
Air Quality Data	Date	The dates from 1st August 2019 to October 31st 2019
	Air Pollution Index	API readings from ground-based monitoring stations
Aerosol Optical Depth Data	Aerosol Optical Depth Data	A quantitative estimation of the concentration of aerosol existing in the atmosphere.
	Angstrom Exponent	A measure which defines how the optical depth of an aerosol depends on the wavelength of the light.
	Spectral AOD Land	A quantitative estimation of the concentration of aerosol existing in the atmosphere.
	Spectral AOD Ocean	A quantitative estimation of the concentration of aerosol existing in the atmosphere.
Fire Data	Distance from the center of fires	The distance from the center of the all fires surrounding the stations.
	Distance from the nearest fires	The from the nearest fire from the stations
	Fire count	The count of fires surrounding the stations at a given time.
Meteorological data	Forecast albedo	The measure of the reflectivity of the Earth's surface.
	Skin temperature	Temperature at the surface of Earth.
	Surface net solar radiation	A measure of solar radiation reaching the surface of the Earth.
	2m dewpoint temperature	Temperature at 2 meters above the surface of the Earth.
	Surface pressure	A measure of the ppressure of the atmosphere at the surface of land.
	Surface solar radiation downwards	A measure of solar radiation reaching the surface of the Earth.

However, after tirelessly searching online for some type of tool which will help in identifying the attributes inside this dataset, a tool named 'Panoply' developed by the National Aeronautics and Space Administration (NASA) was discovered. Using this tool, it confirmed the existence of more than 40 different attributes.

Despite being able to view the attributes, the tool couldn't be used to extract the dataset into a .csv format. Hence a python script was written in order to extract the data from 92 different files each corresponding to a different day of the selected time period. This process was performed for all the selected 7 stations.

The next immediate task was to select the most relevant attributes from the given 40, the selection was made, and 4 columns were selected, the basis of the selection was relevancy and amount of missing data. Although the selected columns had the least amount of missing data among all the columns, they still had a considerable amount of missing data and hence required interpolation. When the API was plotted against different selected attributes of the AOD (Figure 2) it was noticed that the relationship was linear and that the increase and decrease in AOD took place in a linear manner, hence it was established that linear interpolation was the best way to treat the missing data.

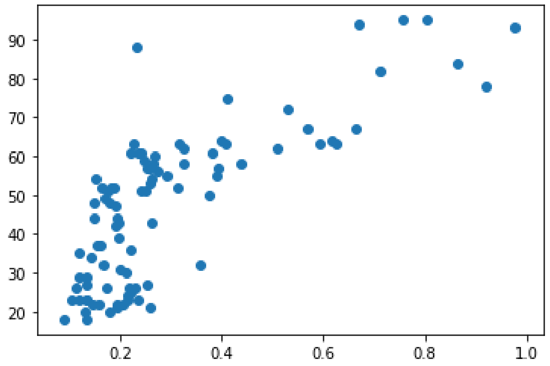

Fig. 2 Scatter plot (AOD vs API)

The other important dataset was the fire dataset, generally if we study the literature, we can notice that fire data is available with name, date, size and location of the fire. However, for Malaysia and Indonesia there has been no convention of naming fires despite there being many instances of wildfire incidents in these regions. Hence for this research it was incumbent to depend on satellite detected hotspots or thermal anomalies, as mentioned earlier hotspots do not necessarily mean there is fire in a certain location, it could also mean there's a volcanoes or gas flares, this isn't a problem as both volcanos and gas flares in some way contribute to pollution in a similar manner to fires.

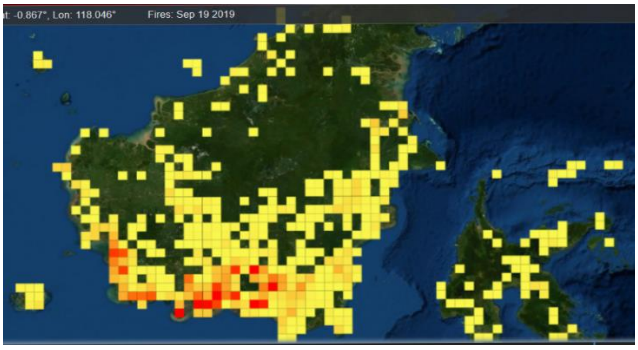

Fig. 3 Fires at a given time

The problem that was encountered when working with fire data is that at a given day there could be multiple fires occurring (Figure 3). However, in our dataset we can only include information of one fire on a given day, but all the fires on a given day play a part in the contribution of pollution. After many discussions and consultations, it was decided that the center of all the fires would be taken as the location of the fire, this was done with assumption that the effect of the fires will be the highest at the center. So, it was decided that the distance from the center of all the fires, the count of the fires in that region and the distance from the nearest fire. In the dataset fires had two main attributes, one being the Confidence which intends to let users to measure the quality of individual hotspot/fire pixels. Confidence values are set to low, nominal and high each indicating the quality of the fires, low confidence fires were removed from the dataset, the other attribute being Fire Radiative Power (FRP), the FRP depicts the intensity of the fire any fire with an FRP less than average FRP was removed in the final dataset. There are close to no missing data in the dataset hence no treatment was done.

The last dataset would be the meteorological data, this dataset like the AOD dataset was in the NetCDF format, hence it had to be converted, however the earlier script would not work for this dataset due to some differences, hence a brand-new python script was used to extract the data. All the datasets were next combined, and a correlation test was performed to study the linear dependence between the predictors and the target. The results suggested which features had low correlation with the target and these features were removed accordingly. A total of 12 features were removed from the dataset. The final dataset had a total of 10 predictors and one target feature.

Once completed with data preprocessing, the next stage would be to perform exploratory data analysis which is a way to analyse datasets in order to be able to summarize and understand their main characteristics, this is often done with visual methods. It is important to understand the types of data we are working with in order to understand the architecture of the data at hand, this will help in selecting the best algorithm.

We first take a look at the behaviour of API over the three selected months, plotted below (Figure 4) is the changes in API from August 2019 to October 2019 in all the selected stations, as we can see between day 30 to day 60 the API seems to be at the peak for all of the stations, however the extent to which the API increased differs from station to station, one example can be Kota Kinabalu where the change in API was very minimal, while Sri Aman had a drastic change. The plot indicates that the geographic location of the stations really matters, each of them might be cities which are slightly near to each other, but the pollution affects them differently.

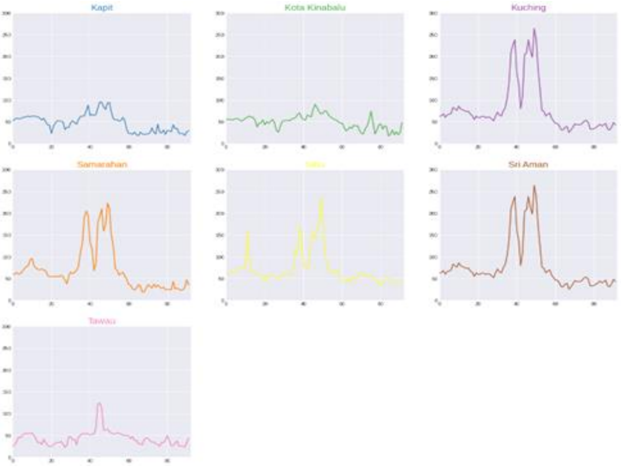

Fig 4: Line graph (API vs Date)

Next came the visualization of the newly created variable "Air Quality Level". In this plotted pie chart (Figure 5) indicates that the class "Very Unhealthy" occurs very few times, which points out that the dataset being worked on is an imbalanced dataset. An imbalanced dataset is a special scenario of a classification problem where the distribution of the classes is not done in uniform manner. Specific algorithms can handle imbalanced datasets well, however there are very few of them, there are other techniques to deal with imbalanced datasets, two of the most popular methods include Random Under/Over-sampling and using class weights during the modelling stage.

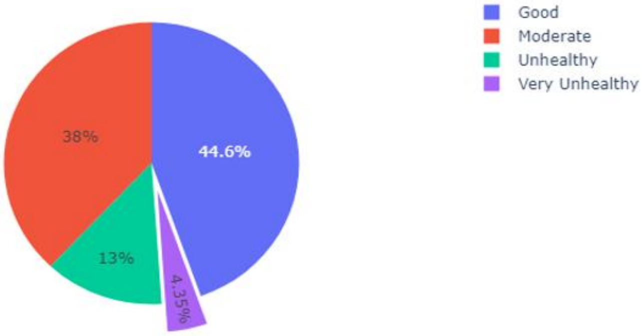

Fig 5: Air Quality Level class distribution

C. *Modelling and Results*

After the exploratory data analysis stage, the immediate next step would be to proceed to the modelling stage, which involves selecting two different classification algorithms and evaluating their performance for each of the selected stations. As a result of the data preprocessing performed earlier the remaining were 9 predictor features and one target (Table 4) were used to perform the modelling. The earlier used "API" feature was removed from the dataset as it will be no longer be used following the introduction of "Air Quality Level".

TABLE IV
SELECTED PREDICTORS AND TARGET VARIABLE

Predictors	Average Aerosol Optical Depth
	Angstrom Exponent
	Spectral AOD Land
	Spectral AOD Ocean
	Fire count
	Skin temperature
	Surface net solar radiation
	Surface solar radiation downwards
	Forecast albedo
Target	Air Quality Level

Two different algorithms were selected and trained accordingly is similar circumstances. The algorithms selected were Random Forest and Gradient Boosting. In order to be able to evaluate the models, a train-test split procedure was performed, this acts an estimate of how well the Machine Learning algorithm is performing when predictions are made on data which is not used to train the model. It is an efficient method to perform, the results of which will allow one to evaluate the performance of the selected Machine Learning algorithms for their predictive modelling problem. It is a well-suited method for supervised learning techniques including both classification and regression problems. The method requires a dataset to be divided it into two subsets. The first subset which is referred to as the training dataset will be used to fit the model. The second subset which is referred to as the testing subset will not be used to train the model; it will instead be used as the input provided to the model, based on the input then predictions are made and compared to the expected values. For this project the common split percentage of Train dataset: 80%, Test dataset: 20% was performed. The estimated accuracy of each of the models is highlighted in Table 5.

TABLE V
EVALUATION RESULTS (ACCURACY)

	Random Forest	Gradient Boosting
Kapit	0.929	0.964
Kota Kinabalu	0.750	0.821
Kuching	0.893	0.857
Samarahan	0.750	0.785
Sibu	0.821	0.714
Sri Aman	0.857	0.857
Tawau	0.928	0.892

Random Forest

Random forest is a supervised learning algorithm, which is an ensemble learning method used to deal with regression and classification problems. It works by constructing a large number of decision trees during the training process and outputs the class that is the mean prediction (regression) of the individual trees or mode of the classes (classification). It has also become one of the most popular algorithms, because of its simplicity and its flexibility in terms or working on both regression and classification problems. Based on literature review conducted, random forest was one of the best performing algorithms when it comes to dealing with air quality data, hence this algorithm was selected, Figure 6 highlights the performance of random forest using the confusion matrix. As we can see there are 3 major misclassifications which occurred (Figure 6). Figure 8 & 9 are the comparisons of how the algorithm fared with data from a different station.

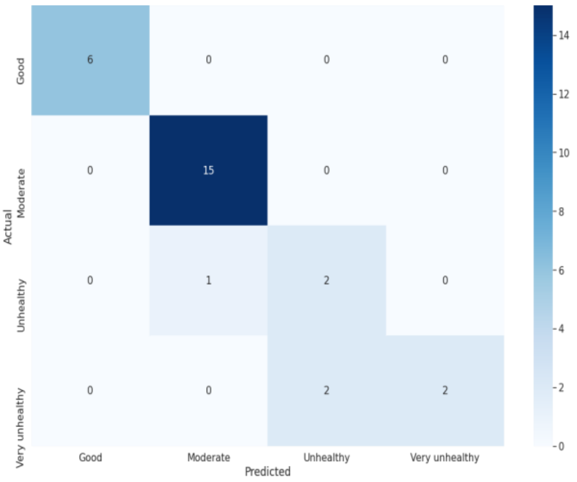

Fig 6: Confusion Matrix (Random Forest)

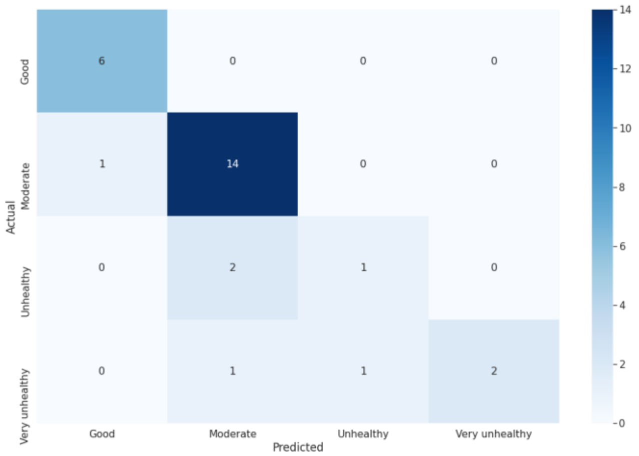

Fig 7: Confusion Matrix (Gradient Boosting)

Gradient Boosting

Like random forest, gradient boosting is a supervised learning algorithm which is also an ensemble learning method for classification and regression. Boosting is known to be a way of converting weak learners into strong learners. When boosting is done, every new decision tree is a fit on a modified version of the original data set. Gradient Boosting can train numerous models in a gradual, additive and sequential manner. The motivation for selecting gradient boosting is that it is classification model that has built-in approaches in order to combat class imbalance, as noticed earlier through the visualizations the dataset is quite imbalanced hence gradient boost will help with this by the construction of successive training sets based on incorrectly classified examples. Figure 7 is the resulting confusion matrix when gradient boosting was performed for a given location. The confusion matrix indicates that there were 4 misclassifications.

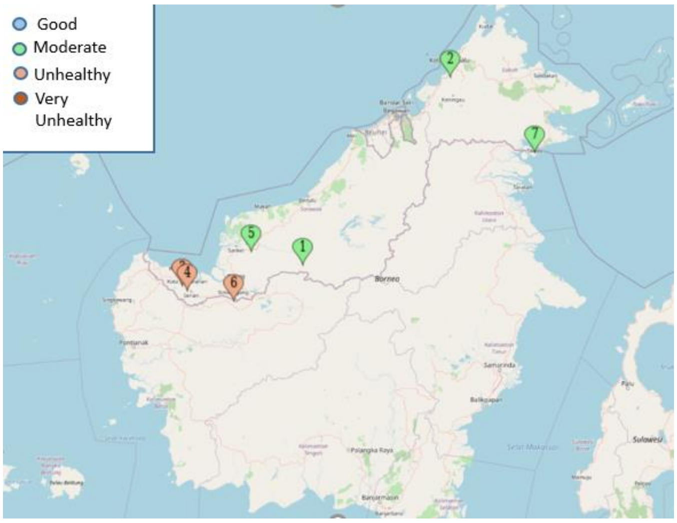

Fig.8 Actual Air Quality Level

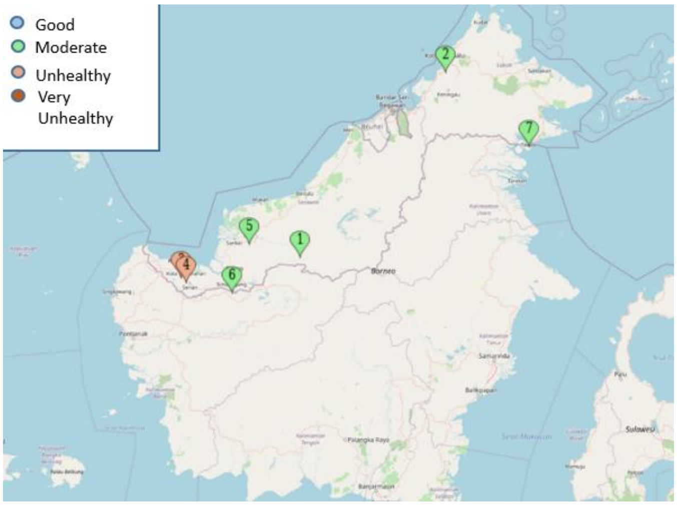

Fig. 9 Predicted Air Quality Level (Random Forest)

IV. Conclusions

After modelling and exploring, we were able to conclude that the air pollution level in areas neighboring wildfires are heavily dependent on the combination of the fire, meteorological and AOD data. Model selection, feature selection and model evaluation were all conducted and lead to reasonable performances by both the algorithms. In the future we aim to use other evaluation techniques to estimate the performance of the current models, to get a better picture of how the model is performing. Another matter which has to be resolved is that a proper measure for distance from fires to the stations must be discovered, as the method used in the initial preprocessing stage resulted in an extremely low correlation with API and more predictor variables may be added in the future if required.

Acknowledgment

We would like to acknowledge the International Islamic University Malaysia for its support. We would also like to extend our gratitude to Dr Rawad Abdulghafor and Dr Norzariyah Binti Yahya for helping us with the research.

References

[1] Mutalib, S. N. S. A., Juahir, H., Azid, A., Sharif, S. M., Latif, M. T., Aris, A. Z., ... & Dominick, D. (2013). Spatial and temporal air quality pattern recognition using environmetric techniques: a case study in Malaysia. Environmental Science: Processes & Impacts, 15(9), 1717-1728.

[2] Latif, M. T., Othman, M., Idris, N., Juneng, L., Abdullah, A. M., Hamzah, W. P., ... & Sahani, M. (2018). Impact of regional haze towards air quality in Malaysia: a review. Atmospheric Environment, 177, 28-44.

[3] Nangoy, F. (2019, October 21). Area burned in 2019 forest fires in Indonesia exceeds 2018 - official. Retrieved August 15, 2020, from https://www.reuters.com/article/us-southeast-asia-haze/area-burned-in-2019-forest-fires-in-indonesia-exceeds-2018-official-idUSKBN1X00VU

[4] CHUA, S. (2019, September 24). MCAQM station arrives in Serian to provide accurate data on air quality in the division. Retrieved August 15, 2020, from https://www.theborneopost.com/2019/09/24/mcaqm-

station-arrives-in-serian-to-provide-accurate-data-on-air-quality-in-the-division/

[5] UN. (n.d.). THE 17 GOALS | Sustainable Development. Retrieved November 12, 2020, from https://sdgs.un.org/goals

[6] Kleine Deters, J., Zalakeviciute, R., Gonzalez, M., & Rybarczyk, Y. (2017). Modeling PM2. 5 urban pollution using Machine Learning and selected meteorological parameters. Journal of Electrical and Computer Engineering, 2017.

[7] Wu, J., Winer, A. M., & Delfino, R. J. (2006). Exposure assessment of particulate matter air pollution before, during, and after the 2003 Southern California wildfires. Atmospheric Environment, 40(18), 3333-3348.

[8] Owusu-Akyaw, A., Li, R., & Moran, C. (2019). Spread of Wildfire Pollutants in California.

[9] Reid, C. E., Jerrett, M., Petersen, M. L., Pfister, G. G., Morefield, P. E., Tager, I. B., ... & Balmes, J. R. (2015). Spatiotemporal prediction of fine particulate matter during the 2008 Northern California wildfires using Machine Learning. Environmental science & technology, 49(6), 3887-3896.

[10] Sukitpaneenit, M., & Oanh, N. T. K. (2014). Satellite monitoring for carbon monoxide and particulate matter during forest fire episodes in Northern Thailand. Environmental monitoring and assessment, 186(4), 2495-2504.

[11] Kanabkaew, T. (2013). Prediction of Hourly Particulate Matter Concentrations in Chiangmai, Thailand Using MODIS Aerosol Optical Depth and Ground-Based Meteorological Data. EnvironmentAsia, 6(2).

[12] Ng, K. Y., & Awang, N. (2018). Multiple linear regression and regression with time series error models in forecasting PM 10 concentrations in Peninsular Malaysia. Environmental monitoring and assessment, 190(2), 63.

[13] Siew, L. Y., Chin, L. Y., & Wee, P. M. J. (2008). ARIMA and integrated ARFIMA models for forecasting air pollution index in Shah Alam, Selangor. Malaysian Journal of Analytical Sciences, 12(1), 257-263.

[14] Azid, A., Juahir, H., Toriman, M. E., Kamarudin, M. K. A., Saudi, A. S. M., Hasnam, C. N. C., ... & Osman, M. R. (2014). Prediction of the level of air pollution using principal component analysis and artificial neural network techniques: a case study in Malaysia. Water, Air, & Soil Pollution, 225(8), 2063.

[15] D. (n.d.). Air Pollutant Index (API). Retrieved August 15, 2020, from https://www.doe.gov.my/portalv1/en/info-umum/english-air-pollutant-index-api/100