International Journal of Advanced Science and Computer Applications (2023); 2(2): 59-68
DOI: 10.47679/ijasca.v2i2.30
https://ijasca.org/index.php/ijasca

RESEARCH ARTICLE

Cardiorespiratory Mortality Prediction Based on Air Pollution Using Tree-Based Ensemble Models

Raja Sher Afgun Usmani[1], Akibu Mahmoud Abdullahi[2*)], Thulasyammal Ramiah Pillai[3], Ibrahim Abaker Targio Hashem[4], Rafiza Binti Shaharudin[5], Mohd Talib Latif[6]

Published online: 10 June 2023

Abstract

Air pollution has a substantial negative impact on human wellbeing and health. Cardiorespiratory mortality is one of the primary effects of air pollution. In this study, we provide analysis of air pollution, cardiorespiratory mortality and the cardiorespiratory mortality is predicted based on air pollution using tree-based ensemble models. The tree-based ensemble models utilized in this study are Voting Regressor (VR), Random Forest (RF), Gradient Tree Boosting (GB), and Extreme Gradient Boosting (XGBoost). The used dataset contains data for five research locations: Shah Alam (SA), Klang (KLN), Putrajaya (PUJ), Cheras, Kuala Lumpur (CKL), and Petaling Jaya (PJ) from January 2006 to December 2016. The results show that XGBoost and VR models outperformed the rest of the models with the best evaluation metric scores in the Klang study area, XGBoost (MAE:0.005, RMSE:0.010, MAPE:0.70%) and VR (MAE:0.005, RMSE:0.011, MAPE:0.70%). The results reveal that the utilized models provided an excellent and accurate prediction of cardiorespiratory mortality based on air pollution and can follow the trend of cardiorespiratory mortality.

Keywords: air pollution, air quality, health, mortality, machine learning, prediction

Civil Engineering, Faculty of Engineering, Tridharma University

[1] Department of Computer Science, FCIT, University of Sialkot, Pakistan

[2*)] School of Computing and Informatics, Albukhary International University, Kedah, Malaysia

[3] School of Computer Science and Engineering, Taylor's University, Selangor, Malaysia

[4] College of Computing and Informatics, Department of Computer Science, University of Sharjah, UAE

[5] Environmental Health Research Centre, Institute for Medical Research, Ministry of Health Malaysia, Setia Alam, Shah Alam, Selangor

[6] Department of Earth Sciences and Environment, Faculty of Science and Technology, Universiti Kebangsaan Malaysia, 43600, Bangi, Selangor, Malaysia

*) *corresponding author*

Akibu Mahmoud Abdullahi
2School of Computing and Informatics, Albukhary International University, Kedah, Malaysia

Email: akibumahmoud@gmail.com

Introduction

According to World Health Organization (WHO), cardiovascular diseases (CVDs) cause 17.9 million lives lost per year and are the leading cause of death worldwide [1]. In 2016, CVDs account for 31% of all worldwide death [2], [3]. CVDs lead the cause of mortality and hospitalization throughout the world. In 2016, CVD accounted for 14 million disabilities and over 900,000 premature deaths in the US. In Iran, CVD was the leading cause of disability and death with 23% of diseases and 40% of all deaths [4]. In Malaysia, CVD has been among the primary cause of hospitalization for the past few years [5].

Scientific research has reported a strong association between cardiovascular diseases and ambient air pollution [6]. Furthermore, many diseases such as asthma, lung cancer, stroke, respiratory diseases are also attributed to exposure to air pollution. Correspondingly, air pollution is one of the primary environmental issues in a developing country. It also increases mortality and is the leading cause of harmful diseases on human health [6]. Air pollution causes over 6.5 million premature deaths worldwide, which outdoor air pollution accounts for over 2 to 4 million deaths every year [1]. Many studies have calculated the impact of air pollution and mortality in countries such as China, Canada, North America and Europe. [7] estimated that in 2015, over 1 million deaths in China were attributable to ambient air pollution, whilst [8]

estimated over 21,000 annual deaths in Canada. Epidemiological studies confirmed that air pollution caused more than 381,000 premature deaths in Europe and 68,000 in North America [9].

Exposure to air pollution is also strongly linked with non-communicable diseases, including CVDs and respiratory illnesses [10]. It is estimated that air pollution exposure contributes more to morbidity and mortality than all other environmental factors combined [10]. A study conducted in Tehran, Iran, reveals that air pollution increases the effect of mortality due to respiratory diseases [11]. Furthermore, the increased risk of mortality has a strong association with PM_{10} in South Korea and China [12], [13]. Ozone (O_3) has a significant association with mortality [14], NO_2, $PM_{2.5}$, and PM_{10} have significant association with increase of mortality also [15].

The increasing availability of environmental and hospitalization data has led to a rise in the dataset available for analysis. These big datasets come with complexity and difficulty, relying on traditional epidemiological, statistical, and environmental models to analyze them. In the past decade, machine learning algorithm has become famous for analyzing and predicting multidimensional, complex, and large amount of dataset [16]. Few studies were found for cardio-respiratory mortality prediction; for example, artificial neural network (ANN) and nonlinear autoregressive (NAR) models were developed to predict the respiratory mortality caused by outdoor air pollution in Ahvaz, Iran. They found CO and NO significantly affected mortality due to respiratory diseases, with NAR model showing better performance compared to ANN [17].

[18] used the dataset from Kaggle to predict the mortality rates affected by air pollution. The authors used Random Forest and Extreme Gradient Boosting (XGBoost) ensemble learning methods. Results show that Random Forest better predicts the mortality rate based on air pollution. [19] used Bayesian networks to model the air pollution, climate, and their health impacts. The environmental factors, exposure levels, and health outcomes were used to recognize the probabilistic dependence structure of the connection between health and the environment. The study results showed that the proposed model has good predictive power, and the model can generalize to new time periods and regions. [20] demonstrated that the public health knowledge, Remote Sensing datasets, and Geographic Information System (GIS) analysis tools could be used to estimate the mortality count due to $PM_{2.5}$ exposure. The study was conducted in Beijing, China, where the concentrations of $PM_{2.5}$ are higher than China's national standards. The study also presented the regions where mortality risk is increasing in Beijing.

To the best of the authors' knowledge, this is the first study of mortality prediction based on air pollution that uses huge amount and variety of datasets, especially in Malaysia. This study aims to predict the cardio-respiratory mortality caused by air pollution at Klang Valley, Malaysia using the cardiovascular mortality and air quality dataset from 2006 to 2016. The results and analysis will give crucial insight into the impact of air pollutants on Klang Valley, Malaysia residents.

Method

Study area

This study was conducted in Klang Valley, Malaysia. Klang Valley is a Malaysian urban region centred around Kuala Lumpur in the state of Selangor. We have used the dataset from 5 study areas in the Klang Valley for this study, namely Putrajaya (PUJ), Shah Alam (SA), Klang (KLN), Petaling Jaya (PJ), and Cheras, Kuala Lumpur (CKL).

The first study area is Putrajaya, with a population of 91,900 [21]. Putrajaya is the federal administrative centre of Malaysia's capital and a planned city. In 1999, due to overcrowding and congested traffic in Kuala Lumpur, the seat of the federal government was moved to Putrajaya. The territory of Putrajaya is confined in the Sepang district of Selangor. Putrajaya is also a part of the Multimedia Super Corridor (MSC) Malaysia, which encompasses the Klang Valley.

The second study area in our study is Shah Alam, the first planned city in Malaysia. It was declared as the new capital of the state of Selangor in 1978, it has a population of 617,149 residents [22]. The third study area is the former capital of Selangor, the royal town of Klang, with a population of 879,867 [23], and it has the 12th busiest transhipment and container port of the world [24].

Petaling Jaya is the fourth study area of the study. Petaling Jaya is a city located in Petaling District, previously, a satellite township for Kuala Lumpur, Malaysia's capital. Petaling Jaya was given the status of a city on 20 June 2006 and it has a population count of 543,415 [25]. The last research location, Cheras, is from the heart of the federal territory of Kuala Lumpur, the capital city of Malaysia. It is Malaysia's largest and most developed city, with a population estimation of 1,809,699 [26], and a total area of 243 km^2 (94 square miles).

Data

In this research, five different locations were explored, which include Shah Alam (SA), Klang (KLN), Petaling Jaya (PJ), Putrajaya (PUJ), and Cheras Kuala Lumpur (CKL). Ten years of data from 2006 to 2016 is used for each study location in the study. The Monthly Air Pollution Mortality (MAPM) dataset is used in this study. Daily and monthly air pollution values from the chosen monitoring stations are included in the dataset, as well as a count of cardiorespiratory death. The Department of Environment (DOE) in Malaysia provided the air quality data, while the Department of Statistics (DOS) in Malaysia provided the cardiorespiratory mortality dataset. The mortality dataset provided by DOS contains the basic information of the deceased, cause of death and address. The air quality dataset from DOE is engineered using the novel feature engineering algorithm [27], and the mortality data provided by DOS is cleaned and combined with air quality engineered data using the

spatial feature engineering algorithm [28]. In terms of causes of death, this study considers cardiovascular and respiratory deaths. Cardiovascular diseases are represented as I00-I99, and respiratory diseases are represented as J00-J99 in the International Classification for Diseases (ICD) [29]. Figure 1 depicts the Air Quality Monitoring (AQM) station's location, which are the focal points of our study areas.

Figure 1: Air Quality Monitoring Stations in Klang Valley, Malaysia

The dataset is generated using our previous work, with a radius value of 10,000 meters [28]. The radius parameter is an important aspect of the inclusion criteria. Researchers can choose the distance between the deceased patient and the AQM station. The patients living outside the radius will be excluded from the dataset by the spatial feature engineering algorithm.

Methods

Decision tree and Artificial Neural Network are two of the most popular machine learning algorithms. They give significantly contrasting predictions if the training dataset utilized in the algorithms have any perturbation [30]. These machine learning prediction algorithms have low bias and high variance. The tree-based ensemble methods are recommended to decrease the bias and/or variance [30]. Different models are created and combined as an ensemble to create one prediction model in the tree-based ensemble models [31]. Four ensemble models are utilized in the current study, i.e., Voting Regressor (VR), Random Forest (RF), Gradient Tree Boosting (GB), and Extreme Gradient Boosting (XGBoost).

Voting Regressor

The Voting Regressor (VR) is created on an intrinsic and simple concept. The concept is to join multiple prediction/forecasting models. A final forecast value is calculated using either their average predicted value or a value predicted by the majority of the machine learning algorithms in the ensemble. The working of VR is presented in Figure 2. VR helps machine learning algorithms with good performance to predict more accurately by balancing out their individual weaknesses.

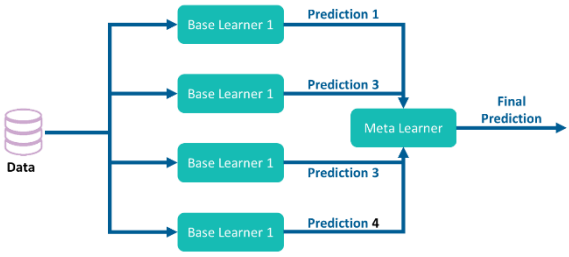


Figure 2: Architecture of Voting Regressor

Random Forest

Random decision forest or Random Forest (RF) is one of the most commonly utilized decision tree-based learning algorithms. It is very popular in both classification and regression problems in machine learning. RF was developed in 2001 by Leo Breiman [32]. Leo Breiman formulated an algorithm of using uncorrelated trees to create a forest using techniques akin to regression and classification of trees. The method included randomized node optimization and bagging. The overall working of RF is presented in Figure 3, with colors denoting different uncorrelated trees. In RF, various trees are trained using

marginally different data and then joined as a robust prediction model. The predictions carried out by the resulting model's committee are precise compared to individual decision trees in the forest.

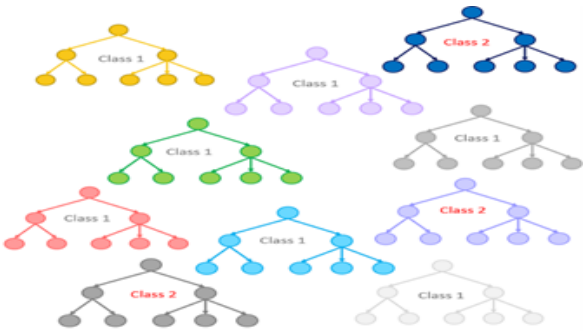


Figure 3: Architecture of Random Forests

Gradient Tree Boosting

Gradient Tree Boosting (GB) or Gradient Boosted Decision Trees is a machine learning algorithm based on a decision tree-based ensemble model. Researchers consider GB as one of the most effective and versatile machine learning predictive models. GB is considered an accurate and efficient method that can be applied for regression and classification problems. In GB, various regression trees are placed iteratively in a sequence. Every regression tree is trained by using the output of the regression tree before it in the iteration. At each step, a new learner is introduced to optimally reduce the loss function. Afterwards, these trees are combined using an additive model, creating a robust and effective ensemble model. The graphical illustration of GB is presented in Figure 4, with colors denoting different trees in a sequence.

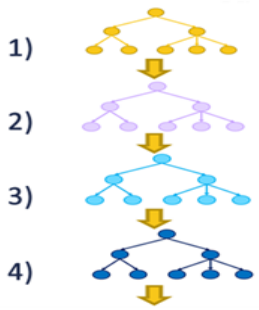


Figure 4: Architecture of Gradient Tree Boosting

Extreme Gradient Boosting

XGBoost, formally known as Extreme Gradient Boosting, extends and implements Gradient Boosted Decision Trees. XGBoost is specifically designed to enhance the speed and performance of the prediction and avoid the phenomenon of overfitting [33]. It is a highly scalable, end-to-end method. It can adapt easily and optimally utilize the resources available in the training phase. Data scientists use XGBoost to tackle various machine learning problems, often producing state-of-the-art results [?].

Results and Discussion

The parameters of the Monthly Air Pollution Mortality (MAPM) dataset and their descriptive statistics are provided in Table 1. The dataset consists of time series data with a time interval of one month. The data for all stations was collected from January 2006 to December 2016. The first parameter, the dataset's station attribute, contains the geocoded location. DOS provides the last parameter, cardiorespiratory mortality count. Rest of the parameters display the monthly statistics air pollutants, i.e., sulphur dioxide (SO_2), nitrogen oxide (NO), nitrogen dioxide (NO_2), nitrogen oxides (NO_x), carbon monoxide (CO), ozone (O_3),, and particulate matter (PM_{10}).

Table 1: Descriptive statistics of Monthly Air Pollution-Mortality datase

Station	Statistic	SO_2*	NO*	NO_2*	NO_x*	CO *	O_3*	PM_{10}**	Mortality
4*KLN	Mean	0.004	0.016	0.021	0.037	0.99	0.018	64	89
	Std	0.001	0.004	0.004	0.007	0.23	0.004	19	20
	Min	0.002	0.007	0.009	0.019	0.52	0.01	41	48
	Max	**0.008**	**0.035**	0.032	0.067	1.64	0.031	**159**	157
4*SA	Mean	0.003	0.017	0.02	0.037	0.79	0.021	52	22
	Std	0.001	0.004	0.004	0.006	0.19	0.004	16	8
	Min	0.001	0.01	0.009	0.021	0.46	0.014	34	6
	Max	0.006	0.026	0.038	0.058	1.33	0.034	147	45
4*PUT	Mean	0.002	0.006	0.014	0.02	0.59	0.021	44	47
	Std	0.001	0.002	0.003	0.004	0.14	0.005	16	15
	Min	0.001	0.001	0.006	0.011	0.27	0.01	23	19
	Max	0.005	0.016	0.023	0.035	1.31	**0.036**	133	77
4*PJ	Mean	0.004	0.033	0.029	0.062	1.29	0.015	49	72
	Std	0.001	0.007	0.004	0.008	0.23	0.003	15	17
	Min	0.002	0.019	0.017	0.044	0.86	0.009	26	40
	Max	0.007	0.05	**0.039**	**0.081**	**1.94**	0.026	126	116
4*CKL	Mean	0.002	0.017	0.021	0.037	0.86	0.02	49	105
	Std	0.001	0.005	0.003	0.006	0.18	0.004	14	74
	Min	0.001	0.005	0.01	0.019	0.5	0.011	30	23
	Max	0.004	0.029	0.03	0.058	1.77	0.034	116	**249**

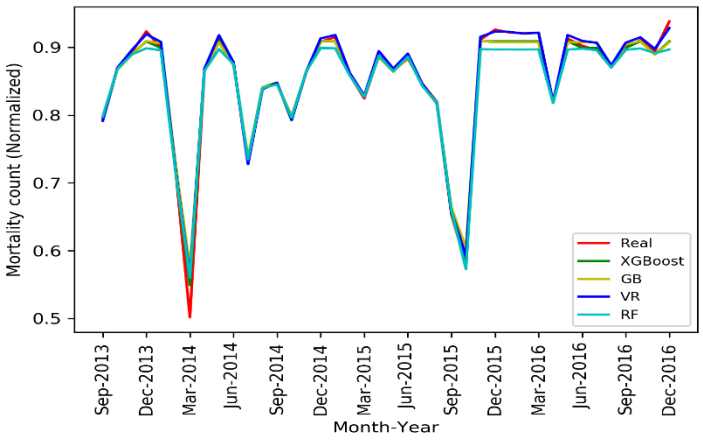
Note: Highest values are presented as bold numbers.
$^{**}\mu g/m^3$, $^{*}ppm$, Min=Minimum, Std=Standard Deviation, Max=Maximum

Table 1 displays the descriptive statistics of the air pollutants and mortality for the five study areas. Maximum, minimum, standard deviation and mean of the parameters can be visualized in Table 1. It can be seen that the highest values for each attribute are highlighted in Table 1. CKL has the highest mean mortality. CKL is based in Kuala Lumpur, which was then developed and the biggest metropolitan in Malaysia. A similar trend is found for the highest mortality count, with CKL leading the mortality count with 249 deaths in a month and SA having the lowest mortality mean (22).

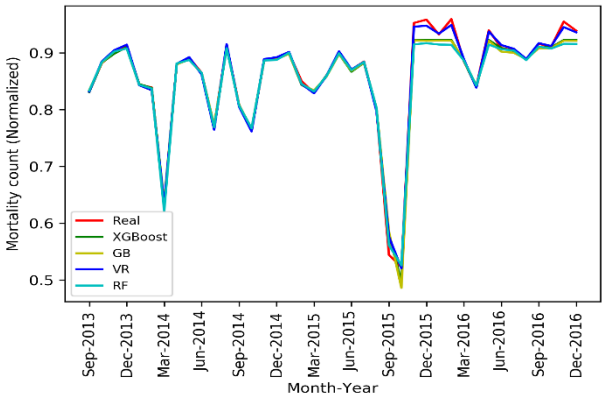
The lowest mean for CO is found in PUT (0.59 ppm), and the highest mean for CO is found in PJ(1.29 ppm). The lowest mean for O_3 is found in PJ (0.015 ppm), and the highest mean for O_3 is found in SA (0.021 ppm). The lowest mean for PM_{10} is found in PUT(44 $\mu g/m^3$), and the highest mean for PM_{10} is found in KLN (64 $\mu g/m^3$). The lowest mean for NO_x (0.020 ppm) and NO_2 (0.014 ppm) is found in PUT and the highest mean for NO_x (0.062 ppm) and NO_2 (0.029 ppm) is found in PJ. The lowest mean for NO is found in PUT(0.006 ppm), and the highest mean for NO is found in PJ (0.033 ppm). The lowest mean for SO_2 is found in PUT (0.002 ppm), CKL (0.002 ppm), and the highest mean is found in KLN (0.004 ppm) and PJ(0.004 ppm). After observing the descriptive statistics of the dataset, we can conclude that there is no pattern found for the lowest or highest mean for the mortality and air pollutant in any study areas.

Model Results and Comparisons

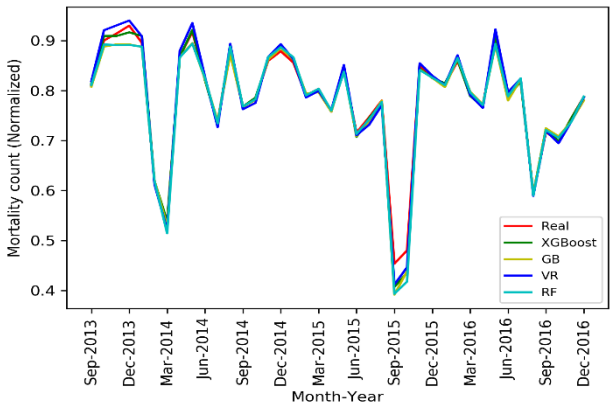
Four tree-based ensemble models are used in this study to predict the cardiorespiratory mortality count based on air pollution. The monthly mortality count and major air pollutants, i.e., CO, O_3, NO_x, PM_{10}, NO, SO_2, and NO_2 are used to train and predict the mortality count. The tree-based ensemble models are trained on 70% of the data (92 months), and prediction is carried out on 30% of the data (40 months).



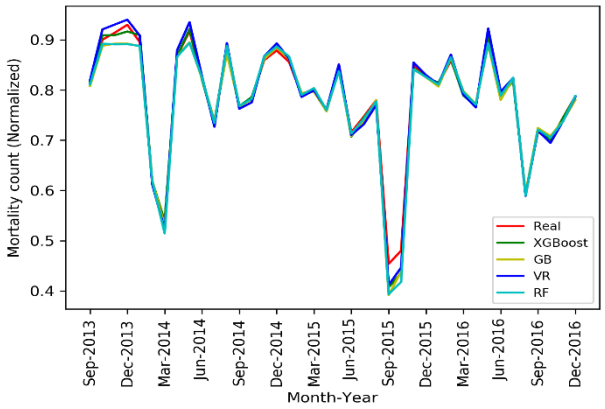
Klang



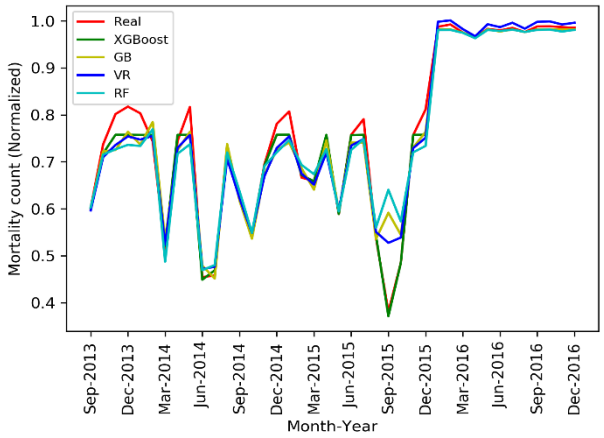
Shah Alam



Putrajaya



Petaling Jaya



Cheras, Kuala Lumpur

Figure 5: Comparison of mortality predictions

Figure 10 presents the results of test data using the tree-based ensemble models. Every graph represents a study area. The results clearly indicate that the developed models (ensemble models) can be used for mortality count prediction using air pollution accurately. The models can also track mortality trends and make predictions based on that data. All tree-based ensemble models performed well in all study areas, but the performance of VR was found better than the others. XGBoost performed better than all models in three study areas, i.e., KLN, PUT and CKL. VR performed better than all models in two research locations, i.e., SA and PJ.

Table 2: MAE, RMSE and MAPE Comparison - Mortality prediction

Station	Model	MAE	RMSE	MAPE (%)
41.2cm**KLN**	GB	0.008	0.015	1.1
	RF	0.009	0.016	1.2
	XGBoost	0.005	0.010	0.7
	VR	0.005	0.011	0.7
41.2cm**SA**	GB	0.021	0.035	3.5
	RF	0.02	0.035	3.3
	XGBoost	0.014	0.025	2.3
	VR	0.009	0.014	1.6
41.2cm**PUT**	GB	0.009	0.015	1.3
	RF	0.009	0.017	1.4
	XGBoost	0.006	0.011	0.9
	VR	0.010	0.013	1.4
41.2cm**PJ**	GB	0.009	0.015	1.1
	RF	0.008	0.015	0.9
	XGBoost	0.007	0.014	0.9
	VR	0.004	0.006	0.5
41.2cm**CKL**	GB	0.026	0.045	4.3
	RF	0.032	0.055	5.3
	XGBoost	0.013	0.022	1.8
	VR	0.026	0.038	4.0

Note: The best results for MAE, RMSE and MAPE are underlined.

Table 2 presents the MAE, RMSE and MAPE values for each model by study area. RMSE is often used as the primary criteria for predictive models in data science. The RMSE values for the tree-based ensemble models indicate that the XGBoost and VR models performed better than the other models. RMSE values of XGBoost were better than all models in three study areas, i.e., KLN, PUT and CKL. RMSE values of VR were better than all models in two study areas, i.e., SA and PJ. Other models used in the study performed quite well too.

The RMSE is always bigger or equal to the MAE values, as the RMSE assigns greater weight to the largest errors. As a result, when significant residual errors are undesirable, comparing the RMSE and MAE values is useful. The MAE values for the tree-based ensemble models employed in this research are also included in Table 2. The same trend for performance is visible in MAE values, with XGBoost and VR performing better than the other models. MAE values for XGBoost and VR show that the models do not have large residual errors, which is significant for our study as we predict mortality count, a crucial and sensitive parameter. All other models also performed well in MAE, with small residual errors.

Table 2 also shows the MAPE values, which is the third performance evaluation matric used in the study. The accuracy of MAPE is expressed as a percentage of the total error. As MAPE is presented as a percentage, it is easier to understand than the other evaluation metrics. The values of MAPE in the Table 2 reiterate that the performance of the utilized approaches is excellent, with XGBoost and VR models performing better than the other models. MAPE values of XGBoost were better than all models in two study areas, i.e., PUT and CKL. MAPE values of VR were better than all models in two research areas, i.e., SA and PJ. XGBoost and VR models had the same MAPE value for the KLN study area. In terms of MAPE, other models used in the study performed good as well.

The results imply that air pollutants readings can be utilized for prediction of cardiorespiratory mortality of residents of Klang Valley, Malaysia correctly. A fascinating part of our study is that various tree-based ensemble models are utilized, and all of them can accurately predict mortality count. As a result, we can indicate that air quality control is critical, and air quality warning systems can help regulate pollutant emissions in the atmosphere [34] and that air quality regulation should be continued and strengthened rigorously as it has a definite effect on cardiorespiratory mortality.

To the best of knowledge of the authors, a study of this volume and scope has never been done previously, especially in the setting of Malaysia. Three studies offer the most accurate attempts to predict the effects of air pollution on mortality rates [18]–[20], and one Kaggle inClass competition [35]. These studies also have limitations, i.e., none or limited use of temporal data [18]–[20], lack of spatial parameters [18], lack of validation with real data [18], [20]. Our results can conclude that mortality prediction can be achieved using historical air quality and mortality data. Using a patient's residential address as an exposure association parameter is one of the study's limitations. It is a general practice to use the residence location as air pollution exposure indication, as it is easier to monitor and collect, and it is used in a variety of researches [36]–[38]. We intend to analyze the relationship between air pollutants and mortality in future studies and quantify the impact of reduced air pollution on cardiorespiratory mortality. Future work for this study could also include reducing misclassification of air pollution exposure by incorporating a quantitative, location-specific, individual-level air pollution exposures and comparing

exposure levels throughout the cohort. Our findings also reveal that daily air quality can be used for mortality prediction and the relationship between latency and mortality, whether it is for daily or monthly predictions.

Conclusions and Recommendations

The study was conducted to predict the cardiorespiratory mortality in Klang Valley, Malaysia using air pollutants. Seven study areas are chosen for the study, ranging from relatively smaller cities to heavily populated urban areas. The descriptive statistics of the dataset generated using air quality data from DOE and mortality data from DOS revealed that the highest values for air pollution and mortality are found in densely populated and urban areas of Klang Valley. The correlation between air pollutants NO, NO_2, CO, O_3, NO_x, PM_{10}, and SO_2 and cardiorespiratory mortality was studied and no obvious pattern was found. The prediction was carried out using four tree-based ensemble approaches, i.e., Voting Regressor (VR), Random Forest (RF), Gradient Tree Boosting (GB), and Extreme Gradient Boosting (XGBoost). The results demonstrated that all tree-based ensemble models in the study could predict cardiorespiratory death. XGBoost and VR models outperformed the rest of the models with the best evaluation metric scores. Therefore, we infer that there is a link between air pollution and cardiorespiratory mortality because we were able to predict cardiorespiratory mortality using air pollutants accurately, and we further propose that continual efforts be made to control ambient air quality, as it has a significant impact on cardiorespiratory mortality among the population in the Klang Valley, Malaysia.

Declarations

All co-authors have seen and agree with the contents of the manuscript and there is no financial interest to report. We certify that the submission is original work and is not under review at any other publication.

Ethical Approval

Not applicable as the research is using secondary datasets provided by DOE and DOS, Malaysia.

Consent to Participate

Not applicable as the research is using secondary datasets provided by DOE and MOH, Malaysia.

Consent for publication

The consent to publish is granted from the DG of health, Malaysia.

Authors Contributions

This research work is part of a project entitled as "Modeling and Visualization of Air-Pollution and its Impacts on Health". R.S.A. Usmani conducted the experiments and wrote the manuscript. TR Pillai conceived and designed the research. IAT Hashem and AM Abdullahi are responsible for feature engineering and analysis. MT Latif and R Shaharuddin are responsible for data collection, statistical analysis, and discussion. All authors reviewed the final manuscript.

Funding

This research is funded by Taylor's University under the research grant application ID (TUFR/2017/004/04) entitled as "Modeling and Visualization of Air-Pollution and its Impacts on Health".

Competing Interests

The authors declare no competing interests.

Availability of data and materials

The datasets used for the experimentation and analysis are part of MOU between Taylor's university, DOE, Malaysia and DOS, Malaysia. These datasets are not available to publish publicly according to the DOE and DOS, Malysia. This research is funded by Taylor's University under the research grant application ID (TUFR/2017/004/04) entitled as "Modeling and Visualization of Air-Pollution and its Impacts on Health". We would like to thank the Director General of Health, Malaysia for his permission to publish this article. We would also like to extend our thank to Department of Environment and Department of Statistics, Malaysia for providing datasets of air quality and mortality respectively.

References

[1] WHO, "World Health Organization - Health Burden." 2021, [Online]. Available: https://www.who.int/health-topics/cardiovascular-diseases.

[2] Y. Liu, J. Sun, Y. Gou, X. Sun, D. Zhang, and F. Xue, "Analysis of short-term effects of air pollution on cardiovascular disease using Bayesian spatio-temporal models," *Int. J. Environ. Res. Public Health*, vol. 17, no. 3, p. 879, 2020.

[3] Q. Chen, Q. Wang, B. Xu, Y. Xu, Z. Ding, and H. Sun, "Air pollution and cardiovascular mortality in Nanjing, China: Evidence highlighting the roles of cumulative exposure and mortality displacement," *Chemosphere*, vol. 265, p. 129035, 2021.

[4] N. Sarrafzadegan and N. Mohammmadifard, "Cardiovascular disease in Iran in the last 40 years: prevalence, mortality, morbidity, challenges and strategies for cardiovascular prevention," *Arch. Iran. Med.*, vol. 22, no. 4, pp. 204–210, 2019.

[5] W. R. W. Mahiyuddin, "AIR POLLUTION AND CARDIORESPIRATORY HOSPITALIZATION TRENDS BASED ON THE DAYS OF THE WEEK IN KLANG VALLEY, MALAYSIA," *Asia Pacific Environ. Occup. Heal. J.*, vol. 4, no. 3, 2018.

[6] C. L. Lokotola, C. Y. Wright, and J. Wichmann, "Temperature as a modifier of the effects of air pollution on cardiovascular disease hospital admissions in Cape Town, South Africa," *Environ. Sci. Pollut. Res.*, vol. 27, no. 14, pp. 16677–16685, 2020.

[7] J. Huang, X. Pan, X. Guo, and G. Li, "Health impact of China's Air Pollution Prevention and Control Action Plan: an analysis of national air quality monitoring and mortality data," *Lancet Planet. Heal.*, vol. 2, no. 7, pp. e313–e323, 2018, doi: 10.1016/S2542-5196(18)30141-4.

[8] D. J. Nowak, S. Hirabayashi, M. Doyle, M. McGovern, and J. Pasher, "Air pollution removal by urban forests in Canada and its effect on air quality and human health," *Urban For. Urban Green.*, vol. 29, no. April 2017, pp. 40–48, 2018, doi: 10.1016/j.ufug.2017.10.019.

[9] U. Im *et al.*, "Assessment and economic valuation of air pollution impacts on human health over Europe and the United States as calculated by a multi-model ensemble in the framework of AQMEII3," *Atmos. Chem. Phys.*, vol. 18, no. 8, pp. 5967–5989, 2018.

[10] S. G. Al-Kindi, R. D. Brook, S. Biswal, and S. Rajagopalan, "Environmental determinants of cardiovascular disease: lessons learned from air pollution," *Nat. Rev. Cardiol.*, vol. 17, no. 10, pp. 656–672, 2020.

[11] S. Jaafari, A. A. Shabani, M. Moeinaddini, A. Danehkar, and Y. Sakieh, "Applying landscape metrics and structural equation modeling to predict the effect of urban green space on air pollution and respiratory mortality in Tehran," *Environ. Monit. Assess.*, vol. 192, pp. 1–15, 2020.

[12] O.-J. Kim, S.-Y. Kim, and H. Kim, "Association between long-term exposure to particulate matter air pollution and mortality in a South Korean National Cohort: comparison across different exposure assessment approaches," *Int. J. Environ. Res. Public Health*, vol. 14, no. 10, p. 1103, 2017.

[13] P. Yin *et al.*, "Particulate air pollution and mortality in 38 of China's largest cities: Time series analysis," *BMJ*, vol. 356, p. j667, 2017, doi: 10.1136/bmj.j667.

[14] Q. Di *et al.*, "Air pollution and mortality in the Medicare population," *N. Engl. J. Med.*, vol. 376, no. 26, pp. 2513–2522, 2017.

[15] Y. Liu *et al.*, "Short-term exposure to ambient air pollution and mortality from myocardial infarction," *J. Am. Coll. Cardiol.*, vol. 77, no. 3, pp. 271–281, 2021.

[16] C. Bellinger, M. S. Mohomed Jabbar, O. Zaïane, and A. Osornio-Vargas, "A systematic review of data mining and machine learning for air pollution epidemiology," *BMC Public Health*, vol. 17, no. 1, pp. 1–19, 2017, doi: 10.1515/JPEM.2001.14.2.151.

[17] D. N. Khojasteh, G. Goudarzi, R. Taghizadeh-Mehrjardi, A. B. Asumadu-Sakyi, and M. Fehresti-Sani, "Long-term effects of outdoor air pollution on mortality and morbidity--prediction using nonlinear autoregressive and artificial neural networks models," *Atmos. Pollut. Res.*, vol. 12, no. 2, pp. 46–56, 2021.

[18] K. C. Dewi, W. F. Mustika, and H. Murfi, "Ensemble learning for predicting mortality rates affected by air quality," in *Journal of Physics: Conference Series*, 2019, doi: 10.1088/1742-6596/1192/1/012021.

[19] C. Vitolo, M. Scutari, M. Ghalaieny, A. Tucker, and A. Russell, "Modeling Air Pollution, Climate, and Health Data Using Bayesian Networks: A Case Study of the English Regions," *Earth Sp. Sci.*, vol. 5, no. 4, pp. 76–88, 2018, doi: 10.1002/2017EA000326.

[20] Y. Li, Z. Chen, and J. Li, "How many people died due to PM2.5 and where the mortality risks increased? A case study in Beijing," in *International Geoscience and Remote Sensing Symposium (IGARSS)*, 2017, doi: 10.1109/IGARSS.2017.8126997.

[21] Theborneopost, "Malaysia's population stood at 32.6 million in Q4 2018." 2018, Accessed: Dec. 31, 2020. [Online]. Available: http://www.theborneopost.com/2019/02/13/malaysias-population-stood-at-32-6-million-in-q4-2018/.

[22] A. Populations, "Shah Alam." 2021, Accessed: Jan. 13, 2021. [Online]. Available: https://all-populations.com/en/my/population-of-shah-alam.html.

[23] WorldoMeters, "Malaysia Population 2020," *Worldometers*. 2020, Accessed: Dec. 31, 2020. [Online]. Available: https://www.worldometers.info/world-population/malaysia-population/.

[24] iContainers, "List of busiest container ports." 2021, Accessed: Jan. 13, 2021. [Online]. Available: https://www.icontainers.com/top-20-ports-in-the-world/.

[25] Wikipedia, "Petaling Jaya." pp. 1–8, 2020, Accessed: Dec. 31, 2020. [Online]. Available: https://en.wikipedia.org/wiki/Petaling%7B%5C_%7DJaya.

[26] A. Population, "Kuala Lumpur." 2021, Accessed: Jan. 13, 2021. [Online]. Available: https://all-populations.com/en/my/population-of-kuala-lumpur.html.

[27] R. S. A. Usmani, W. N. F. B. W. Azmi, A. M. Abdullahi, I. A. T. Hashem, and T. R. Pillai, "A novel feature engineering algorithm for air quality datasets," *Indones. J. Electr. Eng. Comput. Sci.*, vol. 19, no. 3, Sep. 2020.

[28] R. S. A. Usmani, T. R. Pillai, I. A. T. Hashem, N. Z. Jhanjhi, and A. Saeed, "A Spatial Feature Engineering Algorithm for Creating Air Pollution Health Datasets," *International Journal of Cognitive Computing in Engineering*. Elsevier, Nov. 2020, doi: 10.1016/j.ijcce.2020.11.004.

[29] J. Eisfeld, "International Statistical Classification of Diseases and Related Health Problems," *TSQ Transgender Stud. Q.*, 2014, doi: 10.1215/23289252-2399740.

[30] M. A. Hassan, A. Khalil, S. Kaseb, and M. A. Kassem, "Exploring the potential of tree-based ensemble methods in solar radiation modeling," *Appl. Energy*, 2017, doi: 10.1016/j.apenergy.2017.06.104.

[31] T. Hochkirchen, "Modern Multivariate Statistical Techniques: Regression, Classification, and Manifold Learning," *J. R. Stat. Soc. Ser. A (Statistics Soc.*, 2010, doi: 10.1111/j.1467-985x.2009.00634_10.x.

[32] L. Breiman, "Random forests," *Mach. Learn.*, vol. 45, no. 1, pp. 5–32, 2001.

[33] T. Chen and C. Guestrin, "XGBoost: A Scalable Tree Boosting System," in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2016, pp. 785–794, doi: 10.1145/2939672.2939785.

[34] R. S. A. Usmani, I. A. T. Hashem, T. R. Pillai, A. Saeed, and A. M. Abdullahi, "Geographic Information System and Big Spatial Data," *Int. J. Enterp. Inf. Syst.*, vol. 16, no. 4, 2020.

[35] Kaggle, "Predict impact of air quality on mortality rates - Overview." 2016, [Online]. Available: https://www.kaggle.com/c/predict-impact-of-air-quality-on-death-rates/overview.

[36] Y. Jie, "Air pollution associated with sumatran forest fires and mortality on the malay peninsula," *Polish J. Environ. Stud.*, vol. 26, no. 1, pp. 163–171, 2017, doi: 10.15244/pjoes/64642.

[37] N. A. Mabahwi, O. L. H. Leh, S. N. A. M. Musthafa, and K. Aiyub, "Air quality-related human health in an urban region. Case study: State of Selangor, Malaysia," *EnvironmentAsia*, vol. 11, no. 1, pp. 194–216, 2018, doi: 10.14456/ea.2018.15.

[38] M. A. B. A. Tajudin *et al.*, "Risk of concentrations of major air pollutants on the prevalence of cardiovascular and respiratory diseases in urbanized area of Kuala Lumpur, Malaysia," *Ecotoxicol. Environ. Saf.*, vol. 171, no. June 2018, pp. 290–300, 2019, doi: 10.1016/j.ecoenv.2018.12.057.