International Journal of Environmental Science and Technology (2023) 20:6103–6116
https://doi.org/10.1007/s13762-022-04367-6

ORIGINAL PAPER

Environmental hazard assessment and monitoring for air pollution using machine learning and remote sensing

S. Abu El-Magd[1] · G. Soliman[2] · M. Morsy[2] · S. Kharbish[1]

Received: 3 March 2022 / Revised: 12 May 2022 / Accepted: 19 June 2022 / Published online: 30 July 2022

Abstract

The prediction and study of air pollution is a complex process due to the presence of controlling factors, different land use, and different sources for the elaboration of pollution. In this study, we applied the machine learning technique (Random Forest) with time series of particulate matter pollution records to predict and develop a particulate matter pollution susceptibility map. The applied method is to strict measures and to better manage particulate matter pollution in Ras Garib city, Egypt as a case study. Air pollution data for the period between 2018 and 2021 is collected using five air quality stations. Some of these stations are located near highly urbanized locations and could be dense with the current rates of development in the future. The random forest was applied to verify and visualize the relationships between the particulate matter and different independent variables. Spectral bands of Landsat OLI 8 imaginary and land cover/land use indices were used to prepare independent variables. Analysis of the results reveals that the proper air quality distribution monitoring stations would provide a deep insight into the pollution distribution over the study site. Distance from the roads and the land surface temperature has a significant effect on the distribution of air quality distribution. The obtained probability and classification maps were assessed using the area under the receiver operating characteristic curve. The outcome prediction maps are reasonable and will be helpful for future air quality monitoring and improvements. Furthermore, the applied method of pollutant concentration prediction is able to improve decision-making and provide appropriate solutions.

Keywords Machine learning · Random forest model · Air pollution · Particulate matter · Petroleum site · Egypt

Introduction

Environmental pollution problems interest is increasing recently, with the increase of industrialization, urbanization, and other human activities. Air pollution is considered to occur whenever an excessive quantity of pollutants is released into the environment. Thus, air pollution has a direct influence on human health due to the exposure to pollutants and particulates (Hvidtfeldt et al. 2018; Pimpin et al. 2018; Gonzalez et al. 2017).

Ras Gharib area (RG) is one of the important oil production Provinces in Egypt, which is located on the Red Sea coast about 150 km to the north of Hurghada city. In general, the Gulf of Suez area has excellent potential hydrocarbon with a sedimentary basin covering about approximately 19,000 km^2. This basin has more than 80 oil fields (Ramadan et al. 2012). Further, the existence of several oil fields in the area, and thousands of people working inside these fields, they are possible could be under air pollution or environmental impact. Excessive emissions of air pollutants in the RG have obviously been recorded, including particulate matter (PM), sulfur dioxide (SO_2), carbon monoxide (CO), and nitrogen oxides (NOx). However, many regulations and rules are put in the place to monitor and reduce air pollution, still, some of them imposed concentrations above the limits. Significant controlling factors are potentially impacting the air pollutant levels including air temperature, wind direction, wind speed, humidity, as well as topography and terrain. ML approaches were implemented over the past years to help overcome the limitations imposed by data scarcity and insufficient spatial distribution of the datasets.

Editorial responsibility: Maryam Shabani.

S. Abu El-Magd
sherif.abuelmagd@sci.suezuni.edu.eg

[1] Geology Department, Faculty of Science, Suez University, Suez, Egypt

[2] Dara Petroleum Company, Ras Gharib, Egypt

A machine learning (ML) classifier (Random Forest (RF)) was implemented to evaluate and map air pollution. RF due to its good controllability, thus it has been applied successfully in many fields to assess and solve many issues. Therefore, over the past two decades, RF applications include lithology identification (Xie et al. 2018), microarray data classification (Diaz and Andrés 2006; Moorthy and Mohamad 2011), flood-prone areas identification (Zhao et al. 2018; Abu El-Magd 2022), soil texture and pH prediction (Pahlavan-Rad and Akbarimoghaddam 2018), air quality prediction and modeling (Yu et al. 2016; Shamsoddini et al. 2017; Joharestani et al. 2019; AlThuwaynee et al. 2021). RF can blend the concepts of bagging and random feature selection leading to better performance than other algorithms (Archer and Kimes 2008). Furthermore, the advantage of RF is its resistance to overtraining and its capability to grow huge numbers of random trees without the risk of overfitting (Shahabi and Hashim 2015) in addition to learning fast. Thus, RF can automatically handle the missing values or input, where it does not require transforming, rescaling, or modifying (Kamińska 2018). The implemented model relies on field data collection of PM10 during the years 2018–2021 and initial dependent features including air temperature, normalized difference vegetation index (NDVI), Soil-adjusted vegetation (SAVI) index and multiple air pollutants will be used to generate the RF prediction model. The resultant model performance was evaluated using an assessing metric namely, the receiver operating (ROC) curve.

Literature using MLs were employed in different fields such as air and water pollution, soil, floods, and landslides (Boonphun et al. 2019; Abu El-Magd et al. 2021a, b; Althuwaynee et al. 2021; Campanile et al. 2021; Abu El-Magd 2022). Air pollution has become a big concern on the planet, and it is also one of the leading causes of death (Doreswamy et al. 2020). Several studies applied machine learning for air quality mapping and forecasting (Raimondo et al. 2007; Muhammad and Yan 2015; Garcia et al. 2016; Yu et al. 2016; Park et al. 2018). Due to incomplete information or dataset, machine learning shows its capabilities to handle this issue. Constructing, mapping, and prediction, based on the pollution concentration levels of individual pollutants, will help to predict air quality hourly in the investigated area for air quality monitoring. They may play a key role in health alerts when air pollution levels might exceed the recommended levels. To the best of our knowledge, no prior investigations applied similar machine learning approaches in the area of the study for air pollution mapping and prediction. In this context, probability, and classification index mapping of PM10 using an air quality station dataset is the main concern. Besides, this study investigates the importance of the different independent variables. However, the main objectives of the present study are concluded (1) build a prediction model for hourly air quality in the area around Ras Gharib. To achieve this, one of the most powerful ML approaches was applied, i.e., RF. (2) Furthermore, developed a spatial and temporal hazard classes model to determine the air quality in such coastal. As per safety precaution, this mapping and predictive model may be considered as a basis for applying pollution monitoring and control processes.

Study area

Description of the area

The area is characterized by arid conditions, where the rainfall is rare with high storm events in a short time. RG area (Fig. 1) is characterized topographically by high hills (up to 1600 m amsl.) on the west, with a lowland on the coastal strip (to the east). The urbanized area and petroleum activities are located in and around the area. However, the drainage runs from the west through the high land to the east on the coast.

Geological setting

Geologically RG area that belongs to the Northern Egyptian Eastern Desert (Fig. 1b) comprises a wide range of rock units of Neoproterozoic age (Ball 1952; Abdallah et al. 1963; Meshref et al. 1983; GPC 1985; Stern and Hedge 1985; Stern 1994,). The main rock units exposed in the investigated RG area are wadi deposits (Qw), Sabkha deposits (Qb) (silt, clay, and evaporites), and Quaternary (Q) (sand, gravels, and recent coastal deposits). Tertiary deposits are represented by Pliocene deposits (Tpl) and a transgressive–regressive sequence of nummulitic limestone (Tm).

Materials and methods

Materials

The dataset of air quality pollution presented in this study was measured and collected from 5 stations in the area (Fig. 1). All dataset collected contains hourly data of the PM10 were collected from the stations. Remote sensing (RS) data were extracted from the USGS website (https://earthexplorer.usgs.gov). The remote sensing dataset includes thematic layers of bands 2 to 6, SAVI, and land surface temperature (LST). Tables 1, 2 and 3 report a summary of the site measured parameter and a detailed variable description used in the study.

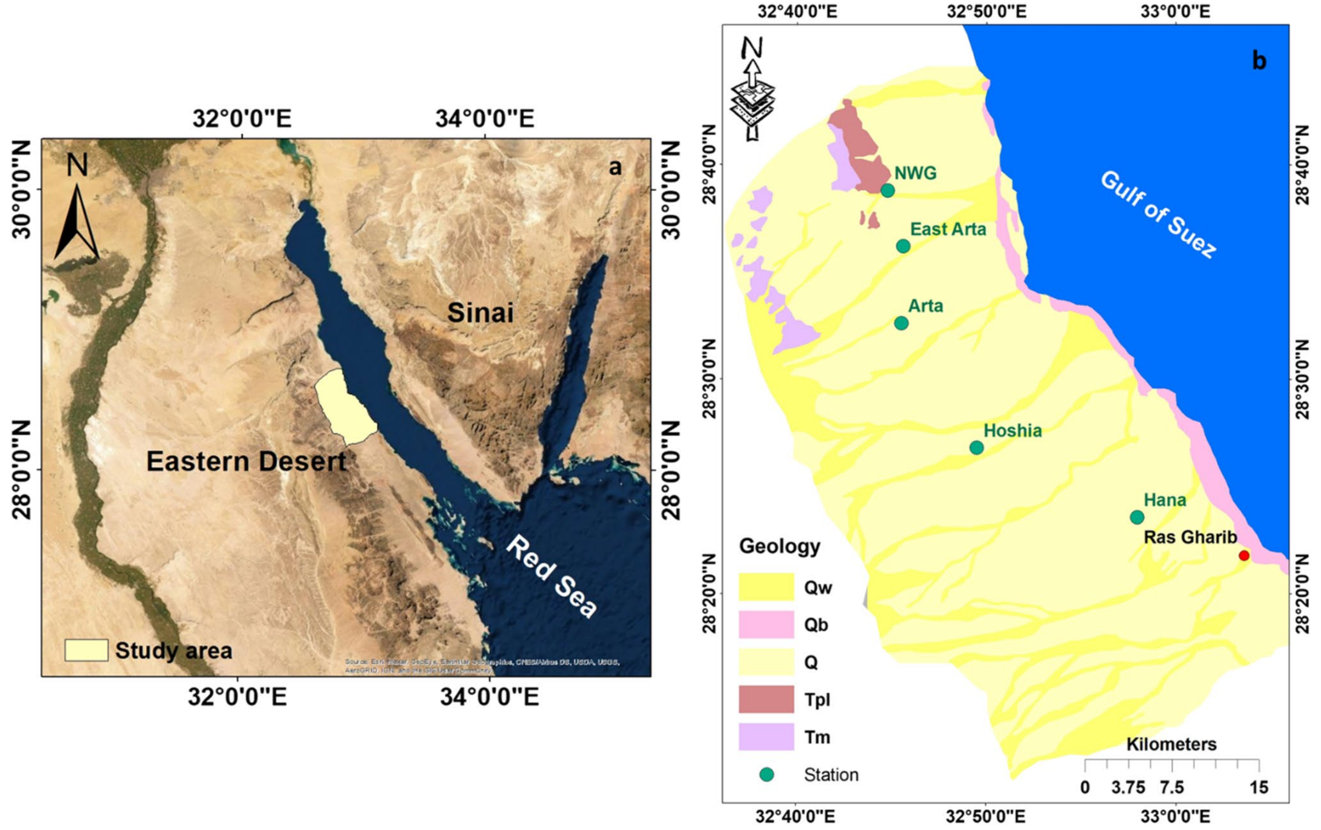

Fig. 1 Location map of the Study site (**a**) and sample locations (**b**)

Methods

Once the raw datasets were available in a manageable format, data was prepared and pre-processed in Quantum GIS (QGIS) for preparing the database for further mapping and prediction in the R programming package. Ten independent variables for air pollution (PM10) were used namely, band 2, band 3, band 4, band 5, band 6, LST, SAVI, BU, and distance from the roads. The R was used to implement the approach RF for air pollution mapping and prediction. RF consists of a set number of simple decision trees. RF is known as a technique for generating an ensemble (or forest) of tree-structured classifiers. It combines and extends the capabilities of CART decision trees (Izenman 2008; Nisbet et al. 2009; Steinberg and Golovnya 2013). A tree similar to CART is built with a bootstrap sampling for random subsets.

Table 1 Summary statistics of the PM10 measurement in the stations (2018–2021)

Fields	Hana	Hoshia	Arta	East Arta	NWG
Min	13.00	11.00	12.50	8.00	12.00
Max	95.00	103.00	620.00	37	105.00
Mean	41.57	58.96	120.47	27.33	46.40

Values in μgm^{-3}

Table 2 Maximum value of PM10 in the present work

Area	Station	PM10 (μgm^{-3})
Gharib	Hana	95
	Hoshia	103
	East Arta	37
	Arta Field	620
	NWG	105

Table 3 PM10 limits according to US, EPA, (1997) in μgm^{-3}

Category	Limits	Class
I	0 – 54	Good
II	55 – 154	Moderate
III	155 – 254	Unhealthy for sensitive
IV	255 – 354	Unhealthy
V	355 – 424	Very unhealthy
VI	> 424	Hazard

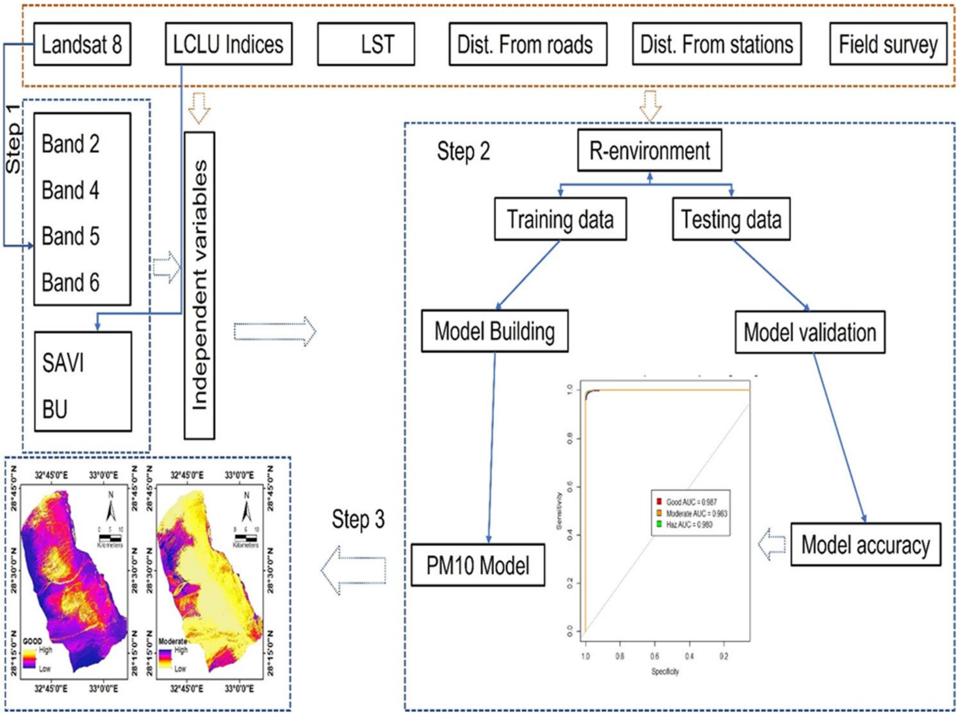

Fig. 2 Flow chart of the applied methods in the present work

The research methodology can be summarized in three steps (Fig. 2).

Results and discussion

Feature importance

After training the RF model, it is necessary to look at which variables or features have the most predictive power for the model. Thus, variables with high importance percentages are drivers of the model outcome. Therefore, these drivers variable values have a significant impact on the outcome model values. Understanding and calculating the importance of the variables can help in choosing the relevant features for the model. Meanwhile, the processing time is strongly affected by the number of the variables (AlThuwaynee et al. 2021). The selected variables for the model were employed with a different number of variables (10 variables). Figure 3 reflects the importance of the variables during different variables. A significant difference in variable importance was observed between air quality classes (good, moderate, unhealthy sensitive, and hazard classes. In the current model from the 10 input variables, distance from the road and LST followed by SAVI are the most important variables for the model. This no change in importance contributed to the removal of band 4 which has no impact on the model importance.

Independent variables

LST: Generally, the incoming solar radiation and energy interact with the ground surface and heat the ground. Therefore, the LST measures the thermal radiance emission from the land surface. LST depicts the average yearly land surface temperature (Fig. 4) in degrees Celsius as measured using the spectroradiometer imaging (Land sat 8). LST in the area ranges between 27.68 and 46.44 °C, the most obvious LST pattern that the map show is the land surface temperature in the study site is relatively moderate to high. Mathematically, LST can be calculated from the following equation (Eq. 1);

$$LST = \left(\frac{BT}{1 + \left(0.00115 * \frac{BT}{1.4388} \right) * Ln(\varepsilon)} \right) \tag{1}$$

where LST donates land surface temperature, BT donates brightness temperature and ε donates emissivity.

Landsat bands: Landsat (8) Operational Land Imager (OLI) and Thermal Infrared Sensor (TIRS) images consist of 9 spectral bands with 30 m resolution for Bands 1 to 7 and 9. Five bands of Land sat (8) were acquired freely for 5 years (http: //earthexplorer.usgs.gov) including (B2, B3, B4, B5, and B6). In the R environment bands were corrected and calibrated to be used later on in the estimation of the indices.

SAVI: in the areas where the vegetation cover is low, the SAVI is used to correct the influence of soil brightness in the normalized difference vegetation index, (NDVI). In

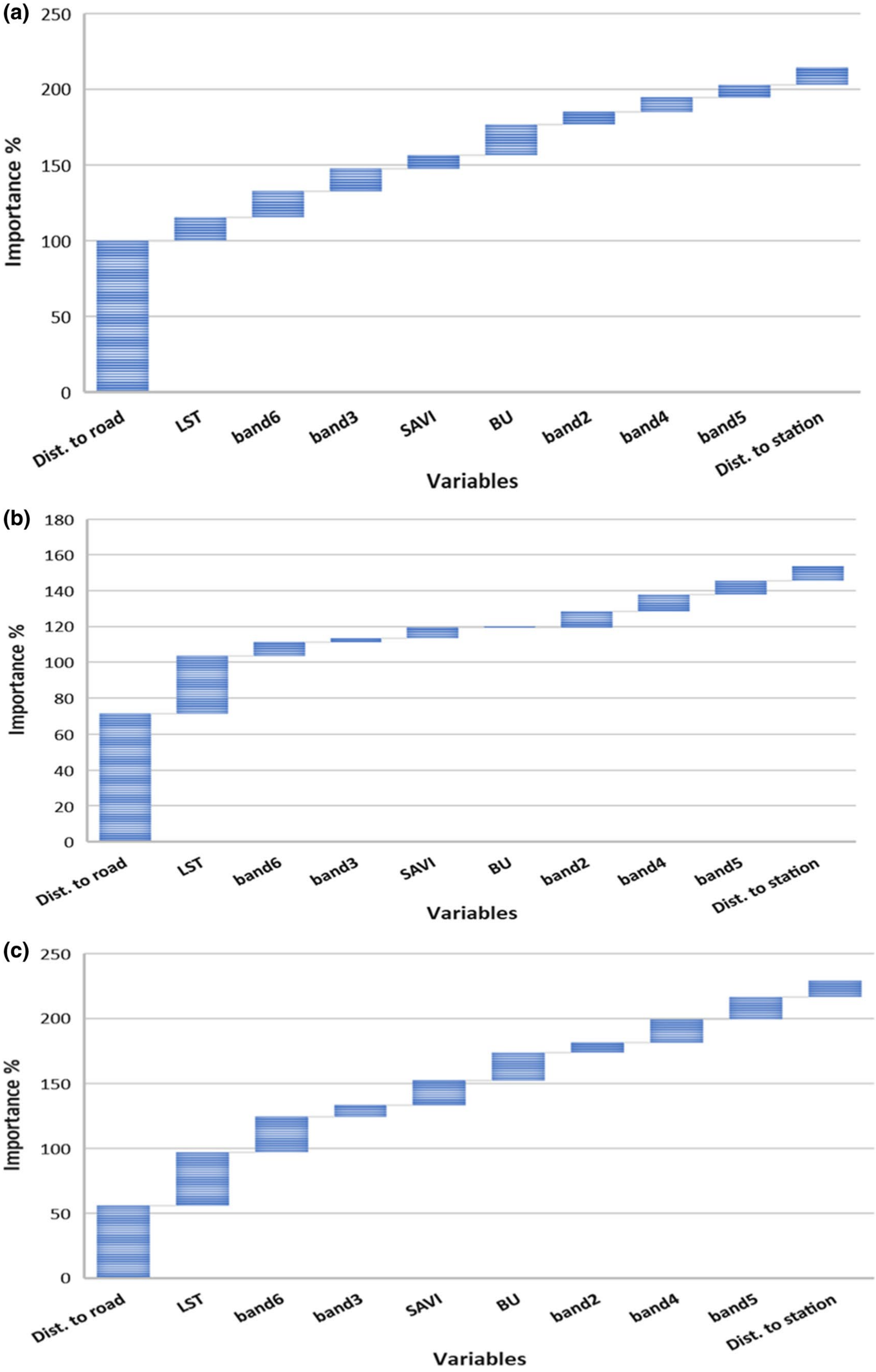

Fig. 3 Variable importance for applied features; **a** good, **b** moderate, and **c** hazard

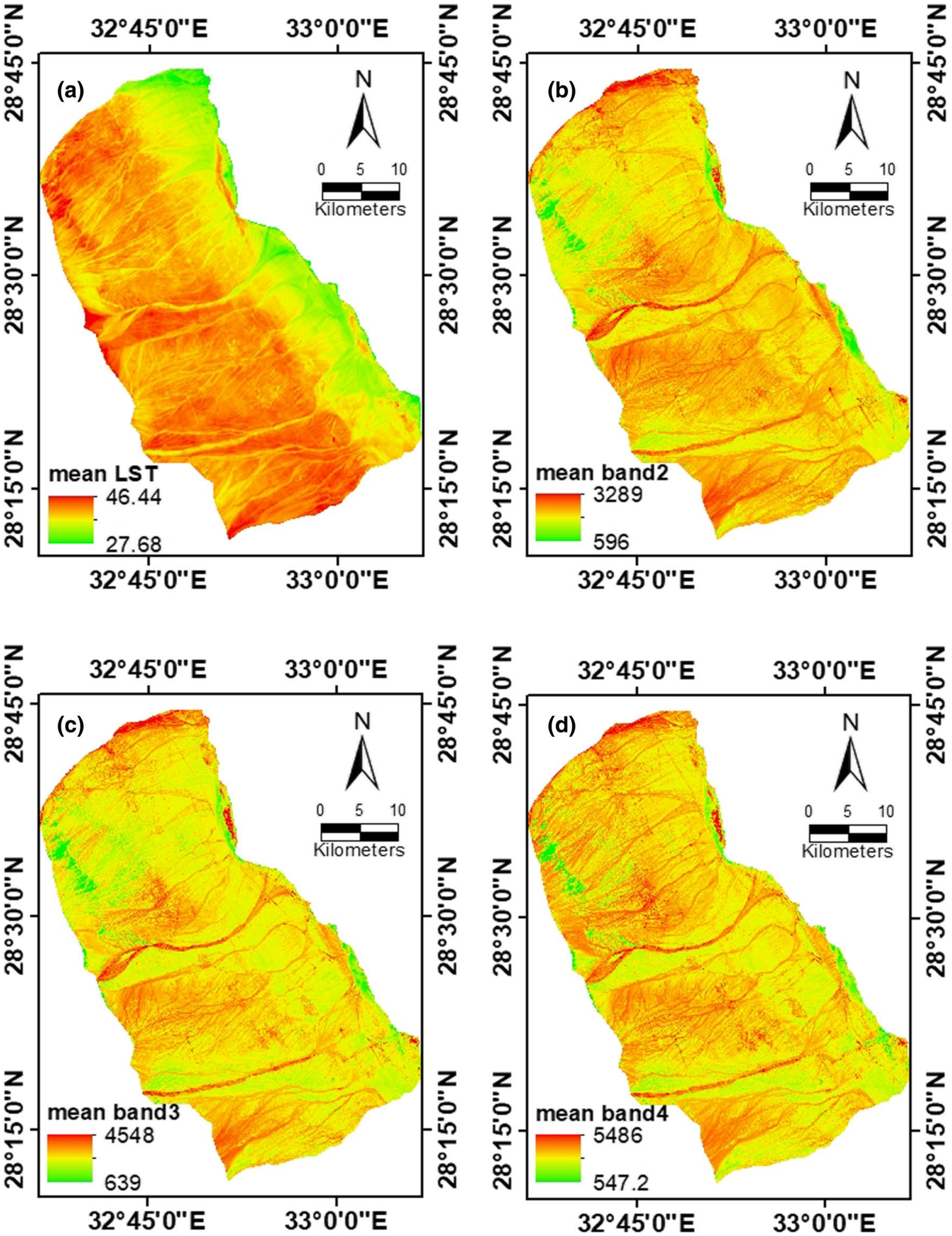

Fig. 4 Controlling factors of air pollution mapping; mean LST (**a**), mean band 2 (**b**), mean band 3 (**c**), and mean band 4 (**d**), Mean band 5 (**e**), mean band 6 (**f**), mean SAVI (**g**), and mean BU (**h**), Mean NDVI (**I**), distance from roads (**j**), and distance from stations (**k**)

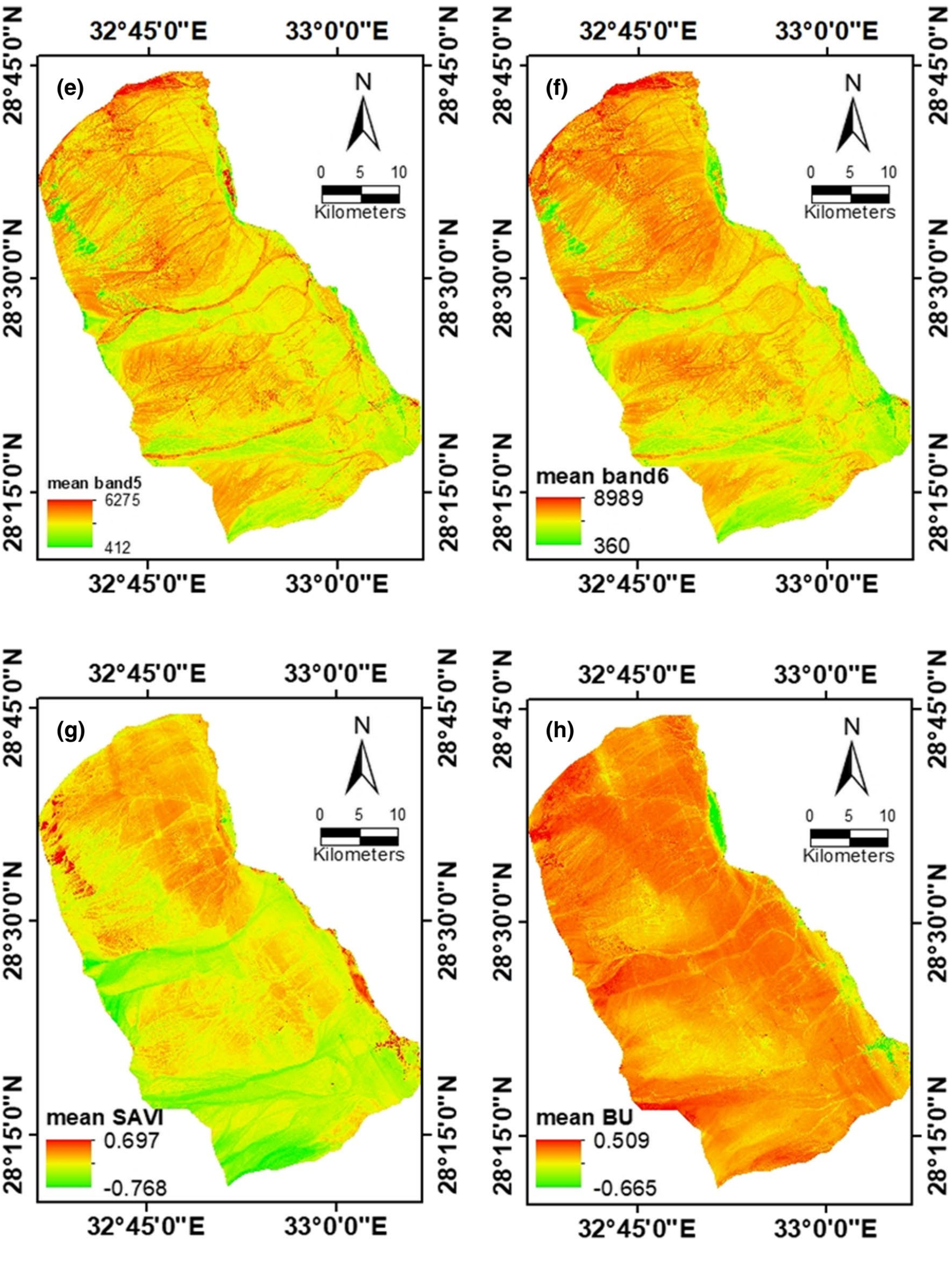

Fig. 4 (continued)

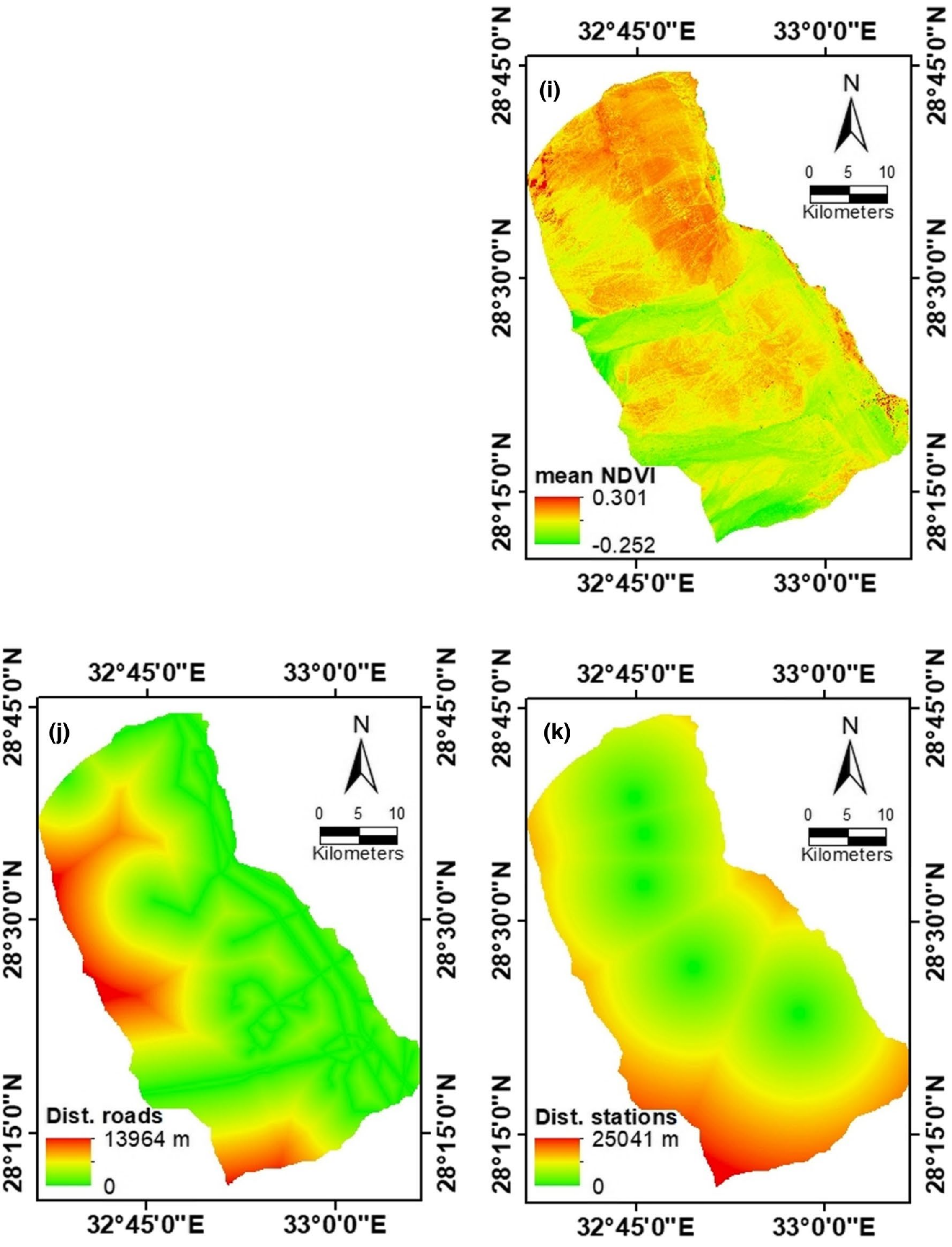

Fig. 4 (continued)

this area, the *SAVI* values range between − 0.77 and 0.69. SAVI can be calculated mathematically, from the following equation (Eq. 2).

$$SAVI = \frac{(NIR - R)}{(NIR + R + L)} \times (1 + L) \tag{2}$$

L donates the soil brightness correction factor and the accommodate value of *L* is 0.5 for most land cover types.

BU index: According to Warren et al. (2018) BU index (Eq. 3) is a result of subtraction of the NDVI (Eq. 5) from an NDB index (Eq. 4). Where NDB index refers to the normalized difference built-up index and can be calculated from equation (Eq. 5) (Zha et al. 2003). The Built-up index ranges in the area from -0.66 to 0.51, and the map shows a moderate to high distribution pattern in the area (Fig. 4).

$$BU = NDVI - NDB \tag{3}$$

$$NDB = \frac{SWIR - NIR}{SWIR + NIR} \tag{4}$$

Distance from the roads: The primary roads network in the areas was extracted from the open street map (www.openstreetmap.org). The extracted vector layers of the roads network were processed using ArcMap to create a thematic layer of distance from the roads network (Fig. 4). Euclidean distance tool in ArcMap is used to compute the distance from the station that evolves the air pollutants and measuring points (Fig. 4).

NDVI: can be calculated according to (Eq. 5) it refers to normalized difference vegetation index and its values always ranges from − 1 to + 1. Thus, the value of + 1 indicates a high possibility of dense vegetation, while the values of NDVI close to Zero mostly refers to urbanization. Where, the green vegetation reflects more near-infrared (NIR) and green light from the other wavelengths. The mean NDVI value of the study site ranges from −0.25 to + 0.30 reflecting absence of dense vegetation.

$$NDVI = \frac{NIR - Red}{NIR + Red} \tag{5}$$

Parameters optimization

Machine learning techniques are usually used for parameter optimization to improve the prediction model accuracy. The present study grid search method was applied for tuning parameters or input variables. The grid search approach is a classifier that can accurately predict the unlabeled dataset (Ataei, and Osanloo 2004). One of the key important advantages of RF is that includes a multitude of decision trees during the training time, that avoid overfitting with a single decision tree (Breiman 2001; Sahin et al. 2020). AlThuwaynee et al. (2021) concluded that the model tuning in RF, uses the parameters mtry, maxnodes, ntree, and nodesize. The optimal parameter for the best classifier is the number of trees which was in the present work is 300 that achieves the highest accuracy. Experimental hyperparameter results on the PM10 dataset show that the highest accuracy obtained with the Random Forest approach is (0.9829) which is higher than the accuracy (0.9655) of the prediction model without parameter tuning. Furthermore, the computational time within the modeling required to find good parameters using the grid-search method is not much.

Air pollution prediction

Five stations in the study site (North Ras Gharib area) were used to collect the PM10 data for 4 years period (2018–2021). The collected spatial dataset was prepared, stacked, and then normalized to be classified into a model training dataset and testing dataset. Then the seed value has to be assigned to ensure similar random distributions in each loop of the algorithm process to avoid optimization failure (AlThuwaynee et al. 2021). The East Arta station was classified as a good class, where the average measurements of PM10 during the study period were within the limits of clear and good air (EPA 1997) (Table 2). Three stations (Hana field, Hoshia field, and Northwest Gharib field) were classified as a moderate class, where the PM10 measurement values were within the moderate air quality limits (EPA 1997). In contrast, one station was classified as hazard class (Arta field), where the highest measured value was above the value of 424 μgm^{-3} recommended by EPA (1997). The RF model was developed with 10 variables and three PM10 index classes including good, moderate and hazard classes (Fig. 5). Eventually, the landsat 8 spectral bands 2, 3, 4, 5,

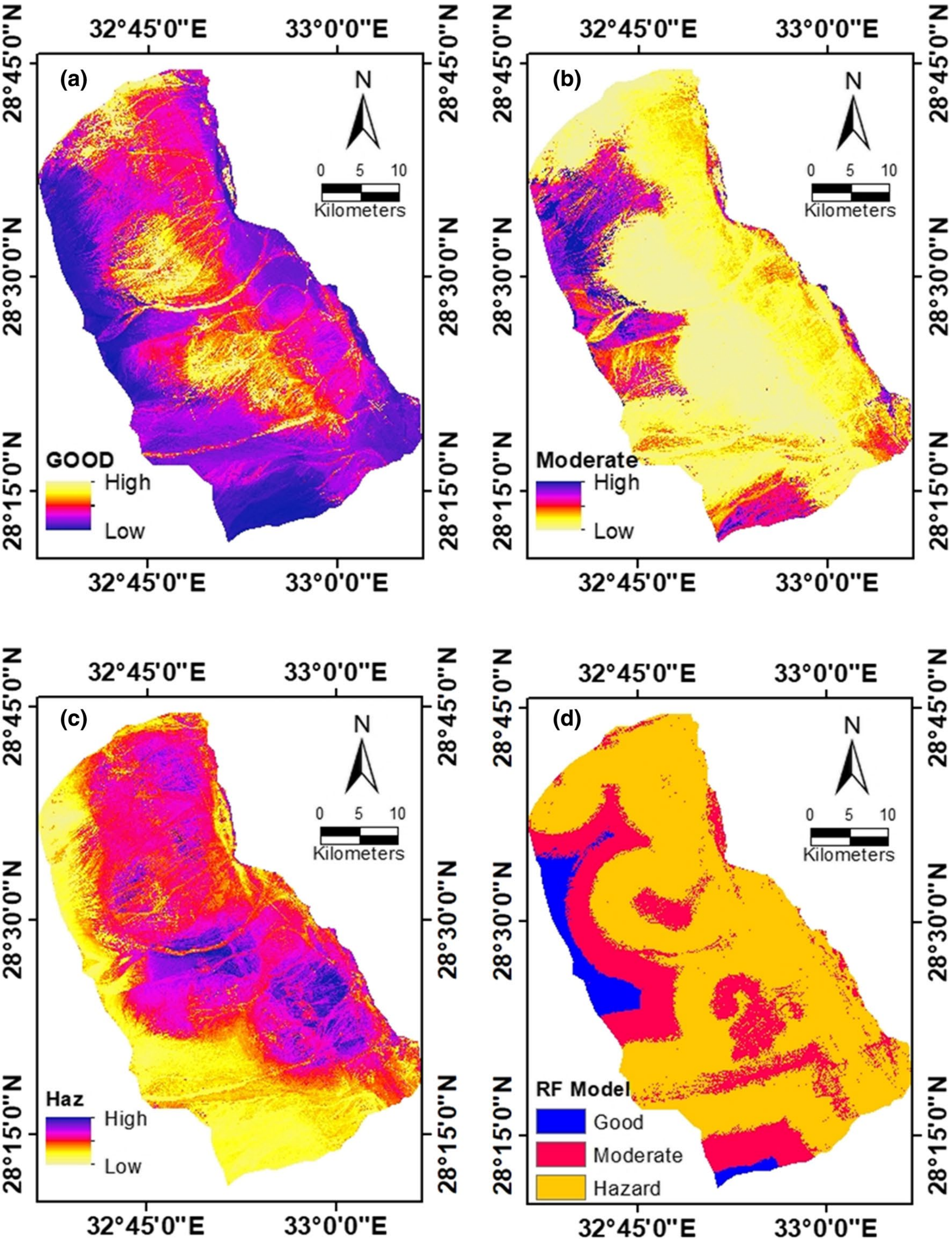

Fig. 5 Prediction models for different air quality classes; good class prediction (**a**), moderate class prediction (**b**), hazard class prediction, and random forest prediction model (**d**)

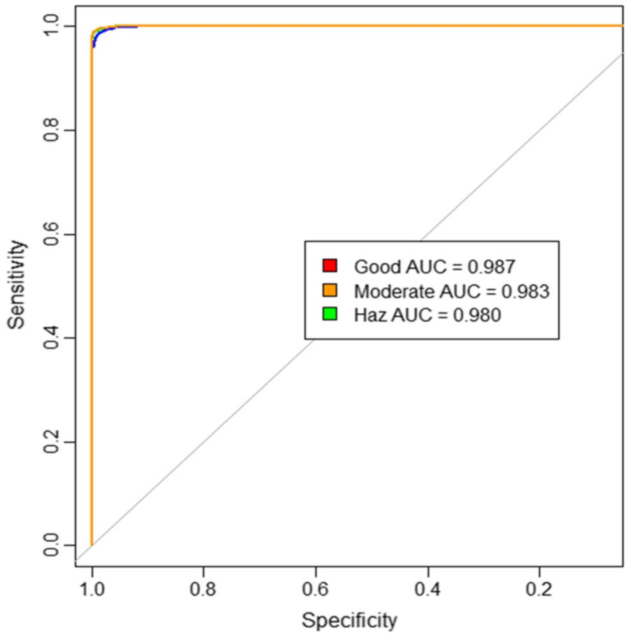

Fig. 6 Validation model of the dataset

Table 4 Summary of the prediction sub-classes percentage in RG area

Class	Area (Km^2)	%
Good	85	5%
Moderate	436	27%
Hazard	1095	68%

and 6 showed variable sensitivity to all PM10 classifications without any observed significant prediction capacity. Eventually, the landsat 8 spectral bands 2, 3, 4, 5, and 6 showed various response sensitivity patterns to all classifications of PM10 without any observed significant prediction capacity.

Spectral bands 2 and 5 show less response to the model, while band 6 has a high response to most of the predictive classes. Moreover, the distance from the roads thematic layer and land surface temperature (LTS) tends to show higher sensitivity to the prediction model. Of the three predictive modeling classes, the low values are located in the western part of all classes (Fig. 5). Also, this indicates that prediction model performance over different regions is similar.

Table 5 Accuracy and Kappa details of the predicted model

ML model	Accuracy details		Accuracy SD
	Accuracy	Kappa	
RF	98.29	96.55	0.0066

Land cover indices (SAVI, BU) show different response patterns to the prediction subclasses. However, SAVI indicates less response for the sub-classes among the land cove indices. Distance from the stations (inside the petroleum activities) has a clear relationship with the PM10 intensity. The predictive model results of the PM10 classes show the majority of the area of hazard class (Table 3). The predictive model indicates that the Arta station PM10 value highly impacts the area. The modeling hyperparameter optimization slightly enhances the results, increasing the model accuracy. However, the final prediction results match the measurement expectation since the uncertainty increases could be increased with the longer period due to the necessity of accurate classification of land cover land use of industrialization and other independent variables.

ML model validation

According to Erzin and Cetin (2013), the best fit of the model can be considered based on the higher accuracy of AUC of ROC values. However, AlThuwaynee et al. (2021) stated that the good performance on seen data doesn't necessarily suggest good performance on the unseen data. ROC approach was used by many researchers to evaluate the air quality pollution modes performance (Djalalova et al. 2010; Tamas et al. 2016; Zhang et al. 2017). Thus, the ROC considered the main accuracy measure in most literature. Accordingly, the ROC used to evaluate model performance (Fig. 6) achieved 98.29 AUC for the periods of 4-year period of data measurements. Table 4 summarizes the results of the model performance and accuracy. Nonetheless, the performance of Kappa was very reasonable comparing to the model accuracy. The standard deviation of different model accuracies was 0.006 (Table 4).

Monitoring program and future research

The lack of a comprehensive plan for monitoring with ignoring the environmental sustainability in the area, makes air pollution issues exacerbated. Extreme pollution was released from some stations in the present study. Air pollution monitoring can be achieved by creating general awareness about the impacts of air pollution and how it affects human well-being. This required that the information on air quality pollution be formed at all levels and is not compromised for pollution control. However, the current capabilities in air pollution monitoring need to be enhanced through the history recording of the dataset. Promoting the integration of the forecasting approach with real data can effectively improve risk management. Furthermore, low-cost sensors for air quality monitoring will be greatly increased the quality of the collected data, in order to improve the aggregation of data and management (Table 5).

Various fields of study related to or influenced air quality pollution need to be covered. Such significant future research in the area demands include air quality on the ecosystem in the area, incorporating with examinations and effects of different pollutants. Finally, any relevant health risks associated with air quality pollution should be communicated rather than communicated the measured pollution concentrations.

Conclusion

In the present work, RF algorithm was applied to assess and identify the distribution of PM10 in the North Ras Gharib area, Egypt, and the surrounding area. Such algorithms have the capabilities to visualize and optimize time-series data. The authors evaluate the relationship between PM10 with spectral bands of Landsat 8 OLI and land use indices. To achieve the aim of the present study time series data for the period between 2018 and 2021 were used. The PM10 dataset was collected from five measurement stations within the study site. Ten variables were applied in the predicted model including bands 2, 3, 4, 5, 6 and LST, SAVI, BU, distance from station, and distance from the roads. Results from the RF model indicated that most of the study site is of hazard or very high susceptibility for PM10. The prediction susceptibility is affected by the distance from the roads and land surface temperature. Further, the high PM10 concentration of Arta field has the have a major impact on the air quality pollution (PM10) for the distribution map on the area. However, band 4 and band 5 of Landsat imaginary were found to have little effect on the PM10 distribution model. The largest sources of PM10 in the area occurred in distant areas with large petroleum activities and production. Despite limited numbers of measuring stations for the PM10, high accuracy was achieved. The predicted model could be improved by increasing the data collection points in the area and considering wind speed and direction.

Acknowledgements The authors express their gratitude to Dara petroleum company, Egypt for providing air quality data records and measurements.

Author contribution SA: writing—original draft preparation, formal analysis conceptualization, methodology, writing-review and editing. MM: data curation, conceptualization. GS: writing—original draft preparation, review and editing. SK: writing-review and editing. All authors read and approved the final manuscript.

Funding Open access funding provided by The Science, Technology & Innovation Funding Authority (STDF) in cooperation with The Egyptian Knowledge Bank (EKB). We confirm that there is no financial or funding was received during this work.

Declarations

Conflict of interest This manuscript has not been published or presented elsewhere in part or in entirety and is not under consideration by another journal.

Ethical approval This article does not contain any studies with human participants or animals performed by any of the authors.

References

Abdallah AM, El-Adindany FN (1963) Stratigraphy of the lower Mesozoic Rocks, Western Side of Gulf of Suez, Egypt, Goel Surv Egypt 10(21)

Abu El-Magd SA (2022) Random forest and naïve bayes approaches as tools for flash flood hazard susceptibility prediction, South Ras El-Zait, Gulf of Suez Coast, Egypt. Arabian J Geosci.

Abu El-Magd SA, Pradhan B, Alamri A (2021) Machine learning algorithm for flash flood prediction mapping in Wadi El-Laqeita and Surroundings, Central Eastern Desert, Egypt. Arab J Geosci. https://doi.org/10.1007/s12517-021-06466-z

Abu El-Magd SA, Sk A, Pham QB (2021) Spatial modeling and susceptibility zonation of landslides using random forest, naïve bayes and K-nearest neighbor in a complicated terrain. Earth Sci Inf. https://doi.org/10.1007/s12145-021-00653-y

AlThuwaynee OF, Kim S, Najemaden MA, Aydda A, Balogun A, Fayyadh MM (2021) Park H (2021) Demystifying uncertainty in PM10 susceptibility mapping using variable drop-off in extreme-gradient boosting (XGB) and random forest (RF) algorithms. Environ Sci Pollut Res 28:43544–43566. https://doi.org/10.1007/s11356-021-13255-4

Ataei M (2004) Osanloo M (2004) Using a combination of genetic algorithm and the grid search method to determine optimum cutoff grades of multiple metal deposits. Int J Surf Min Reclam Environ 18(1):60–78

Ball J (1952) Contributions to The Geography of Egypt, Cairo

Breiman L (2001) Random forests. Mach Learn 45(1):5–32

Campanile L, Cantiello P, Iacono M, Lotito R, Marulli F, Mastroianni M (2021) Applying machine learning to weather and pollution data analysis for a better management of local areas: the case of Napoli, Italy. In: Proceedings of the 6th international conference on internet of things, big data and security (IoTBDS 2021), pp 354–363. https://doi.org/10.5220/0010540003540363

Diaz-Uriarte R, Andrés AD (2006) Gene selection and classification of microarray data using Random Forest. BMC Bioinformatics 7:3

Djalalova I, Wilczak J, McKeen S, Grell G, Peckham S, Pagowski M, DelleMonache L, McQueen J, Tang Y, Lee P, McHenry J (2010) Ensemble and bias-correction techniques for air quality model forecasts of surface O3 and PM2.5 during the TEXAQS-II experiment of 2006. Atmos Environ 44(4):455–467

Doreswamy, Harishkumar K S, Yogesh KM, Gad I (2020) Forecasting air pollution particulate matter (PM2.5) using machine learning regression models. In: Third international conference on computing and network communications (CoCoNet'19). Procedia computer science vol 171, pp 2057–2066

Erzin Y, Cetin T (2013) The prediction of the critical factor of safety of homogeneous finite slopes using neural networks and multiple regressions. Comput Geosci 51:305–313

Garcia JM, Teodoro F, Cerdeira R, Coelho RM, Kumar P, Carvalho MG (2016) Developing a methodology to predict PM10 concentrations in urban areas using generalized linear models. Environ Technol 37:2316–2325

General Petroleum Company (GPC) (1985) Stratigraphic succession of Ras Gharib area, Gulf of Suez, Egypt

Gonzalez Y, Carranza C, Iniguez M et al (2017) (2017) "Inhaled air pollution particulate matter in alveolar macrophages alters local pro-inflammatory cytokine and peripheral IFN production in response to mycobacterium tuberculosis." Am J Respir Crit Care Med 195:S29

Hvidtfeldt UA, Ketzel M, Sørensen M et al (2018) Evaluation of the Danish AirGIS air pollution modeling system against measured concentrations of PM2.5, PM10, and black carbon. Environ Epidemiol 2(2):2018

Izenman AJ (2008) Modern multivariate statistical techniques regression, classification, and manifold learning. Springer, New York

Jirat B, Chalat K, Papis W (2019) Machine learning algorithms for predicting air pollutants. E3S Web Conf. 120:03004. https://doi.org/10.1051/e3sconf/20191

Meshref WM, El-Gindy AK, Abdel-Rahman I (1983) Petrophysical study on subsurface Miocene formations of West Ras Gharib-Ras Shukheir area, Eastern Desert, Egypt : 8th Intern. Cong. Statist. Co. Sci. Soc. and Demograph Res., Ain Shams Univ., Cairo, pp 295–316.

Moorthy K, Mohamad MS (2011) Random Forest for gene selection and microarray data classification. Bioinformation 7(3):142–146

Muhammad I, Yan Z (2015) Supervised machine learning approaches: a survey. Ictact J Soft Comput. 5:946–952

Nisbet R, Elder J, Miner G (2009) Handbook of statistical analysis and data mining applications. Elsevier Academic Press, Burlington

Pahlavan-Rad MR, Akbarimoghaddam A (2018) Spatial variability of soil texture fractions and pH in a flood plain (case study from eastern Iran). CATENA 160:275–281

Park S, Kim M, Kim M, Namgung HG, Kim KT, Cho KH, Kwon SB (2018) Predicting PM10 concentration in Seoul metropolitan subway stations using artificial neural network (ANN). J Hazard Mater 341:75–82

Pimpin L, Retat L, Fecht D et al (2018) Estimating the costs of air pollution to the National Health Service and social care: an assessment and forecast up to 2035. PLoS Med 15(7):1–16

Raimondo G, Montuori A, Moniaci W, Pasero E, Almkvist E (2007) A machine learning tool to forecast PM10 Level. In: Proceedings of the fifth conference on artificial intelligence applications to environmental science, San Antonio, TX, USA, 14–18 January 2007; pp 1–9

Ramadan FS, El Nady MM, Hammad MM, Lotfy NM (2012) Subsurface study and source rocks evaluation of Ras Gharib onshore oil field in the central Gulf of Suez. Egypt Aust J Basic & Appl Sci 6(13):334–353

Ruiyun Y, Yang Y, Yang L, Guangjie H, Oguti AM (2016) RAQ–a random forest approach for predicting air quality in urban sensing systems. Sensors 16:86. https://doi.org/10.3390/s16010086

Steinberg D, Golovnya M (2013) Tree ensembles and extensions, an overview of tree net, random forests, ISLE model compression and rule learner (Salford-Systems, San Diego, CA, 2013), available at http://cdn2.hubspot.net/hub/160602/file-246947114-pdf/docs/JSM_2013_CTW_Slides/2013_TN_RF_ISLE_RL_CTW.pdf

Stern RJ (1994) Arc assembly and continental collision in the Neoproterozoic east African orogen: implications for the consolidation of Gondwanaland. Annu Rev Earth Planet Sci 22:319–351

Stern RJ, Hedge CE (1985) Geochronologic constraints on late Precambrian crustal evolution in the eastern desert of Egypt. Am J Sci 285:97e127

Tamas W, Notton G, Paoli C, Nivet ML, Voyant C (2016) Hybridization of air quality forecasting models using machine learning and clustering: An original approach to detect pollutant peaks. Aerosol AirQual Res 16(2):405–416

US Environmental Protection Agency (US EPA) (2015) Criteria air pollutants, America's Children and the Environment, US EPA, Washington, DC, USA

Xie Y, Zhu C, Zhou W, Li Z, Liu X, Tu M (2018) Evaluation of machine learning methods for formation lithology identification: A comparison of tuning processes and model performances. J Petrol Sci Eng 160:182–193

Yu R, Yang Y, Yang L, Han G, Move OA (2016) RAQ–A Random forest approach for predicting air quality in urban sensing systems. Sensors 16:86

Zha Y, Gao J, Ni S (2003) Use of normalized difference built-up index in automatically mapping urban areas from TM imagery. Int J Remote Sens 24(3):583–594

Zhang ZH, Hu MG, Ren J, Zhang ZY, Christakos G, Wang JF (2017) Probabilistic assessment of high concentrations of particulate matter (PM10) in Beijing. China Atmosph Pollut Res 8(6):1143–1150

Zhao G, Pang B, Xu Z, Yue J, Tu T (2018) Mapping flood susceptibility in mountainous areas on a national scale in China. Sci Total Environ 615:1133–1142