Review Article

Application ofMachine Learning to Air Pollution Studies: A Systematic Review

Abstract

Air pollution is a serious global issue that threatens human life and health, as well as the environment. Machine learning aAlgorithms can be used to predict air pollution level data from both natural and anthropogenic activities. Environmental and government agencies can use these speculations to issue air pollution alerts. This review work is an attempt at the recent status and development of scientific studies on the use of machine learning algorithms to model air pollution challenges. This study uses the scientific web as a primary search engine and covers over 100 highly peer-reviewed articles from 2000-2022. Therefore, this review paper aims to highlight the various application methods of machine learning, notably data mining, in air pollution control and monitoring. It also comprehensively analyses published works by renowned scholars and authors worldwide, discussing how machine learning has been used in mitigating air pollution. By examining the chronological trends of machine learning in air pollution, this review paper provides an up-to-date account of the successes achieved in regulating air pollution using machine learning techniques. Additionally, it identifies areas that require further research, critically analyzing the current state of knowledge and potential research directions.

Keywords: Machine Learning, Environmental Pollution, Air Pollution, Algorithm, Anthropogenic

1. Introduction

In developed and developing countries, increases in industrial development and urbanization are the key factors that have caused air quality reduction. This affects human health and the environment at large. Some studies have been done on air quality forecasting using machine learning and artificial intelligence to control air pollution. As such, there is significant research and review work that has been done on this study.

Environmental pollution refers to harmful substances or pollutants in the natural environment, adversely impacting living organisms and the environment. The diverse forms of pollution, such as air, water, soil, and radioactive decay, pose significant threats to public health and environmental sustainability (Remoundou and Phoebe, 2009).

Machine learning is a sub-division of aArtificial Iintelligence based on the bBiological lLearning process that incorporates various complex procedures.Machine learning has been widely adopted in pollution control and management, ranging from health and environmental studies to agriculture. Specifically, machine learning has been applied in pollution investigation, mitigation, and control. For instance, machine learning is used in water pollution to develop a competent and robust approach to estimating water quality characteristics (Mengyuan et al., 2022; Aniwetalu et al., 2018; Nwaezeapu et al., 2018; Ibekwe et al., 2023). In addition, recent

trends and advances in computer-aided environmental data engineering have also facilitated the adoption of machine learning in environmental pollution management.

Air pollution is one of the world's most significant environmental and health problems in indoor and outdoor contexts. According to recent statistics, air pollution contributes to 11.65% of global deaths (Hannah and Roser., 2017). Machine learning can address this problem by identifying pollution sources, predicting air quality, developing early warning systems, and monitoring air quality.

The sources of pollution are diverse, encompassing industrial, mechanical, agricultural, domestic, and natural factors, which release pollutants into the air, water, and land, among others. Moreover, anthropogenic and natural activities contribute to pollution, resulting in adverse environmental consequences such as climate change, health hazards, and ecological imbalances.

One aspect of machine learning that will be reviewed in this paper is deep learning, using data mining. Data mining involves the integration of statistics, artificial intelligence, and machine learning to predict outcomes by discovering patterns, trends, and insights from large data sets. By leveraging this approach, environmental pollution management can be more efficient, effective, and sustainable. However, machine learning can be applied to air pollution in (a) Identification of pollution sources, (b) prediction of air quality, (c) Development of early warning systems (c) monitoring of air quality.

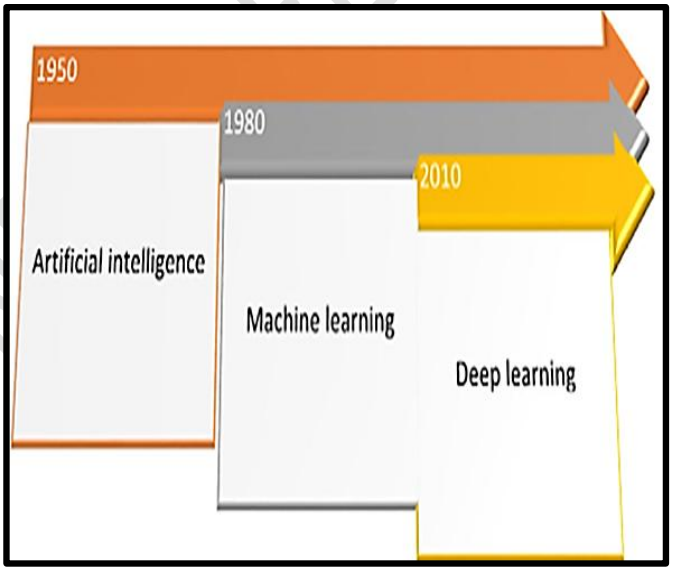

Fig.1. Evolution of Artificial Intelligence, Machine Learning, and Deep Learning. (Copeland, 2016; Ayturan et al., 2018).

2. Sources of Air Pollution

These sources damage the human respiratory system and increase a person's vulnerability to respiratory infections and asthma. Large amounts of nitrogen dioxide are also harmful to plants, and they can damage the plant's leaves and reduce crop growth (Ochando et al., 2015).

PM10: These suspended particles have an aerodynamic diameter of fewer than 10 micrometres. These particles reduce visibility in the city. They are produced mainly by industrial activities, which increase in relatively humid weather. These particles enter the body through the throat and nose, cause serious side effects on the lungs, and increase cancer risk (Alkasassbeh et al.,2013; Vitolo et al., 2018).

PM2.5: This pollutant is one of the most dangerous particles. The size of these particles is so small that the body's natural filters cannot purify them. If a person is exposed to this particle for a long time, the probability of death will be high (Ochando et al., 2015).

O_3 (Ozone): The ozone in the air can harm health, particularly on warm shining days when ozone attains toxic levels. Even relatively low levels of ozone can have health consequences. Deep breathing can be difficult, depending on the amount of ozone exposure. Chest pain occurs during deep breathing. The frequency of asthma attacks increases, and the lungs are more predisposed to infection. These consequences have been detected even in healthy people (Jekin and Clemitsha, 2000).

Formatted: Subscript

3. Collection of Reviewed Papers

We selected articles for this review work from reputable science journals and extracted our information from journals dating between 2000 and 2020. Below shows the number of studied articles by year of publication, and the percentage of articles based on Sources of air pollution can be classified into two major categories:

(1) Natural Sources: Natural air pollution includes forest fires, volcanic eruptions, sandstorms, and gas leaks. These events are natural, and the pollution caused by them cannot be controlled. Humans do not play a significant role in producing this air pollution (Alireza and Mohammed, 2022; Oguadinma et al., 2014).

(2) Man-made Sources Human-made plays a more significant role in air pollution caused by human activities (Bai et al., 2018; Nunez, 2019; Ibekwe et al., 2023). These sources of air pollution include:

(i) Home resources: This type of pollution occurs due to indoor activities such as cooking and lighting a fireplace.

(ii) Industrial resources: This type of pollution is caused by the activities of factories, power plants and petrochemicals factories.

(iii) Dynamic resource: This type of pollution is caused by transportation and it is affected by the development of cities (Alireza and Mohammed, 2022).

(iv) Agricultural resources: This type of pollution occurs when soil is ploughed, and roots are exposed to air and ~~oxidised~~oxidized to carbon dioxide. This is good for weed control but also releases carbon. Various factors are affecting the spread of air pollution in cities.~~described in the Table below:~~

Table 1 shows t~~T~~he sources, acceptable concentrations and effects of the six significant pollutants ~~are summarized in the table below~~.

Table 1~~shows the~~ S~~s~~ix major pollutants and their effects (U.S Environmental Protection Agency, 2023).

Pollutant	Common sources	Maximum acceptable concentration	Environmental risks	Human health risk
Carbon monoxide (CO)	automobile emissions, fires, industrial processes	35 ppm (1-hour period); 9 ppm (8-hour period)	contributes to smog formation	exacerbates symptoms of heart disease, such as chest pain; may cause vision problems and reduce physical and mental capabilities in healthy people
Nitrogen Oxides (NO and NO_2)	Automobile emission and electricity generation.	0.053 ppm (1-year period)	Damage to foliage; contribute to smog formation	Inflammation and irritation of breathing passages
Sulphur Oxide (SO_2)	E~~e~~lectricity generation, fossil-fuel combustion, industrial processes, automobile emissions	0.03 ppm (1-year period); 0.14 ppm (24-hour period)	T~~t~~he primary cause of haze; contributes to acid rain formation, which subsequently damages foliage, buildings, and monuments; reacts to form particulate matter	B~~b~~reathing difficulties, particularly for people with asthma and heart disease
Ozone (O_3)	Nitrogen oxides and volatile organic compounds from industrial and automobile emissions	0.075 ppm (8-hour period)	Interferes with the ability of certain plants to respire, leading to the susceptibility to other environmental stressors (e.g. disease)	Reduced lung function; irritation and inflammation of breathing passages
Particulate matter	sources of primary particles include fires, smokestacks, construction sites, and unpaved roads; sources of secondary particles include reactions between gaseous chemicals emitted by power plants	150 µg/m^3 (24-hour period for particles <10 µm); 35 µg/m^3 (24-hour period for particles <2.5 µm)	C~~c~~ontributes to the formation of haze as well as acid rain, which changes the pH balance of waterways and damages foliage, buildings, and monuments	I~~i~~rritation of breathing passages, aggravation of asthma, irregular heartbeat
Lead (Pb)	Metal processing, waste incineration, fossil fuel combustion	0.15 µg/m^3 (3 months average)	Loss of biodiversity, decreased reproduction	The adverse effect on multiple body systems; may contribute to learning disabilities when exposed to young children

Table 2~~shows the t~~Types of factors that affect the spread of air pollution in various cities (Alireza and Mohammed, 2022).

Types	Meaning of factors
Meteorological parameters	This parameter refers to variables that describe the chemistry of the atmosphere. These include wind speed, wind direction, and relative humidity, which have a broad impact on the dispersion and concentration of pollutants. An increase in their temperature can cause the pollutants to move.
City morphology	This includes road network, population density, and land use type.
Types of pollutant	The type of pollutant can be either gas or particle.
Street traffic	The volume, speed, and duration of traffic in each environment can cause different degrees of dispersion in pollutants.

Pollutants are suspended particles that darken the air in an environment. The higher the number of particles in the air, the more pollution occurs. The durability of these particles in the air is between a few instances to a few months. The types of air pollutants are as follows:

(i) CO (carbon monoxide): Carbon monoxide is a toxic, odorless and colorless gas. Its commonly released into the environment through the exhaust of automobiles. This gas blocks the stream of oxygen to the heart and brain by blocking the blood, which can eventually lead to death (Ochando et al., 2015; Ghadi et al., 2019).

(ii) NOx (Nitrogen oxide): Nitrogen oxide comes in two forms, nitrogen monoxide (NO) and nitrogen dioxide (NO_2), which are gasses produced from natural sources, motor vehicles, and other combustion processes. Increased nitrogen dioxide levels can on the type of pollutant studied.

Formatted: Subscript

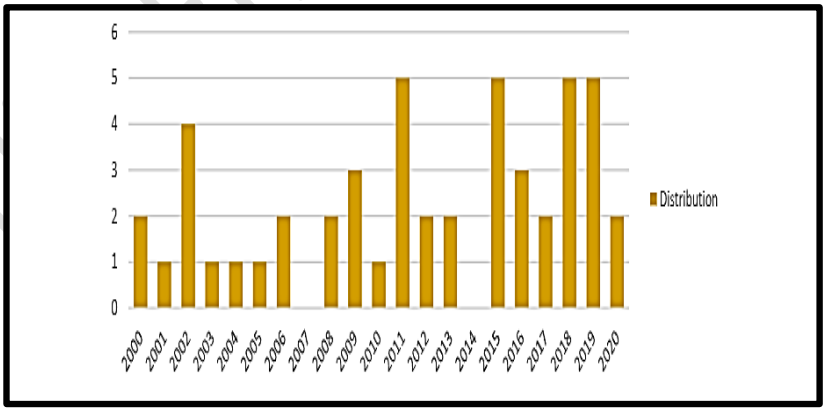

Fig.2. Number of reviewed articles by year of publication (Alireza and Mohammed, 2000).

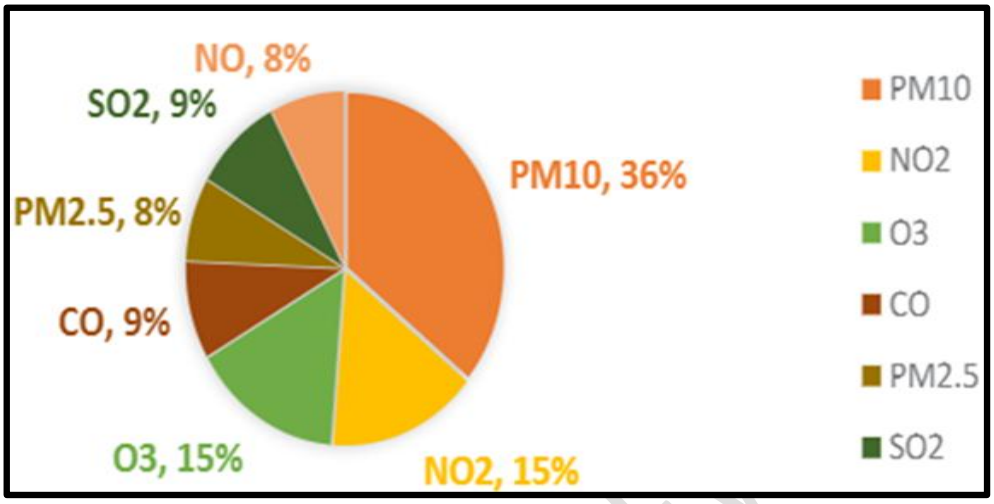

Fig.3. Percentage of reviewed articles based on the type of pollutant(Alireza and Mohammed, 2000).

4. Air Pollution by Oil and Gas Exploitation

Air pollution resulting from oil and gas reservoir exploitation (Oguadinma et al., 2014; Oguadinma et al., 2016; Oguadinma et al., 2017; Nwaezeapu et al., 2018; Oguadinma et al., 2021) is a significant environmental concern. Throughout the various stages of oil and gas production, including extraction, processing, and transportation, a wide range of pollutants are released into the atmosphere, posing risks to human health and the environment.

The extraction and production of oil and gas release substantial amounts of pollutants into the air. During drilling operations, diesel-powered machinery emits nitrogen oxides (NOx), volatile organic compounds (VOCs), and particulate matter (PM) (Smith et al., 2018). These emissions contribute to the formation of ground-level ozone, a key component of smog, and have detrimental effects on air quality and respiratory health (Jones & Brown, 2019). Methane, a potent greenhouse gas, is also released during the extraction and flaring of natural gas, exacerbating climate change (Nwaezeapu et al., 2018; Ibekwe et al., 2023; Aniwetalu et al., 2018).

The processing and refining of oil and gas contribute to air pollution. Refineries emit pollutants such as sulfur dioxide (SO_2), nitrogen dioxide (NO_2), and carbon monoxide (CO) during the refining process (Ibekwe et al., 2023). These emissions have been associated with respiratory and cardiovascular health issues. Volatile organic compounds (VOCs) produced during the refining

Formatted: Subscript

Formatted: Subscript

process contribute to the formation of ground-level ozone, which poses risks to human health and ecosystems (Smith et al., 2018).

Transportation and Distribution: The transportation and distribution of oil and gas also contribute to air pollution. Diesel-powered trucks and vessels used for transportation emit nitrogen oxides (NOx) and particulate matter (PM), leading to poor air quality and adverse health effects (Pwavodi et al., 2023). Pipeline leaks and accidental releases further exacerbate air pollution, with methane emissions contributing to climate change (Smith & Davis, 2019).

Health Impacts: The release of pollutants during oil and gas reservoir exploitation has severe implications for human health. Exposure to particulate matter (PM) emitted from these activities is associated with increased risks of respiratory diseases such as asthma and chronic obstructive pulmonary disease (COPD) (Ibekwe et al., 2023).

5. Pollution prediction approaches

Kridgestatistical method: It is an interpolation method, a geostatistical tool used in the machine learning approach. It can be used to analyze and predict unknown values. Kridge developed this model in 1951. This is a statistical method used to estimate random values at unknown points from known observable points. This tool uses a Variogram (Shad et al., 2009; Stein et al., 2001).

5.1. How machine learning works

A machine learning system learns from historical data, builds a prediction model, and when it receives a new data, it predicts its outcome using the already existing information (Pwavodi et al., 2023). The accuracy of its prediction will be dependent on the volume of data. A large amount of data helps to build a better model which predicts outcomes more accurately. Using algorithms helps build the machine logic per data and predict the outcome. ~~Below is a~~A block diagram that explains the working principle of the machine learning algorithm is presented in the Fig.4.~~:~~

5.2. Features of machine learning

(i) Machine learning uses data to detect patterns in a given dataset.

(ii) It can learn from past data and improve automatically.

(iii) It deals with large amounts of data, similar to data mining.

(iv) It is a data-driven technology.

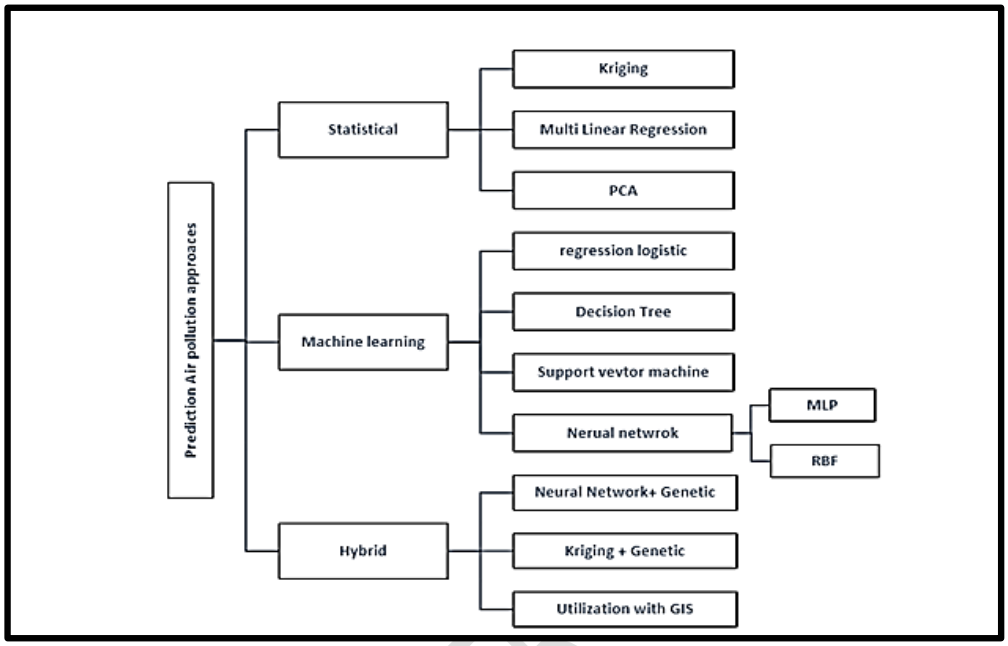

Fig.4. Flowchart of air pollution prediction methods (Alireza and Mohammed, 2000).

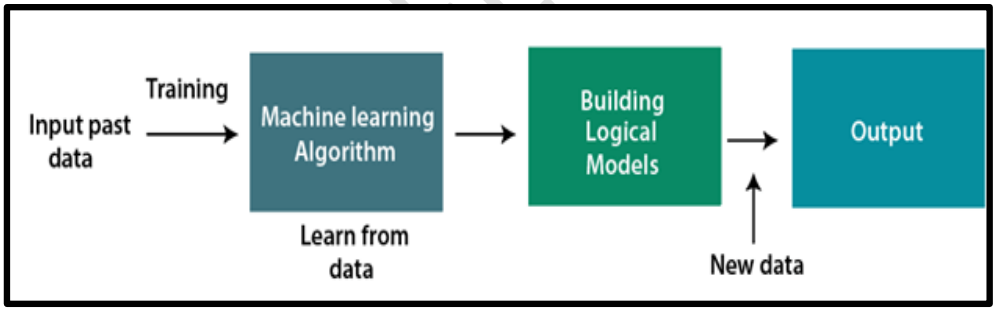

Fig.5. Shows the working principle of machine learning(Sonno Jaiswal, 2023).

5.3. Importance of machine learning

It solves complex problems, which is difficult for man.

It is used for decision-making in various sectors.

Increment in the production of data.

Finding hidden patterns and extracting useful information from them.

5.4. Classification of machine learning

Machine learning is broadly classified into three categories; supervised learning, reinforcement learning and unsupervised learning.

(1) Supervised learning: This type uses provided sampled data to predict outputs.

(2) Reinforcement learning: This is a feedback-based learning method in which the learning agent gets a reward for each right action and gets a penalty for each wrong action as well.

(3) Unsupervised learning: This is a type of learning in which the machine learns without supervision. The goal is to restructure the input data into a new feature with a similar pattern.

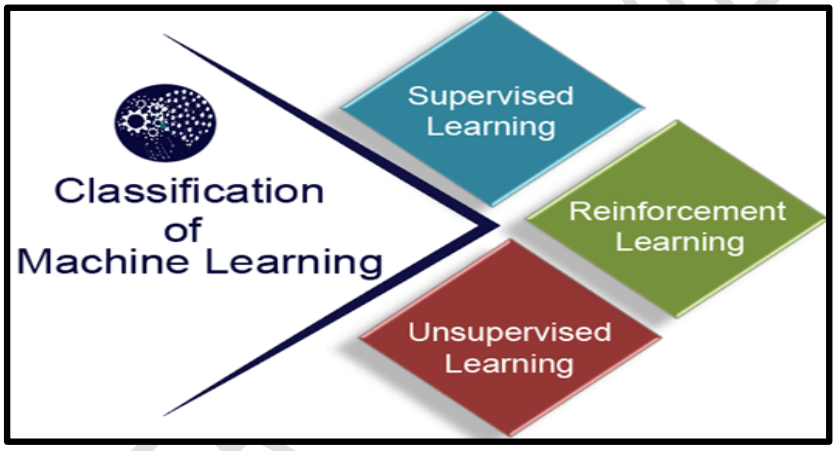

Fig. 6. Shows the classification of machine learning (Sonno Jaiswal, 2023).

5.5. Logistic regression method

Logistic regression is a robust supervised machine learning algorithm for classification while tackling binary problems. It essentially uses a logistic function to model binary output variables (Tolles and Meurer, 2016). It does not require a linear relationship between input and output variables. Logistic Regression Logistic regression (LR) is a distribution algorithm that estimates discrete values based on a given set of independent variables. This algorithm determines the prediction probability, so its costs are between 0 and 1. Backwards and forward stepwise variable selection algorithms are used for building models.

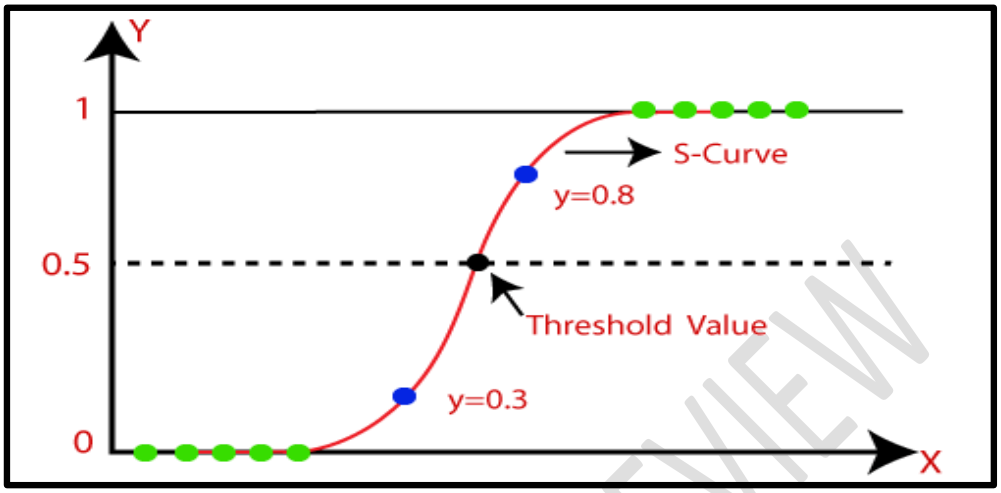

Fig. 7. Shows the linear regression model of machine learning (Sonno Jaiswal, 2023).

5.5.1. Logistic function (Sigmund function) characteristics

(i) The Sigmund function uses its mathematic functions to map predicted values to probabilities.

(ii) It maps absolute values into another value within 0 and 1.

(iii) The value of the logistic regression ranges between 0 and 1. This limit cannot be exceeded; it forms a curve like the S form.

5.5.2. Assumptions for Logistic Regression

The dependent variable must be categorical.

The independent variable should not have multi-collinearity.

5.5.3. Types of logistic regression

Based on category, logistic regression can be classified into three types:Binary, multinomial and ordinal. Each of these categories has a different function.

Conclusion

Recently, research in artificial intelligence and machine learning has been exponentially explored. Prediction of air pollutants can be quite challenging because of our dynamic environment and variability in space and time of the pollutants. Most of the scholars involved are from China and the United States of America. The prediction of air pollutant concentrations is a current research hotspot. The machine learning technique is one of the most effective

methods that can be used to monitor and control air pollution. Proper understanding of both the analytical and statistical tools of machine learning will be helpful in air pollution studies. Over time, more modern approaches have been adopted to improve the air quality of our environment.

References

AlirezaZhalehdoost and Mohammed Taleai, (2022). A Review of the Application of Machine Learning and Geospatial Analysis Method in Air Pollution Analysis. International Journal of Pollution and Environmental Sciences. https://doi.org/10.22059/poll.2022.336044.1300.

Aniwetalu, et al. Spectral analysis of Rayleigh waves in the Southeastern part of Niger Delta, Nigeria. Int J Adv Geosci. 2018;6:51-6. Available: http://dx.doi.org/10.14419/ijag.v6i1.8776

Bai, L., Wang, J., Ma, X. and Lu, H. (2018). Air pollution forecasts: An overview. International journal of environmental research and public health, 15(4), 780. https://doi.org/10.3390/ijerph15040780.

David L. Banks and Stephen E. Fienberg (2003). Data Mining, Statistics. Encyclopedia of Physical Science and Technology (Third Edition). Pp.247-261. https://doi.org/10.1016/B0-12-22741-2.

Ghadi, M. E., Qaderi, F. and Babanezhad, E. (2019). Prediction of mortality resulted from NO 2concentration in Tehran by Air Q+ software and artificial neural network. International Journal ofEnvironmental Science and Technology, 16(3), 1351-1368. https://doi.org/10.1007/s13762-018-1818.

GicelaLuperaCalahorrana, Ahmed ShokryAbdelaleem, Sergio Medina Gonzalez, Antonio Espuna (2018). Ordinary Kriging: a machine learning tool applied to mix-integer multiparametric approach. 28th European Symposium on Computer Aided Chemical Engineering, vol 43 ,pp. 531-536. https://doi.org/10.1016/B978-0-444-64235-6.50094-2.

Hannah and Max Roser., (2017).” Our world in data” Contributes to Raising Awarness and Combating Climate Change. Regional and Business Studies, vol. 11, no. 2, pp. 87-92. doi: 10.33568/rbs.2411.

Ibekwe KN, Arukwe C, Ahaneku C, et al. Enhanced hydrocarbon recovery using the application of seismic attributes in fault detection and direct hydrocarbon indicator in Tomboy Field, western-Offshore Niger Delta Basin. Authorea; 2023.

Ibekwe KN, Oguadinma VO, Okoro VK, Aniwetalu E, Lanisa A, Ahaneku CV. Reservoir characterization review in sedimentary basins. Journal of Energy Research and Reviews. 2023;13(2):20-28.

Jenkin, M. E. and Clemitshaw, K. C. (2000). Ozone and other secondary photochemical pollutants:chemical processes governing their formation in the planetary boundary layer. AtmosphericEnvironment, 34(16), 2499-2527. https://doi.org/10.1016/S1352-2310 (99)00478-1.

Joshua Pwavodi, Ibekwe N. Kelechi, PerekebinaAngalabiri, Sharon ChiomaEmeremgini, Vivian O. Oguadinma, Pore pressure prediction in offshore Niger delta using data-driven approach: Implications on drilling and reservoir quality, Energy Geoscience, Volume 4, Issue 3, 2023.

KyriakiRemoundou and Phoebe Kaundori (2009). Environmental Effects on Public Health: An Economic Perspective. International journal of Environmental Research and Public Health, 6(8), pp.2160-2178. **https://doi.org/10.3390/ijerph6082160**.

Mengyuan Zhu, Jiawei Wang, Xiao Yang, Yu Zhang, Linyu Zhang, Hongqiang Ren, Bing Wu, Lin Ye,6. A review of the application of machine learning in water quality evaluation, Eco-Environment &Health,Volume 1, Issue 2,2022,).

Nwaezeapu VC, Tom IU, David ETA, Vivian OO. Hydrocarbon Reservoir Evaluation: a case study of Tymot field at southwestern offshore Niger Delta Oil Province, Nigeria. NanosciNanotechnol. 2018;2(2).

Oguadinma et al. An integrated approach to hydrocarbon prospect evaluation of the Vin field, Nova Scotia Basin. S.E.G. technical program expanded Abstracts. International Exposition and Annual Meeting, Dallas, Texas. 2016;99-110. Available:10.1190/segam2016-13843545.1

Oguadinma et al. Lithofacies and Textural Attributes of the Nanka Sandstone (Eocene): proxies for evaluating the Depositional Environment and Reservoir Quality. J Earth Sci Geotech Eng. 2014;9660:4(4)2014:1-16ISSN: 1792-9040 (print).

Oguadinma et al. Study of the Pleistocene submarine canyons of the southeastern Niger delta basin: tectonostratigraphic evolution and infilling Conference/Reunion des sciences de la Terre, Lyon, France; 2021

Oguadinma et al. The art of integration: A basic tool in effective hydrocarbon field appraisal, Med-GU Conference, Istanbul. Turkey; 2021.

Oguadinma V, Okoro A, Reynaud J, Evangeline O, Ahaneku C, Emmanuel A et al. The art of integration: A basic tool in effective hydrocarbon field appraisal. Mediterranean Geosciences Union Annual Meeting; 2021.

Oguadinma VO, Aniwetalu EU, Ezenwaka KC, Ilechukwu JN, Amaechi PO, Ejezie EO. Advanced study of seismic and well logs in the hydrocarbon prospectivity of Siram Field, Niger delta basin. Geol Soc Am Admin Programs. 2017;49. DOI: 10.1130/abs/2017AM-296312

Remoundou and Phoebe KoundouriInt., (2009). Environmental Effects on Public Health: An Economic Perspective. International Journal of Environmental Research on Public Health, 6(8): 2160–2178).

SonnoJaismal 2023, Machine learning, java point, https://www.javatpoint.com, March 31, 2023.

Vivian OO, Kelechi IN, Ademola L, et al. Reservoir and sequence stratigraphic analysis using subsurface data. ESS Open Arch. February 09; 2023.

Vivian OO, Kelechi IN, Ademola L, et al. Submarine Canyon: A brief review. ESS Open Arch; 2023.

Yasin Akin AYTURAN.,ZeynepCansu AYTURAN., HüseyinOktay ALTUN (2018). Air Pollution Modelling with Deep Learning. International Journal of Environmental Pollution and Environmental Modelling, vol. 1, pp. 58-62.

Smith, A. B., et al. (2018). Air pollution emissions from onshore oil and gas exploration and production. Atmospheric Environment, 193, 1-10.