International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 11 Issue 12, December 2022

Air Pollution Prediction using Machine Learning Algorithms: A Systematic Review

Vidit Kumar, Sparsh Singh, Zaid Ahmed, Ms. Nikita Verma
Sharda University

***Abstract*:- The quality of the air has been steadily degrading over the last few years, which has raised the number of serious health issues. The government and researchers have taken a particular interest in creating and implementing methods and technologies that can not only aid monitoring present air quality but also being able to forecast the result due to this rapid rise. This study reviews several intriguing modelling strategies based on their general functionality, benefits, drawbacks, etc. Additionally, several data processing techniques are covered in order to increase the general efficiency of any model.**

1.INTRODUCTION

The necessity for urbanisation and an expanding population cause the air quality to worsen as modern civilization develops quickly. As there are more businesses and also automobiles on the road, different air pollutants combine and become even more dangerous. The Economist (2015) reports that air pollution is the leading cause of mortality in India, accounting for almost 1.6 million fatalities annually [1,2]. Additionally, air pollution has a negative psychological and emotional impact on people. People's anxiety and irritation are thought to grow as a result of poor air quality [3,4]. Exposure to air pollution might create anxiety because it increases systemic inflammation and oxidative stress [5].

Psychological effect	Author(s)	Pollutants measured
Anxiety related problems	Cho et al. [6]	Ozone (O_3)
	Sass et al. [7]	PM 2.5
	Rotko et al. [8]	NO_2, PM 2.5
	Xu et al. [9]	Haze
	Lu et al. [10]	Perceived pollution
Mental disorders	Pedersen et al. [11]	Benzene, CO
	Lin et al. [12]	SO_2
Suicidal thoughts	Ng et al [13]	SO_2, NO_2, PM 2.5
	Lee et al [14]	CO, NO_2, PM 10, SO_2

Table 1. Air pollutants and various psychological effects caused by them

1.1 *The major pollutants*

- The principal pollutants are SO2, total suspended particles (TSP) which include dust, PM10, and PM2.5, as well as Ozone gas (O3). Long-term vulnerability to poor air quality has been linked to an increased danger of heart attacks, strokes, lung cancer, and other serious life-threatening conditions, according to several studies [15]. The following are some of the hazardous particulates that can cause indigestion and have an impact on human health when inhaled:
- PM 10 - Particulate matters that are inhalable and are about 10 μm.
- PM 2.5 - Particulates as small as 2.5 microns that are inhalable by humans and represent a greater danger to the human health because some of them may even reach the bloodstream.
- SO2 – Sulphur Dioxide in the air is mostly produced by power plants via the coal combustion, oil, and other fossil fuels.
- O_3 – The photo - chemical reactions of several pollutants, including Nitrogen Oxide, VOCs (Volatile Organic Compounds) from UV radiation, Carbon Monoxide and harsh sunlight, result in the formation of O3 - Ozone gas.

1.2 *Current models*

Several deterministic models may be used to forecast air quality in order to indicate pollution trends in advance and assist people and the government in making educated environmental decisions. Some fundamental markers for the approximative categorization of air quality forecasting models include forecasting approach, input type, and time-scale categorization of forecasting models. As they concentrate on a smaller scale, shorter prediction models can be more exact, in-depth, and produce more accurate findings, but longer forecasting techniques can be used to gather long-term research results and as a tool for long-term pollution management in the future. Numerous air quality monitoring models have recently been created and put into use. The first type of model is called a CTM (chemical transport model), and its purpose is to explain the chemical and meteorological processes that affect the atmosphere, with a focus on the emission, movement, and blending of air pollutant concentrations [16]. To study the atmosphere and make predictions, a variety of WRF (weather research and forecasting) models are utilised. The WRF-Chem and WRF/Chem-MADRID models are a few of examples of WRF models. These models take into account variables including reaction index, reaction products, chemicals used, and dynamic situations. The formation of nonlinearity between the pollution concentration level and its source of distribution is the key component for further advancements. By using the stored past data to make predictions about the future, many models employ statistical methodologies to evaluate relationships between distinct air

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 11 Issue 12, December 2022

quality and air pollution elements in time series. Statistical approaches are thus used instead of traditional ones. They are data-based models based on mathematical statistics, stochastic processes, and probability. ARIMA (Autoregressive Integrates Moving Average) and GM (Grey model) are two examples of statistical models [19] [20]. Then there are regression models like Stepwise Regression [21], MLR (multiple linear regression) , and PCR (principal component regression)that are mostly used to forecast the concentration levels of pollutants [22][23] . Owing to recent advances in machine learning and artificial intelligence, when it comes to air pollution modelling, more intelligent predictors are capable of dealing with non-linearity and interacting correlations while also producing outputs with more accuracy [24]. They do not need to interpret the environmental conditions or other relevant aspects in their immediate surroundings, in contrast to the models outlined above [25]. Artificial neural networks (ANN), Backpropagation Neural Networks (BPNN), and generalised regression neural networks (GRNN) are a few popular examples of these models.

1.3 *Related Work*

In recent years, the society and the government have paid close attention to the deterioration in air quality, the common occurrences of air pollution, and the accompanying health repercussions. As a result, there is an urgent need for relevant and practical forecasting tools in scientific research. The basic forecasting techniques, comprising shallow predictors and deep learning predictors, are described as the basic forecasting models in this work by Hui Liu, Guangxi Yan, Zhu Duan, and Chao Chen. They discuss their foundations, applications, benefits, and drawbacks. For Support Vector Machine, Artificial Neural Network, Forest and Adaptive boosting (AdaBoost), this review's objective is to provide a comprehensive literature evaluation of sophisticated modelling techniques being used air quality forecasting, which may be helpful for future studies [1]. Yun-Chia Liang*, Josue Rodolfo Cuevas Juarez, Angela Hsiang-Ling Chen, and Yona Maimury presented their work, which focuses on the development of AQI prediction models for severe air pollution occurrences 1, 8, and 24 hours in advance. The research of the following machine learning (ML) techniques is done including support vector machines, artificial neural networks, random forest, and stacking ensemble approaches [2]. The recommended work by Madhuri VM, Samyama Gunjal GH, and Savitha Kamalapurkar used the supervised learning approach. Several algorithms, such as Naive Bayes, Linear Regression, SVM, Kernel SVM, Nearest Neighbor, and Random Forest, fall within the category of supervised learning algorithms. Because Random Forest outperforms all the other techniques in terms of outcomes, our technique uses it to reliably forecast air pollution [3].

Selva Prabhakaran constructed an ideal ARIMA model out from foundation up to develop Seasonal ARIMA (SARIMA) and SARIMAX models. Moreover, I t also shows Python programmers how to build auto-arima models. Selva Prabhakaran, starting with forecasting basics, he looked into a variety of ideas, such as AR, MA, SARIMA, ARIMA, and ultimately the SARIMAX model [4]. The stationary stochastic ARMA/ARIMA [Autoregressive Moving (Integrated) Average] modelling technique has been improved by Ujjwal Kumar & V.K. Jain to determine everyday mean ambient air pollutants (CO, O3, NO and NO2) levels at an urban traffic location (ITO) in Delhi, India. Being one of the most polluted cities in the world, Delhi's air pollution problems might be analyzed and examined using the results of this study. Pooja Gopu, Naresh Kumar Nagwani and Rama Ranjan Panda analyse several pollutants in this work and forecast them using the Auto Regressive Integrated Moving Average (ARIMA) model. The effectiveness of this technique is investigated, and its performance is assessed, based on the available data set [6]. Jaelim Cho, Seong-Kyung Cho, JungwooSohn, Mina Suh, Yoon Jung Choi, Kyoung Hwa Ha, Changsoo Kim, and Dong Chun Shin carried out a time-series study with spline variables (relative humidity, daily mean temperature, and visit date) and parametric variables (national holiday , daily mean air pollutant concentration , and day of the week) using a generic additive model with Poisson distribution . They investigated the relationship between the risk of ER visits brought on by panic attacks and ambient air pollution [7]. The association between past air pollution exposure and present anxiety symptoms was studied by Melinda C. Power, Marianthi-Anna Kioumourtzoglou, Francine Laden, Olivia I. Okereke, Jaime E. Hart, and Marc G. Weisskopf. There was no correlation among anxiety and exposure to PM 2.5-10 in this study [8]. Researchers Tuulia Rotkoa*, Mark J. Nieuwenhuijsen, Nino Kunzli, Paolo Carrer, Lucy Oglesby, and Matti Jantunen investigated and compared the levels and characteristics affecting air pollution irritation among the adult populations of six European cities. They determine the connection between exposure levels to fine particles (PM2.5), nitrogen dioxide (NO2), as well as micro and the subjective irritation of air pollution [9]. This study by Sabrina Llop* finds modifying variables and their association with exposure to ambient nitrogen dioxide, in addition to reporting the degree of irritation brought on by air pollution and noise in pregnant women in a birth cohort (NO2). Environmental irritation may have physiological consequences that impact fetal growth or even psychological impact that reduce quality of life [10]. With a focus on highlighting the clinical consequences for researchers and healthcare professionals, Robert D. Brook carried out a study with the updated American Heart Association scientific statement as its aim to present a thorough review of the latest information linking PM vulnerability with cardiovascular disease [11]. A study by Victoria Sass found that public health initiatives to reduce the personal and societal costs of mental illness should focus not only on unique individual characteristics and factors related to the social environment, but also on under-researched aspects of the physical environment should also be taken into account for instance Air Pollution [12]. Wei Xu1 conducted a study to look into the role that stress plays in mediating the connection between people's perceptions of haze and their regular negative emotions. The interaction between stress and affective responses to haze was thoroughly explored in this study [13]. Jackson G. Lu claims that anxiety brought on by air pollution may enhance unlawful and immoral behaviour. In investigations of a 9-year panel of 9,360 U.S. cities, it was discovered that air pollution predicted six key categories of crime. These estimates, which included a variety of controls (including population, fixed effects for cities, law enforcement and years), also passed several robustness checks [14]. In a study by Carsten Bcker Pedersen, the potential link between air pollution from traffic and an increase in the prevalence of schizophrenia is investigated. This study also examines if

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 11 Issue 12, December 2022

this potential link could explain the differences in schizophrenia risk among urban and rural regions. Data on 7455 children's air pollution exposure at birth were gathered by the Danish Cancer Society in order to assess whether traffic-related air pollution has a role in the development of cancer. On such data, this study cohort was built. According to a study by Yanfen Lin, mental stress throughout pregnancy and air pollution are related in a dose-dependent manner. In order to investigate the connection between air pollution and maternal stress throughout pregnancy, we enlisted 1,931 women in Shanghai in 2010 and followed them during their mid- to late-stage pregnancies [16]. In ten significant cities in Taiwan, South Korea, and Japan, Yoonhee Kim examined the connection between air pollution and suicide. Over several days, elevated NO2, SO2, PM10, and PM102.5 concentrations were associated with a greater likelihood of suicide. The findings of this study suggest that increased air pollution levels can be associated to suicide, and further research is required to understand the mechanisms involved [17]. Mieczysław Szyszkowitz's study examined the impact of ambient air pollution on Emergency department visits for attempted suicides. The results suggest a potential link among air pollution and emergency room visits for attempted suicides. The study recommends utilising generalised linear mixed models, a relatively new statistical technique, to investigate the relationship between vulnerability to ambient circumstances and the rate of emergency department (ED) visits for attempted suicides [18]. Adil Masood's paper offers a comprehensive overview of the most widely used AI-based techniques for forecasting air pollution, comprising Deep Neural Networks (DNN), Artificial Neural Networks (ANN), Support Vector Machines (SVM), and Fuzzy Logic, through a systematic assessment of the literature. The chosen 90 papers in total were circulated between 2003 and 2021 [19]. I.B. Konovalov's research examines the advantages of integrating deterministic and statistical methods for PM10 forecasting across Europe one day in front. The recommended technique statistically postprocesses deterministic forecasts utilizing monitoring data for PM10 [20]. Guangqiang Zhou suggested a study that included statistical air quality forecasts and a comprehensive examination of data over a period of two full years (2014–2015). The evaluation's results show that the RAEMS is effective at foretelling the temporal changes and spatial distribution of important air pollutants across eastern China. The accuracy is the same for varying forecast durations of 24 hours, 48 hours, and 72 hours [21]. The Real-Time Air Quality Forecast (RT-AQF) system, which is centered on a three-dimensional air quality framework, is an efficient way to forecast air quality and advise the public on necessary preventive measures. Ming-Tung Chuang develops and instals a new RT-AQF system in the American Southeast as part of this endeavour [22]. The Air Quality Index (AQI), which may be determined using a formula begins with a thorough analysis of the level of air pollutants, is a tool that government organisations can use to describe the condition of the air quality at a particular place. Developing a forecasting model for daily AQI prediction as the basis for environmental science decision-making is the goal of the current research by Anikender Kumar [23]. In connection with the uncertainty challenge of contaminants diffusing in the environment of Wuan City, Ji-hong Zhou performed research, developed a unique theory about the air ground pollution belt, and came up with a two-dimensional bounded difference equation of air pollution. In Wuan City, implement interim air pollution control measures in compliance with the simulation's findings [24]. Regarding changes in terrestrial arthropod variety, abundance, and resilience in ecosystems affected by point pollutants, Elena L. Zvereva recommended a study to examine the broad trends and identify the origins of variation. There were discovered 134 pertinent studies that were written during 1965 and 2007. They used meta-analysis to check for general impacts and contrast between various polluter types and arthropod classifications, as well as linear regression to describe the latitudinal gradient and quantify linkages between pollution and arthropod behaviours [25]. Aoife Donnelly's model can forecast air quality in real time with outstanding computational accuracy and efficiency. Utilizing past relationships between meteorology and nitrogen dioxide (NO2) levels as well as temporal variations in NO2 levels, it is possible to estimate air quality 48 hours beforehand. The model just requires basic input data and very little computing resources. It was found to be a trustworthy and practical technique for making real-time forecasts of air quality [26]. Real-time air quality forecasting (RT-AQF) is a new area of atmospheric sciences that Yang Zhang introduced. It is one of the most important advances and potential implementation of science and engineering, presents unmatched technical, scientific, and computational dilemmas, and provides considerable prospects for scientific dissemination and community involvement. This two-part study presents a complete overview of the history, present state, major research and outreach difficulties, and future directions of RT-AQF with a concentration on the use and improvement of three-dimensional (3-D) deterministic RT-AQF models [27].

2. BACKGROUND

In this section, we go through how different facets of society are impacted by air quality:

2.1 *Medical problems*

Individuals who are vulnerable to air pollution experience numerous detrimental health effects. Impacts can be categorised into two categories: short-term effects and long-term consequences. Bronchitis and pneumonia are two examples of impermanent short-term repercussions. Also included are discomforts such rashes on the eyes, skin, nose, or throat.

The long-term impacts of air pollution could endure a person for their entire life or for a very long time. They could even be the reason someone dies. Lung cancer, heart disease and respiratory conditions like emphysema are just a few of the long-term medical repercussions of air pollution. In addition to harming a person's kidneys, liver, brain, nerves, and other organs over time, air pollution can also cause damage to them. Some researchers believe that the principal reason of birth anomalies is Air Pollution.

2.2 *Effects on plants*

There are a number of ways to show how air pollution damages plants. Foliage damage may immediately develop as necrotic lesions (dead tissue), or it may take time to evolve and manifest as chlorosis or yellowing of the leaves. It's possible that different plant portions thrive more slowly than usual. Although plants can be entirely harmed, they typically don't die until they have been damaged numerous times.

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 11 Issue 12, December 2022

2.3 *Agriculture*

Motor vehicle usage and an expanding population both contribute to photochemical air pollution, which has an impact on neighbouring rural areas as well as metropolitan centres. The intermixture of pollutants from all sectors, including agriculture, has released pollutants into the air, including hydrocarbons, aldehydes, organic acids, ozone, peroxyacetyl nitrates, pesticides, and radionuclides. These contaminants can have different effects on food, fibre, forage, and forest crops, based on their concentration, location, and weather. Of course, monetary impact also arises from crop damage by air pollution [28].

2.4 *Economy*

The consumption of more products and services by people, which results in increased air pollution, is one of the immediate repercussions of more nations expanding quickly. The demand for products and services reaches customers through international trade, which allows the flow of commodities around the world at reasonable pricing. As a result, trade contributes to a role in the expanding environmental deterioration [29].

The underlying effect of ambient air pollution from burning fossil fuels includes long-term economic impact and public health problems like lung cancer, respiratory ailments, and work loss. The opportunity cost of productivity loss owing to premature mortality, the direct expenses of illness management at all levels, along with the indirect expenses of informal care, must all be taken into consideration.

3. TIME SERIES AND ARIMA

Time series forecasting is the process of projecting future values over an extended period of time. It entails developing models based on past data and using them to make conclusions and direct future tactical judgements. The future is forecasted or evaluated based on the past. Time series add an additional time order dependence among observations. This dependency serves as both a knowledge source and a knowledge barrier. Time series can be encountered in a variety of phenomena, from economics to the physical sciences, and general methods can be applied without having to fully understand the underlying causal linkages between the independent and dependent variables (viz., a black-box approach). The observed demeanor cannot be replicated, and the insights are typically dependent upon time since only one realisation happens at a given moment in each experiment. Air quality measures are a good way to display environmental time series. The methodology that is usually applied for the estimate of environmental parameters is based on traditional descriptive statistics because of the significant variability of air quality data and the poor signal-to-noise ratio of the relevant measurements. Time-series analysis might be an useful tactic to overcome these difficulties since it makes it possible to uncover underlying deterministic behaviour and thereby contributes to the understanding of cause-and-effect relations in environmental problems.

3.1 *Stationarity*

The variance ,mean, autocorrelation, and other statistical characteristics of a stationary time series should remain the same during the course of the event. The time series can be made approximately stationary by applying quantitative adjustments, which is the basic assumption of the bulk of statistical forecasting methods. Stationarized series are relatively easy to predict. All you have to do is consider that its statistical properties stand the same as in the past. Then, by reversing any previous mathematical manipulations, the estimates for the stationarized series can be "untransformed" to produce foretelling for the original series. Because of this, figuring out the series of transformations needed to stationarize a time series frequently provides helpful clues when looking for an appropriate forecasting model. When constructing an ARIMA model, a time series must first be stationarized by differencing [39].

Another purpose to try to stationarize a time series is the capacity to obtain meaningful sample statistics, such as means, variances, and correlations with other parameters. Such statistics can only be utilized to forecast behaviour in the future if a series is stationary. The sample mean and variance, for instance, will rise with sample size and consistently undervalue the mean and variance in succeeding periods if the series is increasing continuously over time. Moreover, the series' mean and variance are not specifically articulated if the mean, variance, and correlations with other variables are not. For this reason, consider caution when extrapolating regression models fitted to nonstationary data.

3.2 *Differencing*

A technique for turning a non-stationary time series into a stationary one is differentiation. It can be used to get rid of the series' so-called temporal reliance on time. Structures like patterns and seasonality are included in this. It can be used to get rid of the series' so-called temporal reliance on time. Structures like patterns and seasonality are included in this. The difference between the present time period and the preceding time period makes up the initial differencing value. If these numbers don't rotate around a constant mean and variance, we use the results of the first differencing to calculate the second. This is imitated till the series becomes dormant. The most efficient way to judge whether or not the differenced series is sufficiently distinct is to plot it and look at its mean and variance to determine if they stay unchanged. To analyse linear trends, first-order differencing applies the transformation zi = yi-yi-1. Zi = (yi- yi-1) - (yi-1-yi-2) is a first-order difference of a first-order difference that is used in second-order differencing. It is equivalent to zi = yi-2yi-1+ yi-2 and so on [38]. Quadratic patterns are addressed with this technique.

3.3 *Unit Root Test*

The term "random walk with drift" is widely used to describe a stochastic trend in a time series called as a unit root, also termed as a unit root process or a difference stationary process. If a time series has a unit root, it will show an unexpected systematic pattern. A unit root is utilized in time series analysis to ascertain whether a time series is stationary or not. The null hypothesis asserts that time series have a unit root, while the alternative hypothesis asserts that time series are stationary [37].

The following is a mathematical expression for the unit root test:

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 11 Issue 12, December 2022

$$y_t = D_t + z_t + \varepsilon_t$$

where,

- D_t is the deterministic component (trend, seasonal component, etc.)
- z_t is the stochastic component.
- ε_t is the stationary error process.

The fundamental purpose of the unit root test is to check whether or not the zt (stochastic component) contains a unit root.

ADF Test

The Augmented Dickey-Fuller test (ADF) explores the occasion that a unit root exists in a time series sample. The alternative hypothesis may differ based on the testing version being used, but it is frequently stationarity or trend-stationarity. It serves as an enhancement to the Dickey-Fuller test for a larger and more complex set of time series models. The ADF test expands the Dickey Fuller test equation to include high order regressive processes. The Augmented Dickey-Fuller (ADF) statistic for the test yields a negative result. At least to a certain degree of certainty, the stronger the rejection of the unit root hypothesis, the further negative it is.

Let yt be a time series. We implement ADF to a model, and it has the following mathematical representation:

$$\Delta y_t = \alpha + \beta t + \gamma y_{t-1} + \delta_1 \Delta y_{t-1} + \cdots + \delta_{p-1} \Delta y_{t-p+1} + \varepsilon_t,$$

Where

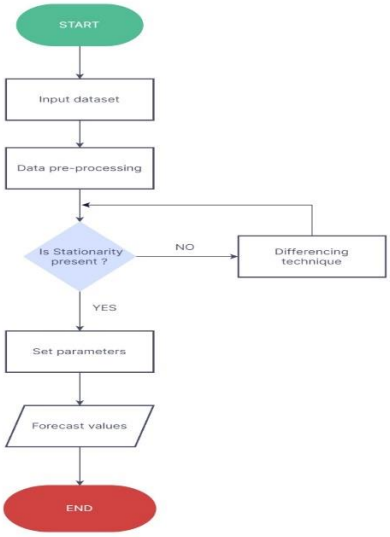
- α is a constant
- ???? is the coefficient at time.
- p is the lag order of the autoregressive process.

We have now included the differencing variables that modify how ADF and the Dickey-Fuller test compare mathematically. An important thing to keep in mind is that the test's p-value must be smaller than the significance level (let's say 0.05) in order to reject the null hypothesis since the null hypothesis implies the existence of a unit root. Resulting in the conclusion that the series is stationary.

ARIMA MODEL

A class of models known as "Auto Regressive Integrated Moving Average," or just "ARIMA," uses a time series' own historical values—more particularly, its own lags and lagged prediction errors—to "describe" the time series and then applies that equation to estimate future results. [34] ARIMA models may be used to model any "non-seasonal" time series containing patterns and more than random noise. The terms p, d, and q define an ARIMA model. The AR word is positioned in 'p' order where, the 'MA' word is in q-order. 'd' denotes how much differencing must occur for the time series to become stationary. If a time series exhibits seasonal patterns, seasonal keywords need to be included, thus the abbreviation SARIMA. When a time series does not appear to be covariance stationary, the differencing method can be employed to establish it stationary. The ARMA (p, q) model may be used to represent the stationary differenced time series; the resultant is referred to as the ARIMA (p, d, q) model, where d is the order of differencing (Shumway and Stoffer 2006; Brockwell and Davis 2002). [35] There are two distinct types of tests that may be used to assess if a model is stationary or not. They are the Rolling Statistics (RS) and the Augmented Dickey Fuller Test (ADFT). As It rolls, plot the statistical moving average or moving variance to determine if it changes with time. The model is non-stationary if the response is yes. If not, it is a stationary model. The null hypothesis of the Augmented Dickey Fuller Test is that the time series is non-stationary. The findings of the dickey Fuller test may be used to establish whether or not the model is stationary. If the estimated p-value is relatively tiny and the critical values are greater than the test statistic, the model is considered to be stationary. [36] After performing the tests for stationarity, if indeed the model is not stationary, apply differencing to make the time series stationary. Non-stationary time series are changed into stationary time series through the technique of differencing. The initial differencing quantity is the difference between the current time period and the previous time period. If the model is not yet stationary, move on to the second differentiation. This should be done repeatedly until the model becomes stationary. [36]

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 11 Issue 12, December 2022



4. MACHINE LEARNING MODELS

A thorough and in-depth examination of the literature reveals that deep learning and flat predictors are the most popular smart predictors for predicting air quality. In continual development, every predictor is optimised. Because structures and algorithms are comparable, this section categorizes and describes the updated simple intelligent models.[31].

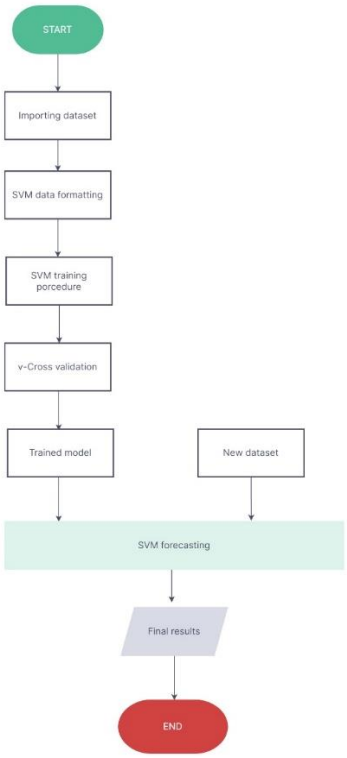
4.1 *Artificial Neural Network*

According to the needs of the application, the ANN simulates the neuronal architecture of the human brain to create a simplified model with various networks. The fundamental ANN is a massively connected prediction model with each node representing an activation function. The ANN is the most popular algorithm used in a variety of engineering fields, and it uses a large number of neurons to learn nonlinear information from the input data and infer the complex relationship between the unknown data to build models that can generalise and predict unknown data of air pollutants. Many enhanced models of the conventional ANN have been proposed and used for forecasting air quality with continuous application and updating. The backpropagation neural network (BPNN) model, which uses feed-forward multilayer networks, is a popular machine learning technique. An algorithm has been developed to enhance ANN. Artificial neurons in the three-layer BPNN structure transport the input data from the input layer to the hidden layer, allowing the information to flow forward and delivering the results to the output layer. However, as a result of the feedback, the network's errors will be transmitted backward. The BPNN model has also been used to address air quality prediction issues. The wavelet neural networks (WNN) use a wavelet basis function as the activation function in conjunction with ANN and wavelet analysis. So that it can benefit from both networks and increase its learning capacity and air pollution prediction accuracy. The Elman neural network (ENN) was proposed by Elman in]. By incorporating a context layer as a one-step delay operator into the hidden layer, ENN achieves the goal of improved memory, the ability to adapt to time-varying identity, and increased network stability. Numerous researchers improved the ENN model for better predictive performance and used it to forecast pollutant concentrations.[31]

4.2 *Support Vector Machine*

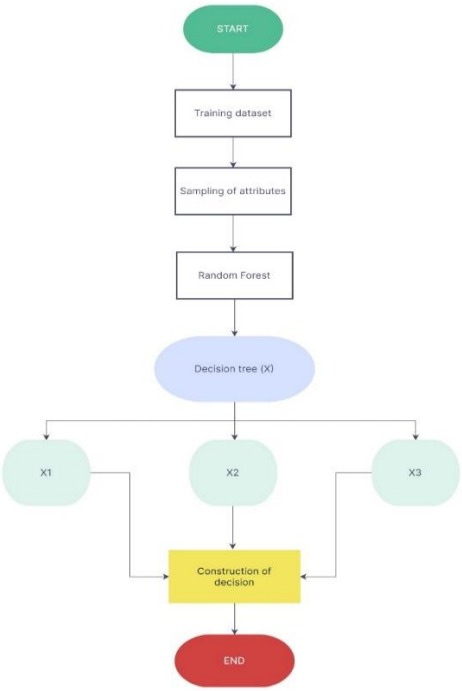
The support vector machine (SVM), a machine learning technique designed to reduce generalisation error from training error, is based on the theory of structural risk minimization (SRM). However, SVM differs significantly from ANNs in terms of its mathematical structures and processing strategies. SVM builds hyperplanes to divide up various classes. Regression analysis can take the place of classification for an output variable that is continuous. Support vector regression (SVR), a form of optimised modelling, was successfully used by Nieto et al. to calculate the air quality in the northern Spanish city of Oviedo and to find a rough solution to highly nonlinear problems. [31]

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 11 Issue 12, December 2022



4.3 *Random Forest*

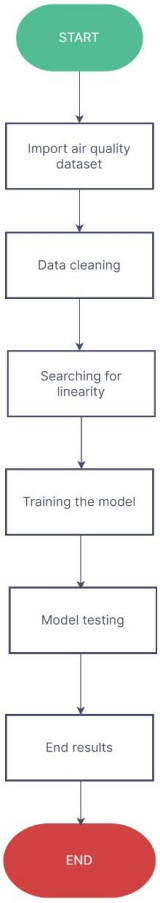
Random forest is a well-known supervised learning ensemble algorithm that combines multiple decision trees to create a forest and the bagging concept to introduce randomness into the model-building process. The individual tree is divided using a random selection of features, and each decision tree's training data subset is made using a random selection of instances. At each decision node in every tree, the variable from the random number of features is taken into account to determine the best split. The most frequent prediction will be used by random forests if the target attribute is categorical. If it is numerical, however, the average of all predictions will be used. The random forest can handle both classification and regression cases, much like SVM. Every test data point is run through the forest's decision trees in order to make a prediction. The prediction is then made based on the outcome of a vote among the trees, and as a result, a stronger and more reliable single learner is produced. Because Random forests can get around each decision tree's prediction variance such that the prediction average will nearly match the categorization or actual value (regression). [32]



International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 11 Issue 12, December 2022

4.4 *Linear Regression*

Most academicians likely began their first experiences with machine learning with linear regression. The process of fitting a line in n dimensions with a dependent variable and one or more independent variables constitutes the main working principle. N typically represents the quantity of variables in a dataset. According to the creators of this line, fitting all the instances into it would result in a maximum number of errors being minimised. The technique most frequently used for optimization is gradient descent. It works by partly deriving the loss function, and to update all parameters, the previous value is subtracted from the derivative and multiplied by a pre-set learning rate.. The simplest method, known as the rule of thumb (trial and error), can be used to adjust learning rate, or a more complex rule, such as a meta-heuristic. Lazo and ridge regression are two types of regularisations utilised in linear regression. By setting a feature's coefficient to zero, lasso regularisation will remove less significant features while keeping more significant ones. Ridge regularisation, on the other hand, tries to reduce the size of the coefficients to reduce the model's variance rather than trying to eliminate a feature.[33]



4.5 *Decision Tree*

The decision tree is one of the algorithms used in supervised learning to describe the choice made depending on the circumstance. It is used in both regression and classification. The decision tree should always be constructed top to bottom. The uppermost node is known as the "root node." The chain's last node is known as a leaf node. Internal nodes exist in the space between the root node and leaf nodes. A condition-based division of the internal nodes is followed by decision-making. Real-time algorithm complexity and tree size increase as the number of variables rises. Regression and classification trees are the two types of decision trees that exist. The dataset is classified using a classification tree so that data analysis is simple. But with this algorithm, we are unable to predict anything. The main purpose of the regression tree is to forecast continuous values. The attributes used to make predictions, the split condition, and the timing of the tree's growth are just a few examples of the factors that affect the tree's growth.[32].

5. CONCLUSION AND FUTURE SCOPE

The climatic variables such as wind direction, atmospheric wind speed, relative humidity, and temperature control the concentration of air contaminants in ambient air. The (AQI) Air Quality Index is a tool used to gauge air quality. Furthermore, proposed statistical models may be incorporated as part of the data-assimilation process of the deterministic air quality models, which are currently popular and offer useful spatial scenarios of air quality status but perform less well than statistical models when it comes to future prediction. It is anticipated that such inclusion may enhance this air quality models' capacity for prediction. Even though they are very useful for predicting the future, time series models cannot reproduce spatial scenarios; therefore, it is

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 11 Issue 12, December 2022

anticipated that their incorporation into deterministic air quality models will result in a much better spatio-temporal scenario that policymakers can use to regulate air pollution. According to this review, the majority of researches have focused on forecasting AQI and pollutants concentration levels, which will provide an accurate picture of AQI. Many researchers choose Artificial Neural Network (ANN), ARIMA Model, Linear Regression, and Logistic Regression for the prediction of AQI and air pollutants concentration. When projecting the AQI or the future concentration level of various pollutants, the future scope may take into account all elements, including meteorological parameters and air contaminants. As the data changes at specific intervals of time, we can also use real-time data analysis via the cloud to get better results for increased performance. To process enormous amounts of data and combine two or more machine learning algorithms, we can obtain more accurate results.

REFERENCES

[1] Cho, J., Choi, Y. J., Sohn, J., Suh, M., Cho, S. K., Ha, K. H., ... & Shin, D. C. (2015). Ambient ozone concentration and emergency department visits for panic attacks. *Journal of Psychiatric Research*, *62*, 130-135.
[2] Power, M. C., Kioumourtzoglou, M. A., Hart, J. E., Okereke, O. I., Laden, F., & Weisskopf, M. G. (2015). The relation between past exposure to fine particulate air pollution and prevalent anxiety: observational cohort study. *bmj*, *350*.
[3] Rotko, T., Oglesby, L., Künzli, N., Carrer, P., Nieuwenhuijsen, M. J., & Jantunen, M. (2002). Determinants of perceived air pollution annoyance and association between annoyance scores and air pollution (PM2. 5, NO2) concentrations in the European EXPOLIS study. *Atmospheric Environment*, *36*(29), 4593-4602.
[4] Llop, S., Ballester, F., Estarlich, M., Esplugues, A., Fernández-Patier, R., Ramón, R., ... & Iñiguez, C. (2008). Ambient air pollution and annoyance responses from pregnant women. *Atmospheric Environment*, *42*(13), 2982-2992.
[5] Brook, R. D., Rajagopalan, S., Pope III, C. A., Brook, J. R., Bhatnagar, A., Diez-Roux, A. V., ... & Kaufman, J. D. (2010). Particulate matter air pollution and cardiovascular disease: an update to the scientific statement from the American Heart Association. *Circulation*, *121*(21), 2331-2378.
[6] Cho, J., Choi, Y. J., Sohn, J., Suh, M., Cho, S. K., Ha, K. H., ... & Shin, D. C. (2015). Ambient ozone concentration and emergency department visits for panic attacks. *Journal of Psychiatric Research*, *62*, 130-135.
[7] Sass, V., Kravitz-Wirtz, N., Karceski, S. M., Hajat, A., Crowder, K., & Takeuchi, D. (2017). The effects of air pollution on individual psychological distress. *Health & place*, *48*, 72-79.
[8] Rotko, T., Oglesby, L., Künzli, N., Carrer, P., Nieuwenhuijsen, M. J., & Jantunen, M. (2002). Determinants of perceived air pollution annoyance and association between annoyance scores and air pollution (PM2. 5, NO2) concentrations in the European EXPOLIS study. *Atmospheric Environment*, *36*(29), 4593-4602.
[9] Xu, W., Ding, X., Zhuang, Y., Yuan, G., An, Y., Shi, Z., & Hwa Goh, P. (2020). Perceived haze, stress, and negative emotions: An ecological momentary assessment study of the affective responses to haze. *Journal of health psychology*, *25*(4), 450-458.
[10] Lu, J. G., Lee, J. J., Gino, F., & Galinsky, A. D. (2018). Polluted morality: Air pollution predicts criminal activity and unethical behavior. *Psychological science*, *29*(3), 340-355.
[11] Pedersen, C. B., Raaschou-Nielsen, O., Hertel, O., & Mortensen, P. B. (2004). Air pollution from traffic and schizophrenia risk. *Schizophrenia research*, *66*(1), 83-85.
[12] Lin, Y., Zhou, L., Xu, J., Luo, Z., Kan, H., Zhang, J., ... & Zhang, J. (2017). The impacts of air pollution on maternal stress during pregnancy. *Scientific reports*, *7*(1), 1-11.
[13] Kim, Y., Ng, C. F. S., Chung, Y., Kim, H., Honda, Y., Guo, Y. L., ... & Hashizume, M. (2018). Air pollution and suicide in 10 cities in Northeast Asia: a time-stratified case-crossover analysis. *Environmental health perspectives*, *126*(3), 037002.
[14] Szyszkowicz, M., Willey, J. B., Grafstein, E., Rowe, B. H., & Colman, I. (2010). Air pollution and emergency department visits for suicide attempts in Vancouver, Canada. *Environmental health insights*, *4*, EHI-S5662.
[15] Masood, A., & Ahmad, K. (2021). A review on emerging artificial intelligence (AI) techniques for air pollution forecasting: Fundamentals, application and performance. *Journal of Cleaner Production*, *322*, 129072.
[16] Konovalov, I. B., Beekmann, M., Meleux, F., Dutot, A., & Foret, G. (2009). Combining deterministic and statistical approaches for PM10 forecasting in Europe. *Atmospheric Environment*, *43*(40), 6425-6434.
[17] Zhou, G., Xu, J., Xie, Y., Chang, L., Gao, W., Gu, Y., & Zhou, J. (2017). Numerical air quality forecasting over eastern China: An operational application of WRF-Chem. *Atmospheric Environment*, *153*, 94-108.
[18] Chuang, M. T., Zhang, Y., & Kang, D. (2011). Application of WRF/Chem-MADRID for real-time air quality forecasting over the Southeastern United States. *Atmospheric environment*, *45*(34), 6241-6250.
[19] Kumar, A., & Goyal, P. (2011). Forecasting of daily air quality index in Delhi. *Science of the Total Environment*, *409*(24), 5517-5523.
[20] Zhou, J. H., Zhao, J. G., & Li, P. (2010, March). Study on gray numerical model of air pollution in wuan city. In *2010 International Conference on Challenges in Environmental Science and Computer Engineering* (Vol. 1, pp. 321-323). IEEE.
[21] Zvereva, E. L., & Kozlov, M. V. (2010). Responses of terrestrial arthropods to air pollution: a meta-analysis. *Environmental Science and Pollution Research*, *17*(2), 297-311.
[22] Kumar, A., & Goyal, P. (2011). Forecasting of air quality in Delhi using principal component regression technique. *Atmospheric Pollution Research*, *2*(4), 436-444.
[23] Donnelly, A., Misstear, B., & Broderick, B. (2015). Real time air quality forecasting using integrated parametric and non-parametric regression techniques. *Atmospheric Environment*, *103*, 53-65.
[24] Zhang, Y., Bocquet, M., Mallet, V., Seigneur, C., & Baklanov, A. (2012). Real-time air quality forecasting, part I: History, techniques, and current status. *Atmospheric Environment*, *60*, 632-655.
[25] Cabaneros, S. M., Calautit, J. K., & Hughes, B. R. (2019). A review of artificial neural network models for ambient air pollution prediction. *Environmental Modelling & Software*, *119*, 285-304.
[26] Kamal, M. M., Jailani, R., & Shauri, R. L. A. (2006, June). Prediction of ambient air quality based on neural network technique. In *2006 4th Student Conference on Research and Development* (pp. 115-119). IEEE.
[27] Antanasijević, D. Z., Pocajt, V. V., Povrenović, D. S., Ristić, M. Đ., & Perić-Grujić, A. A. (2013). PM10 emission forecasting using artificial neural networks and genetic algorithm input variable optimization. *Science of the Total Environment*, *443*, 511-519.
[28] Aneja, V. P., Schlesinger, W. H., & Erisman, J. W. (2009). Effects of agriculture upon the air quality and climate: research, policy, and regulations.
[29] Alvarado, R., Ortiz, C., Jiménez, N., Ochoa-Jiménez, D., & Tillaguango, B. (2021). Ecological footprint, air quality and research and development: the role of agriculture and international trade. *Journal of Cleaner Production*, *288*, 125589.
[30] Taghizadeh-Hesary, F., & Taghizadeh-Hesary, F. (2020). The impacts of air pollution on health and economy in Southeast Asia. *Energies*, *13*(7), 1812.
[31] Liu, H., Yan, G., Duan, Z., & Chen, C. (2021). Intelligent modeling strategies for forecasting air quality time series: A review. *Applied Soft Computing*, *102*, 106957.
[32] Liang, Y. C., Maimury, Y., Chen, A. H. L., & Juarez, J. R. C. (2020). Machine learning-based prediction of air quality. *Applied Sciences*, *10*(24), 9151.

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 11 Issue 12, December 2022

[33] Madhuri, V. M., Gunjal, G. S., & Kamalapurkar, S. (2020). Air pollution prediction using machine learning supervised learning approach. *International Journal of Scientific and Technology Research*, *9*(4), 118-123.
[34] Prabhakaran, S. (2022, September 3). *Arima model - complete guide to time series forecasting in python: ML+*. Machine Learning Plus. Retrieved October 31, 2022, from https://www.machinelearningplus.com/time-series/arima-model-time-series-forecasting-python/
[35] Kumar, U., & Jain, V. K. (2010). ARIMA forecasting of ambient air pollutants (O3, NO, NO2 and CO). *Stochastic Environmental Research and Risk Assessment*, *24*(5), 751-760.
[36] Gopu, P., Panda, R. R., & Nagwani, N. K. (2021). Time series analysis using ARIMA model for air pollution prediction in Hyderabad city of India. In *Soft Computing and Signal Processing* (pp. 47-56). Springer, Singapore.
[37] Stephanie. (2021, January 01). Unit root: Simple definition, unit root tests from https://www.statisticshowto.com/unit-root/ADF Test
[38] Differencing (of time series). (n.d.) from https://www.statistics.com/glossary/differencing-of-time-series/
[39] Stationarity in time series analysis - towards data science , from https://towardsdatascience.com/stationarity-in-time-series-analysis-90c94f27322
[40] Prabhakaran, S. (2022, April 04). Augmented dickey-fuller (ADF) test - must read guide - ml+ from https://www.machinelearningplus.com/time-series/augmented-dickey-fuller-test/