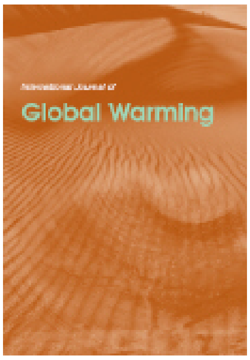


International Journal of Global Warming

ISSN online: 1758-2091 - ISSN print: 1758-2083
https://www.inderscience.com/ijgw

Wavelets and hybrid optimised SVM with random forest-based pollution forecasting

Zaheer Abbas, Princess Raina

DOI: 10.1504/IJGW.2024.10060433

Article History:

Received:	13 April 2023
Last revised:	06 June 2023
Accepted:	07 June 2023
Published online:	06 December 2023

Wavelets and hybrid optimised SVM with random forest-based pollution forecasting

Zaheer Abbas* and Princess Raina

Department of Mathematical Sciences,
Baba Ghulam Shah Badshah University,
Rajouri, India
Email: zaheerabbas142022@gmail.com
Email: princessraina1452@gmail.com
*Corresponding author

Abstract: Due to its detrimental effects on human health, information about meteorological pollutants including CO, NO_2, SO_2, and dust is becoming more and more crucial. In all nations' urban areas, this is particularly true. The instantaneous registration provided by the automatic measurements of these pollutants' concentrations serves as the foundation for the calculation of averaged values. The key issue is early pollution forecasting in order to warn or notify the local population of the impending risk. In this work, machine learning and wavelet decomposition are used to forecast daily air pollution. This research provides the forecasting strategy, utilising the hybrid random forrest with optimised support vector machines (HRFOpSV), depend on the collected data of NO_2, CO, SO_2, and dust, for the past years, and real weather conditions like humidity, wind, temperature and pressure.

Keywords: pollution; machine learning; forecasting; decomposition; human health.

Reference to this paper should be made as follows: Abbas, Z. and Raina, P. (2024) 'Wavelets and hybrid optimised SVM with random forest-based pollution forecasting', *Int. J. Global Warming*, Vol. 32, No. 1, pp.81–101.

Biographical notes: Zaheer Abbas is a faculty member of the Department of Mathematical Sciences, Baba Ghulam Shah Badshah University Rajouri, J and K, India. He has a teaching experience of 17 years and research experience of 15 years. His main areas of interest are approximation theory in pure mathematics and data mining using wavelets in applied mathematics. He has more than 30 research publications to his name and is actively involved in teaching and research.

Princess Raina is a research scholar of the Department of Mathematical Sciences, Baba Ghulam Shah Badshah University Rajouri, J and K, India. She is pursuing Doctoral degree in Applied Mathematics and is studying pollution data using wavelets and other latest tools.

1 Introduction

One of the major problems that limit the sustainable development is the air pollution. Air pollution is a gradual process which will cause disastrous effects when it is not controlled in efficient manner that is caused by human activity and other natural resources (Liu et al., 2021a). Currently, most of the cities suffer from smog which affects people's health and daily life. Smog component is the particulate matter (PM) 2.5 (Tao et al., 2019). As per WHO report, due to exposure of air pollution every year about 4.2 million people die too early. PM2.5 is the greatest risk factor among all pollutants. These particles will reside in the air for a long time as they are small and light. Exposure to air pollution will raise your risk of getting sick or dying. Hence, monitoring and reducing air pollution have become crucial challenges (Liu et al., 2020). To forecast atmospheric pollutants, neural network can be used at present (Li and He, 2020).

The most of the environmental problems caused by air are the volcanic eruptions, increased nitrogen dioxide from vehicle in addition with dust, storms and bush fires. Nitrogen dioxide leads to acid rain which dissolves in water in the atmosphere that react with UV radiation to form photochemical smog (Liu et al., 2021b). For all living things on Earth, air is the main element. As a result of power plants, urbanisation, automobiles, chemical activities, industrialisation and certain other normal happenings pollution has continued to rise over the past 50 years. Out of 85 risk factors PM is the 4th main reason according to global burden of disease (GBD) update that cause over 5 million deaths in 2017 (Harishkumar et al., 2020). Researchers have conducted pollution analysis and predicted pollutant concentrations since human activities have significantly worsened the level of air pollution (Chen et al., 2019). Fine PM (PM2.5), sulphur dioxide (SO_2), respirable PM (PM10), carbon monoxide (CO), nitrogen dioxide (NO_2), and ozone (O_3) are the six major pollutants that contribute to the air quality index (AQI). Air pollution problems can be controlled by the forecasting techniques of AQI (Jin et al., 2021). Respiratory problems are often caused by exposure to air pollution, which can include particles, gases, and chemicals that are released into the air. These pollutants can irritate the lungs and airways, causing symptoms such as coughing, wheezing, and difficulty breathing. Long-term exposure to air pollution can also increase the risk of developing chronic respiratory diseases like asthma and COPD. Air pollution can also contribute to the greenhouse effect, which is the process by which heat is trapped in the Earth's atmosphere. This occurs when certain gases, such as carbon dioxide and methane, are released into the air and then trap heat from the sun, causing the planet's temperature to rise. The greenhouse effect is a major cause of global warming, which is the long-term trend of rising temperatures across the planet. Global warming can have a wide range of negative impacts on the environment and human society, including sea level rise, more frequent and severe weather events, and the loss of plant and animal species.

Through smart sensors and the developing systems which can track and predict the air quality constantly we can assure clean air in cities. The majority of air pollution simulation techniques use simulations in mathematics. For studying air pollution, Statistical prediction methods rely on machine learning (ML) becomes useful tool with the development of artificial intelligence (Jiang et al., 2020). During peak pollution episodes to help the exposed people from the pollutants forecasting tools are employed as early-warning mechanism. Air pollution modelling divided into deterministic and data-driven (Cabaneros et al., 2020). Monitoring of air pollution is continually expanding, and its effects on human health are receiving more attention. For forecasting the potential harm of main pollutants nitrogen dioxide (NO_2) and sulphur dioxide (SO_2) multiple models have been created. But making accurate predictions is nearly impossible (Heydari et al., 2022).

AQI predictions using signal decomposition method for air pollution concentration, the performance of the prediction is enhanced. In order to train forecasting models and increase prediction accuracy, non-stationary issues resulting from the AQI traits or the non-stationary nature of air pollution were resolved using hybrid models. Its distinguishing feature is the further breakdown of the transforming a nonlinear initial time series into a more stable and regular subseries. The results of the final prediction are derived by adding the expected values of all sub-sequences (Wu and Lin, 2019). The commonly used wavelet decomposition (WD) method is based on the basic principle of breaking up a non-smooth discontinuous time series into various sequences with different low-frequency approximate component and high-frequency detail components. The number of layers in the WD process determines the quantity of high-frequency detail components. The crucial issue is the early detection of detrimental pollution only to alert or warn the local population of the impending threat (Fan et al., 2021). To decompose the original time series into different scales, SWT is utilised which represents the wavelet coefficients. A support vector regression (SVR) mode is trained using wavelet coefficients (Li and Tao, 2018). For the prediction of PM 10 concentration a modern processing technique wavelet transform (WT) is extensively used (Qiao et al., 2021). The following are the paper's main contributions:

- In this article, the monitored air pollution is sent to time series data to extract the parameters like humidity, temperature, pressure and wind speed.
- Then these parameters are given into WD method to decompose the data using SWT into the wavelet coefficients
- Then the hybrid optimised SVM is utilised to come up with the predicted values from the parameters.
- The random forest (RF) technique is used for error evaluation in order to achieve a forecast with an acceptable level of accuracy.

Further portions of this paper explore the specifics of this work; Section 2 discusses the findings, and the following section provides an examination of related earlier research on the suggested model. Section 3 demonstrates the suggested technique for the work, which has tested and examined. Section 4 covers the result analysis and comparative discussion of the study. Section 5 describes the outcome of the work.

1.1 Significance of the study

- Health consequences: Air pollution poses a major threat to human health. Prolonged exposure to high pollution levels can increase the risk of heart disease, stroke, and respiratory difficulties. By predicting air pollution levels, authorities may take action to lessen its negative effects on public health, such as sending warnings to locals to stay indoors during times of high pollution.
- Environmental protection: Air pollution may harm wildlife, forests, and agriculture, among other aspects of the environment. Authorities can reduce air pollution's negative effects on the environment by restricting emissions from enterprises and power plants by foreseeing its levels.

- Economic benefits: Air pollution can also have an economic impact, including increased healthcare costs and reduced productivity due to illness. By forecasting air pollution levels, authorities can take steps to reduce its economic impact, such as implementing policies to reduce emissions from vehicles and other sources.
- Planning and decision-making: Air pollution forecasting provides valuable information for planners and decision-makers, enabling them to make informed decisions about land use, transportation, and other infrastructure projects that may impact air quality.

2 Related work

> "Some of the recent research works related to pollution forecasting model were reviewed in this section."

Mishra (2018) has proposed a prediction model, which is really a subgroup of predictive modelling, and a ML model to forecast the degree of air pollution. To determine the best forecasting and predicting models for computing the AQI of four distinct gases, three ML methods in particular were used: O_3, NO_2, CO and SO_2.

Kumar (2018) has proposed a forecasting air pollution via differential evolution and RF. The prediction of concentration values for various contaminants is carried out in this work. Predicting precise numbers and giving accurate forecasting information is the key goal. The multi-label classifier and Bayesian network stand-alone classifiers cannot compete with RF and a mixed differential evolution approach.

Liu et al. (2019) have proposed regression models for forecasting the air quality index using SVR and random forest regression (RFR). This work also demonstrated how effectively and conveniently merging ML for predicting air quality may be used to address several associated environmental issues. Through the development of two prediction models based on a variety of scenarios, this work enhanced the predictability of air indicators and provided help for modelling and analysis of urban air quality.

Li et al. (2019) have proposed a hybrid quantum-behaved particle swarm optimisation – support vector regression (QPSO-SVR) model to forecast atmospheric PM2.5 and NO_2 concentrations. The model performs better in prediction accuracy and computational time. As it has smaller impact on meteorological factors, the QPSO-SVR is more reliable. This technique could be employed for the forecasting of various pollutant concentrations.

Dun et al. (2020) have proposed a hybrid model combining the SVR and the fractional order accumulation model (FGM (0, m)) of the grey multivariable regression model is used. The absolute percentage errors (APEs) are applied to calculate the weights for the SVR and FGM (0, m). For the same place and time period, the air quality contaminants are predicted using the Holt-Winters model.

Leong et al. (2020) have developed a SVM to simulate the air pollution index. Kernel functions model parameters alone were investigated in this model. Coefficient of determination (R^2), mean of sum of squares error (MSSE) and Sum of squares error (SSE) are used to examine this model outcomes. To improve the model's performance, the missing data and outlier were handled by using data screening technique.

Wang et al. (2020b) have presented bivariate empirical mode decomposition (BEMD) technique that is employed to construct an improved interval PM2.5 concentration prediction model. Interval grey incidence analysis (IGIA) is applied to choose input parameters for the model with the aim of obtaining the major affecting elements. From this study, it is suggested that this model is highly applicable and successful in forecasting interval PM2.5 concentration.

Wen et al. (2019) have developed a spatiotemporal long short-term memory neural network extended (C-LSTME) model for forecasting concentrations of air quality. Convolutional neural network (CNN) and long short-term memory neural network (LSTM-NN) were combined aerosol and meteorological data were also merged for high-level spatiotemporal feature extraction and for improving the model prediction performance.

Du et al. (2019) have presented a deep air quality forecasting framework (DAQFF), an integrated model for the problem of air quality prediction. One-dimensional CNNs and bi-directional LSTM are the two deep neural networks that make up DAQFF. It may be trained to recognise patterns of spatial-temporal relationships and local trends in time series data linked to multivariate air quality.

Wang et al. (2020a) have developed a new hybrid approach that employs heuristic intelligent optimisation algorithms and outlier identification and correction algorithms. First, to detect and correct outliers' data pre-processing algorithms were applied. Second, in order to determine each subseries forecasted outcomes heuristic intelligent optimisation algorithm was utilised. Ultimately, experimental outcomes and analysis demonstrate that the offered hybrid model offers precise prediction.

2.1 Problem statement

In air pollution forecasting system, the existing methods have certain drawbacks that are stated below:

- LSTM neural network which uses the concept of gates are limited due to hardware computational limitations. It impacts the experiment due to the missing data. Many hyperparameters have to be optimised for accurate predictions. When the sample is too big, the model will experience the issue of huge computation and poor real-time performance.
- Heuristic intelligent optimisation algorithm considers only AQI time series and does not take into account any other contributing elements. Additionally, it does not consider the multi-objective alternatives and takes into account simply single-objective optimisation methods for forecasting AQI time series. Hence, possible factors are to be added to predict AQI time series in the multi objective versions.
- Autoregressive integrated moving average (ARIMA) time series data is smoothed using lagged moving averages. The deterministic air quality models perform less than the mathematical models for future prediction. Such incorporation may be enhanced to improve the predictions of air quality models.
- To forecast the accurate predictions, dynamics of air pollution represented by a variety of parameters including wind direction, humidity, temperature, rainfall, snowfall, wind speed and so on which will change the air pollution concentration must be studied to know the characteristics of pollutant.

- The problem in monitoring the quality of the environment is the early warning of hazardous pollution in a region. At present, the pollutant characteristics and their scientific evaluation are hardly concentrated on early warning systems.

To overcome these drawbacks in the existing techniques, this paper proposed a ML technique along with the metaheuristic optimisation algorithm is introduced for air pollution forecasting. Data on the air pollutants is provided for wavelet processing by SWT. The hybrid RF transfers the data from the wavelet transformation to the optimised SVM, and the HRFOpSV algorithm is used to forecast the wavelet coefficient. The optimal wavelet coefficient parameter of the SVM is chosen using GWO, and using OpSVM, the analysed data is sent back to the RF, providing accurate forecasting data.

3 Proposed methodology

The objective of air pollution prediction and forecasting is to provide timely and accurate information about the levels of air pollutants in the atmosphere, as well as their potential impacts on human health and the environment. By predicting and forecasting air pollution levels, authorities and individuals can take appropriate measures to mitigate the harmful effects of air pollution, such as reducing outdoor activities or wearing protective masks. This information can also help to inform policy decisions, such as implementing stricter emission controls or increasing public transportation options to reduce vehicle emissions.

Additionally, air pollution prediction and forecasting can provide valuable information for industries and businesses that may be affected by changes in air quality, such as agriculture or outdoor recreation. By understanding how air pollution levels may fluctuate over time, these industries can make informed decisions about their operations and resource management.

The detailed operation of this paper is explained below. This proposed methodology consists of two stages like decomposition and forecasting. The steps to be followed is discussed as,

Step 1 The data preparation is performed using the time series data like NO_2, SO_2, CO, pressure, temperature, humidity and wind speed.

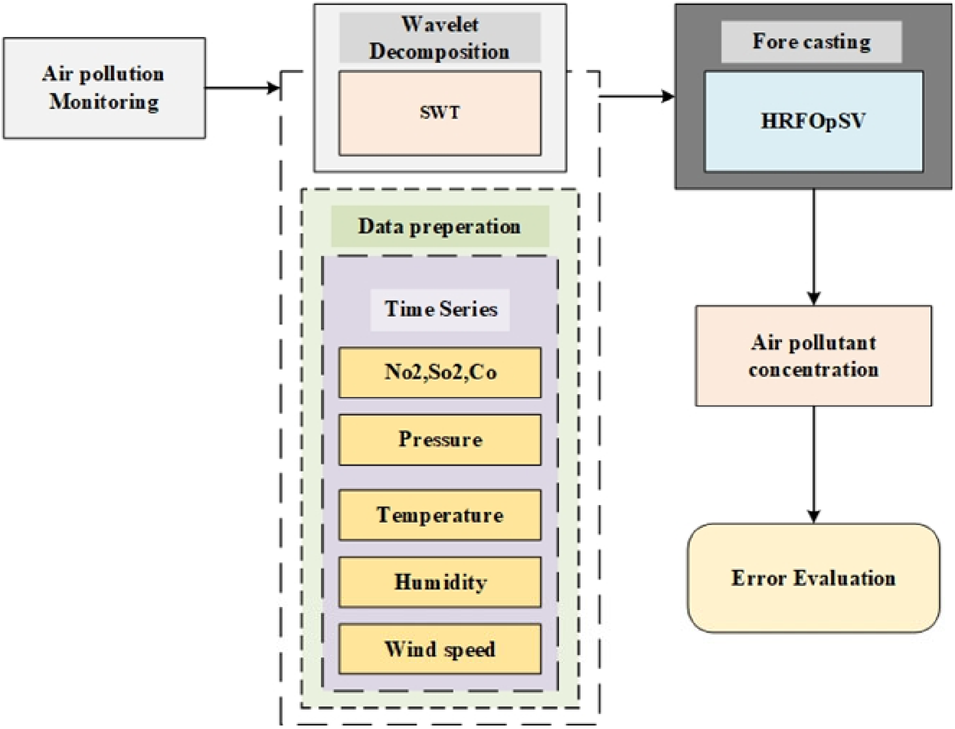
Step 2 In the second stage, the data decomposition is done with the help of WD, the SWT is used in this stage.

Step 3 The forecasting the next stage, the HRFOpSV is used which includes the SVM, RF and GWO.

Figure 1 illustrates the structural design of the proposed system, detailed and explained in this section.

First of all, the air pollutant data is given for wavelet transformation, is done by SWT. During wavelet transformation, the data preparation takes place, the data’s taken at the respective times series is divided into four blocks such as metaheuristic chemical, pressure, temperature, humidity and wind speed is calculated and the wavelet coefficient is obtained by using three level WD. The wavelet coefficient is forecasted by HRFOpSV technique, the hybrid RF sends the data from the wavelet transformation to the optimised SVM. As for the optimisation GWO is utilised, is used to select the best wavelet coefficient parameter of the SVM and through OpSVM the analysed data is send back to RF, it gives the accurate forecasting data.

Figure 1 Structural design of proposed model (see online version for colours)

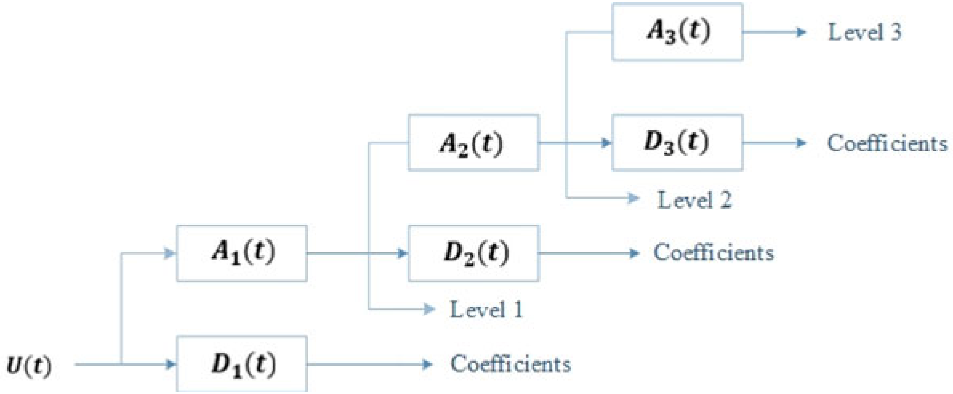


3.1 WD using SWT

Time series data splits into approximation and detail components using WD, enabling for the implementation of specific forecasting models to each component. This feature can improve forecasting performance. It can be used to monitor signals in the time phase, cut down on noise, and keep the essential aspects of the original signals. SWT is used as a wavelet transform in this article.

The discrete wavelet transform's shift-invariance issue is solved by SWT. The source signal is subdivided into levels by the wavelet transform; the resulting sub signals possess wavelengths that are equivalent of the wavelet level before the approximation signal. The down sampling operator is dropped from the standard DWT design by the SWT. It makes it possible to create a four vector containing the signals at each time step have four different wavelet components. The resulting sub signals in the SWT are the same length as the source signal. Even though the output signal is not decimated, the detail coefficients and approximation are sets of high and low-frequency coefficients that SWT uses to partition the data series. While the detail coefficients represent the minute fluctuations in the series, the approximation components show the overall time series trend. A dyadic tree can be used to illustrate the breakdown. An implementation of a three-level decomposition depend on the SWT is shown in Figure 2. The SWT divides the provided signal *u*(*t*) into two coefficients: the detail coefficients *D*(*t*) and the approximation coefficients *A*(*t*). The convolution outcomes of the low- and high-pass filters are represented by these coefficients. Using approximation coefficients as input, this decomposition process is repeated at each succeeding decomposition level.

Figure 2 Functional structure of SWT (see online version for colours)



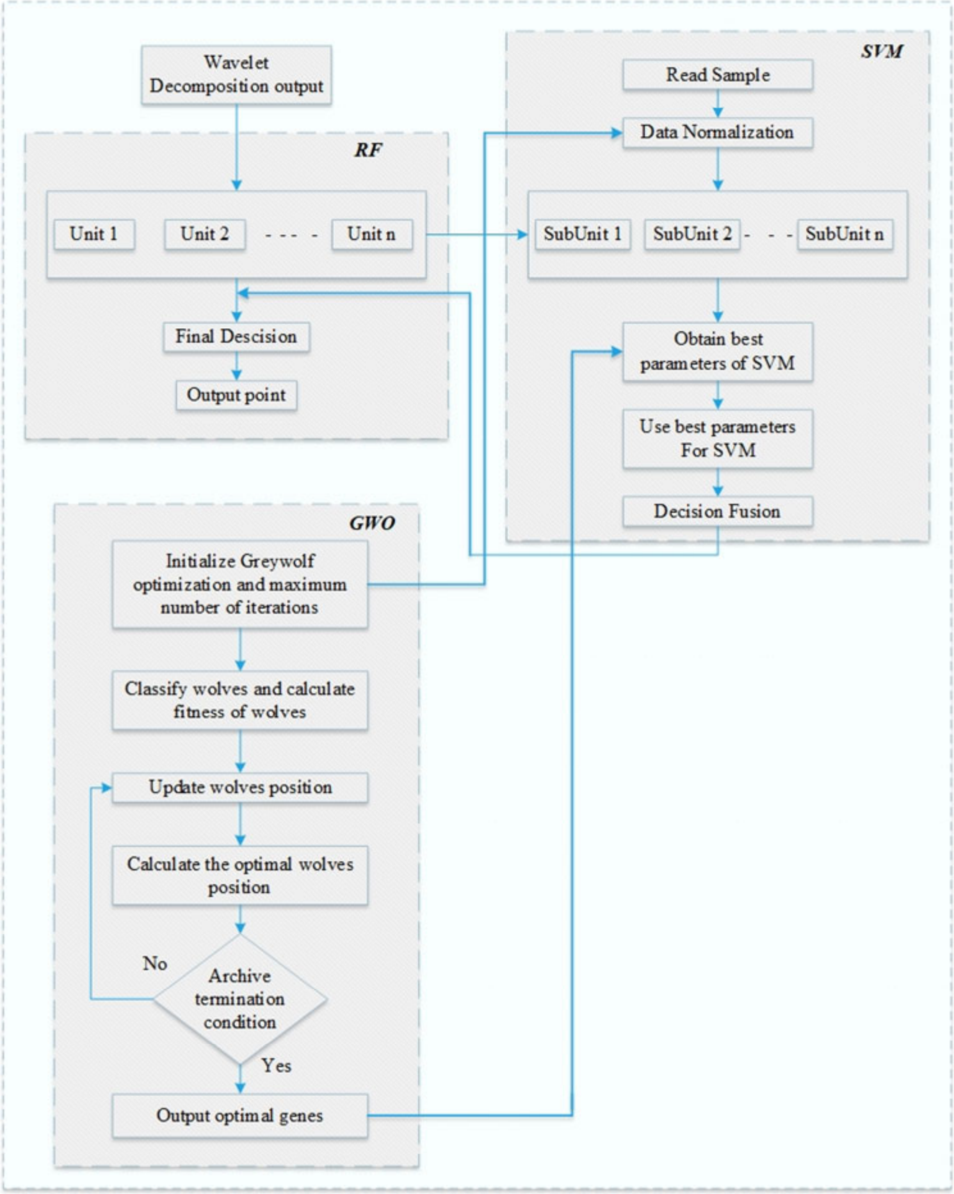
3.2 *Forecasting by HRFOpSV*

In order to tackle the classification problem, the suggested system uses a hybrid strategy and has RF capabilities. The classification system consists of several components, and the sum of their predictions determines the final classification. Each unit is made up of smaller units whose outcomes are determined by the feature extraction unit, kernel, and bootstrap sample. A decision fusion technique is utilised to combine the units' output. In this method, OpSVM subunits are changed while the RF's architecture is retained. Structural flow of HRFOpSV are shown in the Figure 3.

- *Classification unit*: The nature of classification units is similar, and each of these units is made up of a number of smaller units. One decision fusion subunit and *M* subunits make up a complete unit (DFS).
- *Classification subunit*: Distinct classes of test sample are predicted by different categorisation subunits. Typically, during the creation of an algorithm, the number of subunits is a parameter that can be selected. Eight subunits are used in the current investigation.

On to the next step shows the optimisation with SWN and as for optimisation, GWO is used. By automatically determining the ideal feature subset and the appropriate SVM parameter values for the SVM model, one can increase the SVM classifier's accuracy. Three stages of decomposition are used to divide air pollution data into five sub-band signals. The useful features like wind, humidity, temperature, metaheuristic properties are derived from wavelet coefficients. Relevant characteristics are chosen from the retrieved features, and GWO dynamically optimises the SVM parameter values. The output data's is given to the RF and the final decision is obtained and gives out accurate output samples.

Figure 3 Structural flow of HRFOpSV (see online version for colours)



Wavelet Decomposition output
RF
Unit 1
Unit 2
Unit n
Final Descision
Output point
SVM
Read Sample
Data Normalization
SubUnit 1
SubUnit 2
SubUnit n
Obtain best parameters of SVM
Use best parameters For SVM
Decision Fusion
GWO
Initialize Greywolf optimization and maximum number of iterations
Classify wolves and calculate fitness of wolves
Update wolves position
Calculate the optimal wolves position
Archive termination condition
No
Yes
Output optimal genes

The three steps of the hunting process include surrounding, chasing, and attacking in the GWO algorithm, a mathematical modelling technique that replicates the Grey-Wolf population's predatory strategy. Finding the best solution is a step in the process of capturing prey. Alpha (α), beta (β), delta (δ), and omega (ω) are the four social degrees assigned to grey wolves in accordance with the social hierarchy. α serves as the pack's managers and leaders during the hunting process. They are also in charge of all pack decision-making., the following fittest groups are β and δ, the remaining wolves are classified as ω those that aid α, β, and δ in pack management and attack the prey and assist α in pack management. Grey-Wolf group X is made up of N different grey wolves, assuming that the Grey-Wolf's solution space has V dimensions that is, $X = X_1, X_2, \ldots, X_N$. In V-dimensional space, the location of a single Grey-Wolf is X_j ¼ $X1j, X2j, \cdots, XVj$. Following the round t iteration, the distance between the individuals shows where each one has moved in the population.

Attack

For the best outcome, the wolves' assault and seize the prey. By lowering the value, the process is made to happen. The wolves assault their prey when $|Aj|$ 1, which is the best solution. It will keep looking for the best solution until $|Aj| > 1$, at which point it will move on to new possibilities. By adjusting the iterative population position optimisation and the best kernel function parameters as well as penalty parameters for the SVM method were found by modifying the independent variable dimension of the Grey-Wolf algorithm. The Grey-Wolf's location and the location of its prey are determined by fitness; the closer the location, the better the fitness. The following describes the Grey-Wolf optimisation procedure.

Surrounding

In the GWO, The first wolf in the Grey-Wolf pack is specified as being at position X_j in the D-dimensional search space, which is used to construct a mathematical model of surround behaviour. The resulting equations are implied:

$$D_j = \left|c_j.X_1(t) - X_j\right| \tag{1}$$

$$X_j(t+1) = X_1(t) - A_j.D_j \tag{2}$$

t → number of iterations, $X_1(t)$ → prey position after the t iteration, X_j → location of the Grey-Wolf, and D_j → position between the prey and Grey-Wolf. A_j and c_j → coefficients, develops the following formulae,

$$A_j = 2a * R_1 - a \tag{3}$$

$$C_j = 2 * R_2 \tag{4}$$

$R1$ and R_2 → [0, 1]. The parameter a value → [0, 2] illustrates an inverse linear association with the amount of iterations t:

$$a = 2 - \frac{t}{t_{\max}} \tag{5}$$

$t_{\max}$ → maximum number of iterations.

Chasing

The position of the prey is determined by the positions of α, β, and δ during the optimisation process of the Grey-Wolf algorithm. Under the direction of wolves α, β, and δ, wolf ω pursued the prey, updated their positions in accordance with the location of the most effective search unit, and repositioned the prey in accordance with the wolf α, β, and δ locations. The impact of α, β, and δ on the individual locations of grey wolves in the population is taken into account when the distribution in the *t* generation is determined and X_1, X_2, X_3 is as follows:

$$\begin{cases} D_\alpha = \left|C_1 X_\alpha(t) - X(t)\right| \\ D_\beta = \left|C_1 X_\beta(t) - X(t)\right| \\ D_\delta = \left|C_1 X_\delta(t) - X(t)\right| \end{cases} \tag{6}$$

$$\begin{cases} X_1 = X_\alpha(t) = A_1.D_\alpha \\ X_1 = X_\beta(t) - A_2.D_\beta \\ X_1 = X_\delta(t) - A_3.D_\delta \end{cases} \tag{7}$$

$$X(t+1) = \frac{\left(X_1 + X_2 + X_3\right)}{3} \tag{8}$$

$X_\alpha(t)$, $X_\beta(t)$, $X_\delta(t)$ are the locations of α, β, and δ. C_1, C_2, C_3 and A_1, A_2, A_3 are various oscillation factors and convergence factors correspondingly, as the population iterates to the t generation. After round *t*, the updated position of the individual in the population is denoted by $X(t + 1)$.

4 Result and discussion

In this section, the result of this article is discussed. The predicted values of gases CO, NO_2 and SO_2 is estimated by SWT model. The actual and predicted values of each gas at time period of 350 days is calculated and detailed in graph. The forecasting of air pollution is calculated through the matrices such as MAE, mean absolute percentage error (MAPE), root mean squared error (RMSE), NMSE, IOA and SAE and the error is calculated. The overall detailed of this section is given below.

- MAE

  Mean absolute error, the easiest way to gauge forecast accuracy (MAE). The variance between the expected and the actual value, expressed as an absolute number, is the absolute error.

$$MAE = 100 * \left(ABS(Actual - Forecast) / Actual\right) \tag{9}$$

- MAPE

  The MAPE is one of the most commonly used KPIs to measure forecast accuracy. MAPE is the sum of the individual absolute errors divided by the demand (each period separately). It is the average of the percentage errors.

$$MAPE = \frac{1}{N}\sum_{1}^{N}\left|\frac{Actual - Forecast}{Actual}\right| \tag{10}$$

- RMSE

  The square root of the mean of the square of all the errors is known as the RMSE. RMSE is frequently employed and is regarded as a superior all-purpose error metric for numerical forecasts.

$$RMSE = \left[\frac{1}{N} \sum_{1}^{N} (x_i - o_i)^2 \right]^{\frac{1}{2}} \tag{11}$$

- NMSE

  The normalised mean square error statistic, which draws attention to the scatter in the full data set (NMSE). The product's normalisation C_p*C_o guarantees that the NMSE will not favour models that overpredict or underpredict. Better model performance is represented by lower NMSE values. The following gives the expression for the NMSE:

$$NMSE = \frac{\overline{(C_o - C_p)^2}}{\overline{C_o} * \overline{C_p}} \tag{12}$$

- IOA

  IOA is a refined index that offers a reliable and adaptable way to quantify forecast accuracy.

$$IOA = \begin{cases} 1 - \dfrac{\sum_{1}^{N} |X_i - o_i|}{2\sum_{1}^{N} |o_i - \overline{o}|} \text{ when } \sum_{1}^{N} |X_i - o_i| \leq 2\sum_{1}^{N} |o_i - \overline{o}| \\ 1 - \dfrac{\sum_{1}^{N} |o_i - \overline{o}|}{2\sum_{1}^{N} |X_i - o_i|} \text{ when } \sum_{1}^{N} |X_i - o_i| \leq 2\sum_{1}^{N} |o_i - \overline{o}| \end{cases}, -1 \leq IOA \leq 1 \tag{13}$$

- SAE

  By utilising standard deviation, a statistical term known as the standard error evaluates how well a sample distribution represents a population. In statistics, the standard error of the mean is the variation between a sample mean and the actual mean of the population.

$$yt = {}^\wedge yt \mid t - 1 + et.yt = y \wedge t \mid t - 1 + et$$

4.1 Error evaluation of air pollutants

The gases such as SO_2, NO_2 and CO gases are evaluated through the proposed model. The error of each gases defined, is detailed in figure. The gases calculated through 350 days is analysed and the error evaluation rate of SO_2, NO_2, CO is graphically explained through some matrices. Table 1 shows the respective error values calculated by some matrices such as MAE, RMSE, MAPE, NMSE, IOA and SAE.

Table 1 Error evaluation of pollutants using the matrices

Matrices	*CO*	*NO_2*	SO_2
MAE	0.12523	0.12528	0.12578
MAPE	251.8905	251.5097	252.9801
RMSE	0.14456	0.14447	0.14496
NMSE	1.66E-08	7.13E-06	7.07E-05
IOA	0.99995	0.99995	0.99995
SAE	7.0388	0.23398	0.14154

The table displays various statistical measures of three different pollutants – CO, NO_2, and SO_2, with their corresponding values in matrices.

- The first measure, MAE provides an average of the absolute differences between the predicted and actual values. The values of MAE for CO, NO_2, and SO_2 are 0.12523, 0.12528, and 0.12578, respectively.
- The second measure, MAPE expresses the error as a percentage of the actual value. The values of MAPE for CO, NO_2, and SO_2 are 251.8905, 251.5097, and 252.9801, respectively.
- The third measure, RMSE is the square root of the average of the squared differences between the predicted and actual values. The values of RMSE for CO, NO2, and SO_2 are 0.14456, 0.14447, and 0.14496, respectively.
- The fourth measure, NMSE is the RMSE divided by the variance of the actual values. The values of NMSE for CO, NO2, and SO_2 are 1.66E-08, 7.13E-06, and 7.07E-05, respectively.
- The fifth measure, IOA is a statistical index that assesses the agreement between predicted and actual values. The values of IOA for CO, NO_2, and SO_2 are 0.99995, 0.99995, and 0.99995, respectively.
- The last measure, SAE sums up the absolute differences between the predicted and actual values. The values of SAE for CO, NO_2, and SO_2 are 7.0388, 0.23398, and 0.14154, respectively.

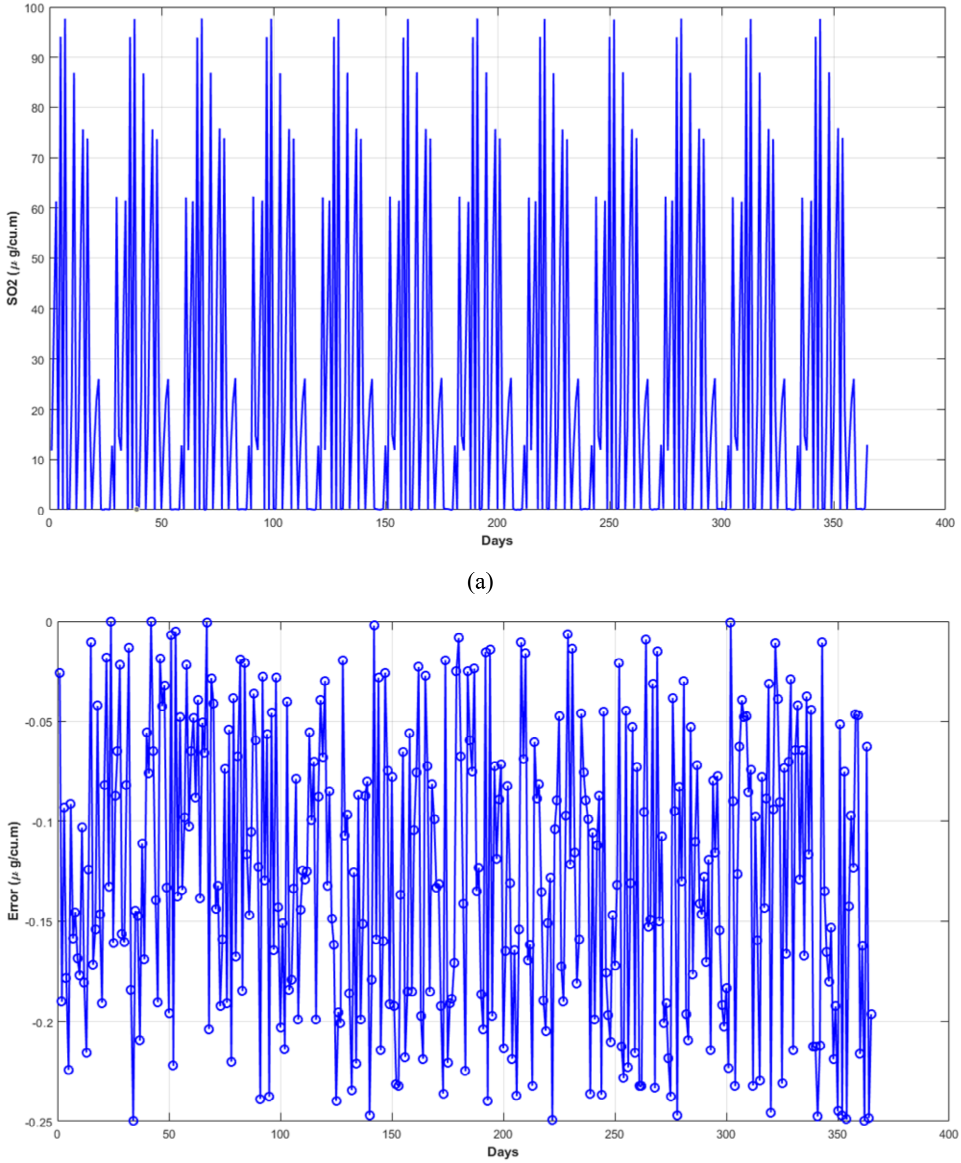
Overall, these measures provide a comprehensive understanding of the accuracy and performance of the models in predicting the concentrations of CO, NO_2, and SO_2 pollutants.

4.1.1 SO_2

The error evolution of SO_2 gas is evaluated with respective time period is detailed in Figure 4.

Figure 4 explains the SO_2 concentration in air pollution data among the time period of 350 days. At the day of 50 the SO_2 concentration in air pollutant is 150 g/cu.m as well as the error rate is about 0.23 ,a mismatch among predicted and the actual value.

Figure 4 Error evaluation of air pollution among time period, (a) quantity of SO_2 (b) error rate of SO_2 (see online version for colours)



(a)

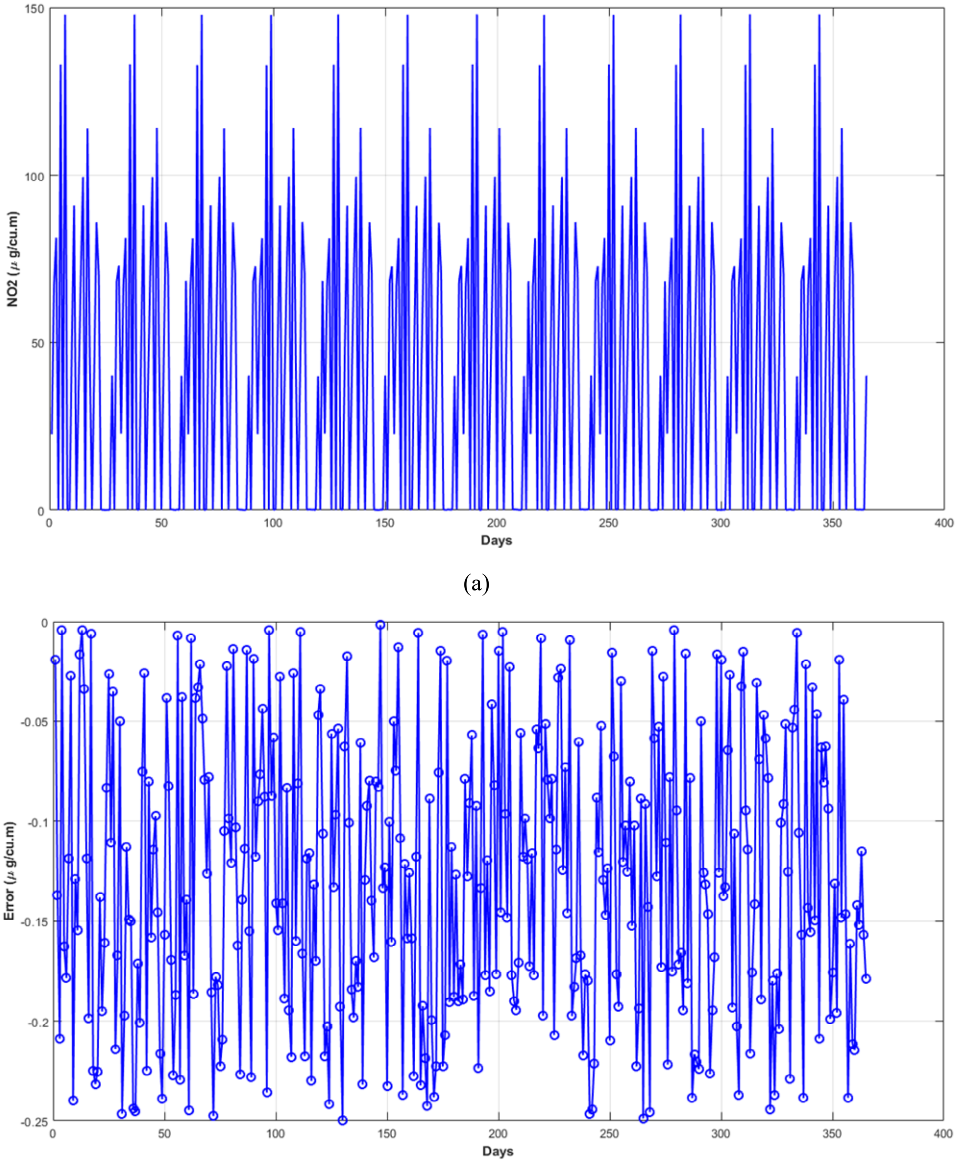
(b)

4.1.2 NO_2

The error evolution of NO_2 gas is evaluated with respective time period is detailed in Figure 5.

Figure 5 explains the NO_2 concentration in air pollution data among the time period of 350 days. At the day of 50 the NO_2 concentration in air pollutant is 150 g/cu.m as well as the error rate is about 0.25 ,a mismatch among predicted and the actual value.

Figure 5 Error evaluation of air pollution among time period, (a) quantity of NO_2 (b) error rate of NO_2 (see online version for colours)



(a)

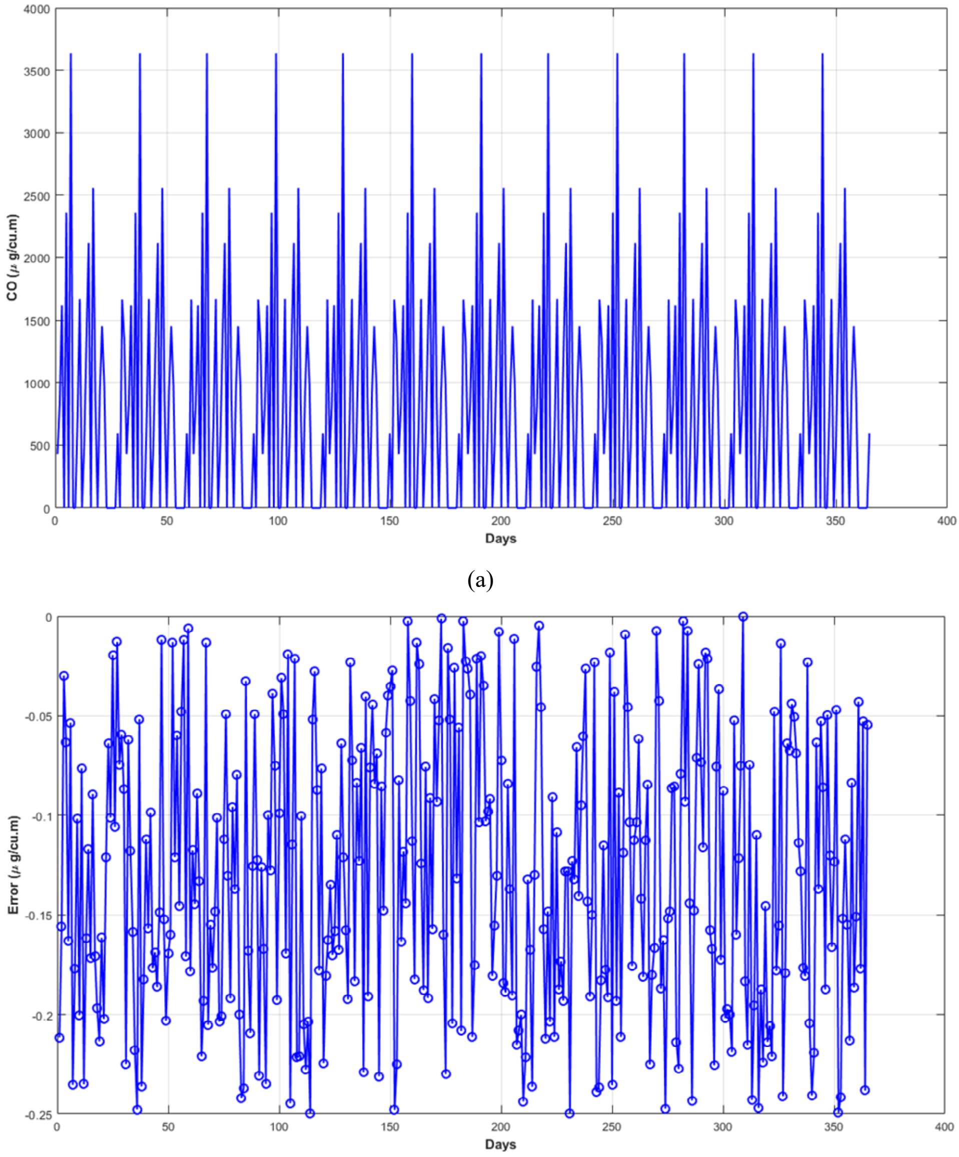
(b)

4.1.3 CO

Figure 6 displayed the error evolution of CO gas is evaluated with respective time period.

The concentration CO in air pollutant among the time period of 350 days is calculated and at the day of 50 the concentration leads to over 3,500 g/cu.m .The error rate at the same day shows the error value of 0.24 and it is show in Figure 6.

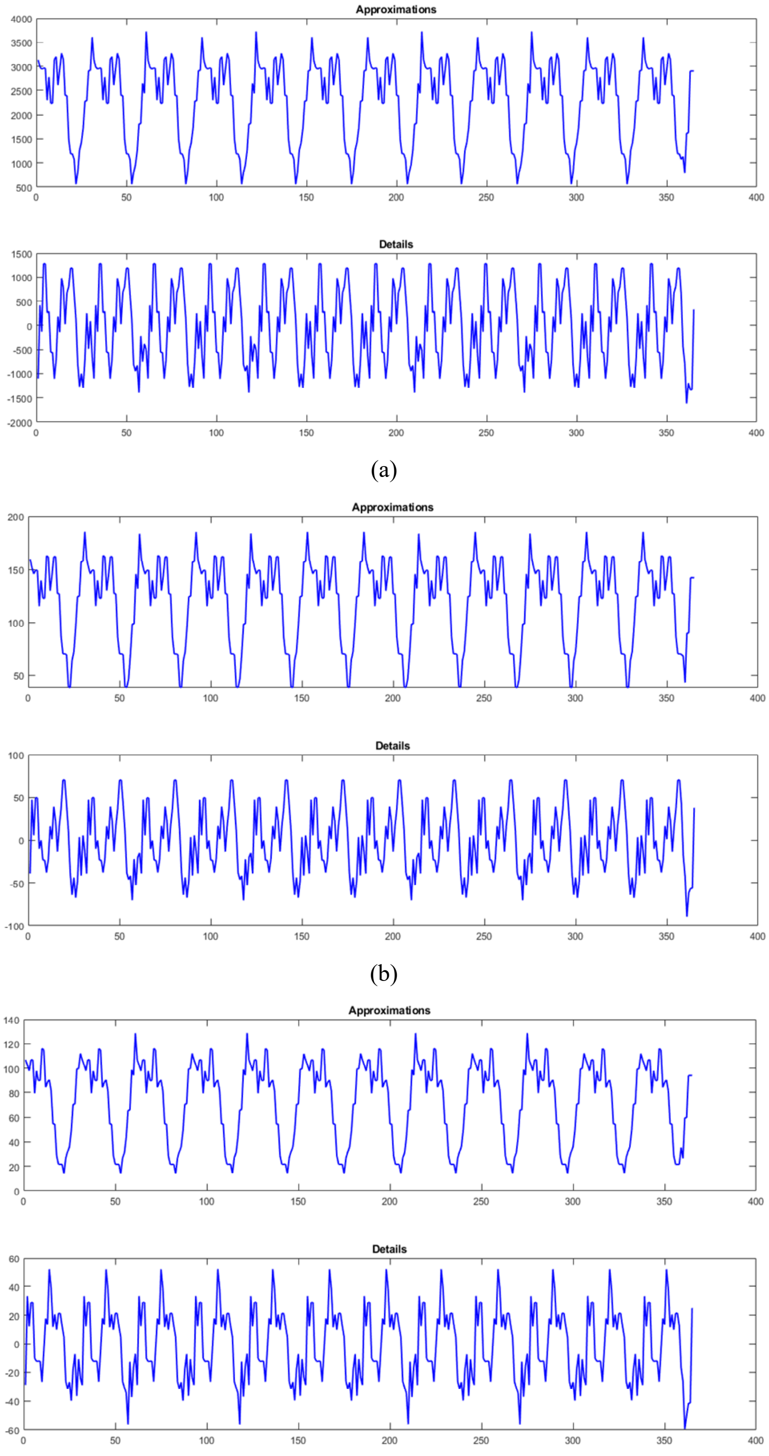
Figure 6 Error evaluation of air pollution among time period, (a) quantity of CO (b) error rate of CO (see online version for colours)



(a)

(b)

Figure 7 Predictive values of (a) CO, (b) NO_2, (c) SO_2 (see online version for colours)

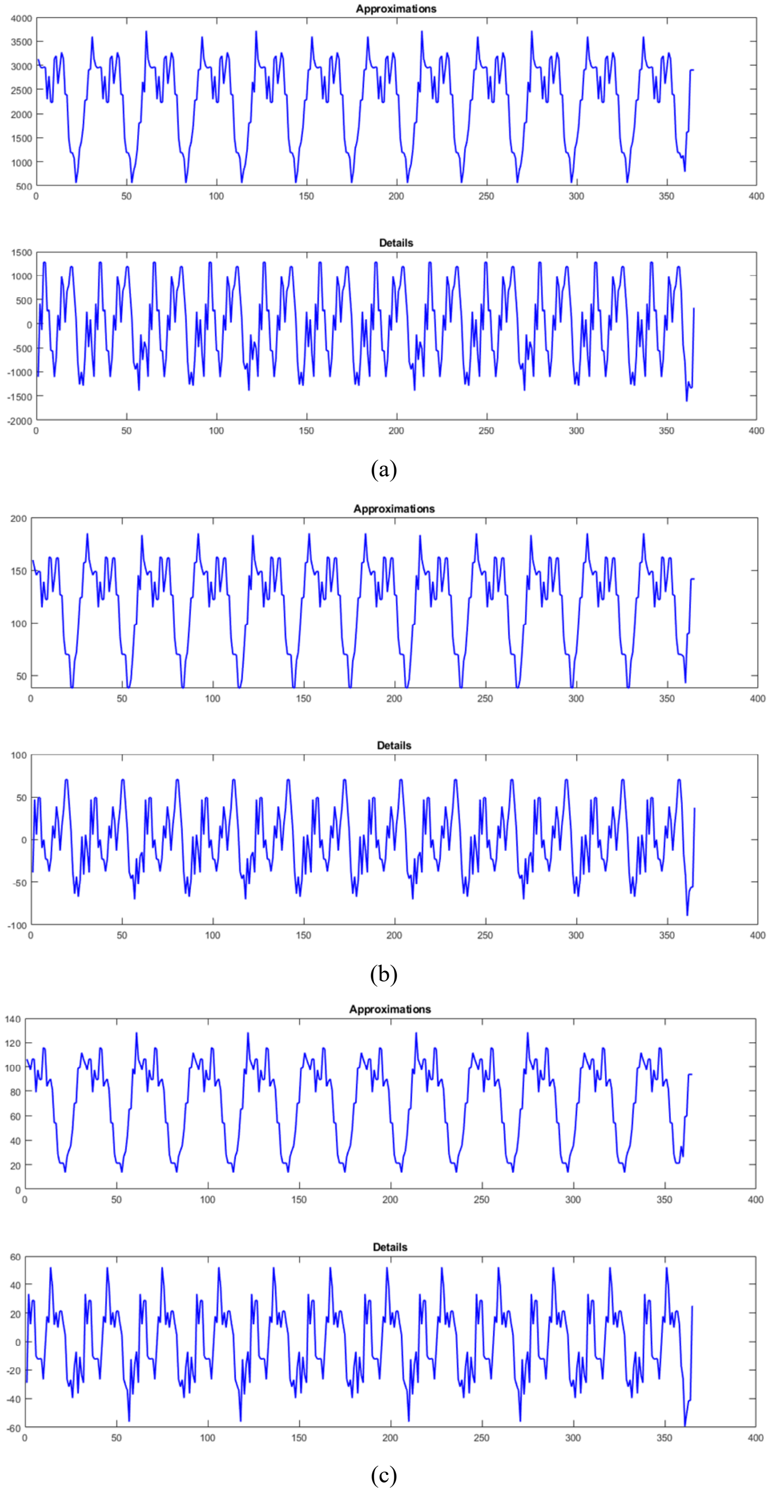


(a)

(b)

(c)

Figure 8 Actual values of (a) CO, (b) NO_2, (c) SO_2 (see online version for colours)



(a)

(b)

(c)

4.2 Experimental result

This section shows the values predicted and actual values obtained by analysing the air pollutant data and each pollutant gas data is predicted through SWT transform model. The values are graphically plotted. Figures 7 and 8 illustrate the predicted and the actual values of gas NO_2, SO_2 and CO. As per the recorded outcomes, the predicted values are found to be highly correlated with the actual value.

The actual values for the CO, NO_2 and SO_2 are shown in Figure 8. Figures 7 and 8 depict the levels of three air pollutants – NO_2, CO, and SO_2. Figure 7 shows both the actual and predicted values of the pollutants, while Figure 8 presents the approximation and details of the actual values. Starting with Figure 7, it is evident that the predictive value of CO is 4,000, which means that the estimated level of CO in the air is 4,000 units. Similarly, the predictive value of NO_2 is 150, indicating that the predicted concentration of NO_2 in the air is 150 units. On the other hand, the pollutant value of SO_2 is high, which implies that the estimated concentration of SO_2 is above the average level.

Now moving on to the actual values, Figure 8 shows that the pollutants' values of NO_2, SO_2, and CO are all in high concentration, with an approximate value of 200 g/cum. The approximation and details of the actual values help in understanding the distribution and sources of these pollutants in the air. In summary, the figures illustrate the levels of three air pollutants, with the predicted values of CO and NO_2 and high concentration of SO_2. The approximation and details of the actual values of these pollutants help in understanding their sources and distribution in the air.

5 Conclusions

The study investigated a daily atmospheric pollution forecast system using HRFOpSV algorithm and WD. The key component of this strategy is the individual wavelet prediction at several levels and the breakdown of the daily data into wavelets. The use of SVM in place of traditional MLP has made it possible to forecast wavelet coefficients and, as a result, the total pollutant concentration with much greater accuracy. Regardless of the pollutant type, the predicted results are well-aligned with the actual results. The main conclusion of these trials is that the applied predicting system has excellent generalisation capabilities and is capable of delivering accurate predictions for all stations for the various types of pollutants (CO, SO_2, CO_2).

References

Cabaneros, S.M., Calautit, J.K. and Hughes, B. (2020) 'Spatial estimation of outdoor NO_2 levels in Central London using deep neural networks and a wavelet decomposition technique', *Ecological Modelling*, Vol. 424, p.109017.

Chen, S., Wang, J-q. and Zhang, H-y. (2019) 'A hybrid PSO-SVM model based on clustering algorithm for short-term atmospheric pollutant concentration forecasting', *Technological Forecasting and Social Change*, Vol. 146, pp.41–54.

Du, S., Li, T., Yang, Y. and Horng, S-J. (2019) 'Deep air quality forecasting using hybrid deep learning framework', *IEEE Transactions on Knowledge and Data Engineering*, Vol. 33, No. 6, pp.2412–2424.

Dun, M., Xu, Z., Chen, Y. and Wu, L. (2020) 'Short-term air quality prediction based on fractional grey linear regression and support vector machine', *Mathematical Problems in Engineering*.

Fan, S., Hao, D., Feng, Y., Xia, K. and Yang, W. (2021) 'A hybrid model for air quality prediction based on data decomposition', *Information*, Vol. 12, No. 5, p.210.

Harishkumar, K.S., Yogesh, K.M. and Gad, I. (2020) 'Forecasting air pollution particulate matter (PM2.5) using machine learning regression models', *Procedia Computer Science*, Vol. 171, pp.2057–2066.

Heydari, A., Nezhad, M.M., Garcia, D.A., Keynia, F. and De Santoli, L. (2022) 'Air pollution forecasting application based on deep learning model and optimization algorithm', *Clean Technologies and Environmental Policy*, Vol. 24, No. 2, pp.607–621.

Jiang, N., Fu, F., Zuo, H., Zheng, X. and Zheng, Q. (2020) 'A municipal PM2.5 forecasting method based on random forest and WRF model', *Engineering Letters*, Vol. 28, No. 2.

Jin, N., Zeng, Y., Yan, K. and Ji, Z. (2021) 'Multivariate air quality forecasting with nested long short term memory neural network', *IEEE Transactions on Industrial Informatics*, Vol. 17, No. 12, pp.8514–8522.

Kumar, D. (2018) 'Evolving differential evolution method with random forest for prediction of air pollution', *Procedia Computer Science*, Vol. 132, pp.824–833.

Leong, W.C., Kelani, R.O. and Ahmad, Z. (2020) 'Prediction of air pollution index (API) using support vector machine (SVM)', *Journal of Environmental Chemical Engineering*, Vol. 8, No. 3, p.103208.

Li, L. and He, Z. (2020) 'Atmospheric pollutant prediction based on wavelet decomposition and long short-term memory network', in *IOP Conference Series: Materials Science and Engineering*, IOP Publishing, Vol. 768, No. 7, p.072059.

Li, X., Luo, A., Li, J. and Li, Y. (2019) 'Air pollutant concentration forecast based on support vector regression and quantum-behaved particle swarm optimization', *Environmental Modeling & Assessment*, Vol. 24, No. 2, pp.205–222.

Li, Y. and Tao, Y. (2018) 'Daily PM 10 concentration forecasting based on multiscale fusion support vector regression', *Journal of Intelligent & Fuzzy Systems*, Vol. 34, No. 6, pp.3833–3844.

Liu, B., Yu, X., Chen, J. and Wang, Q. (2021a) 'Air pollution concentration forecasting based on wavelet transform and combined weighting forecasting model', *Atmospheric Pollution Research*, Vol. 12, No. 8, p.101144.

Liu, B., Zhang, L., Wang, Q. and Chen, J. (2021b) 'A novel method for regional NO_2 concentration prediction using discrete wavelet transform and an LSTM network', *Computational Intelligence and Neuroscience*.

Liu, D-R., Lee, S-J., Huang, Y. and Chiu, C-J. (2020) 'Air pollution forecasting based on attention-based LSTM neural network and ensemble learning', *Expert Systems*, Vol. 37, No. 3, p.e12511.

Liu, H., Li, Q., Yu, D. and Gu, Y. (2019) 'Air quality index and air pollutant concentration prediction based on machine learning algorithms', *Applied Sciences*, Vol. 9, No. 19, p.4069.

Mishra, A. (2018) 'Air pollution monitoring system based on IoT: forecasting and predictive modeling using machine learning', in *International Conference on Applied Electromagnetics, Signal Processing and Communication (AESPC)*.

Qiao, W., Wang, Y., Zhang, J., Tian, W., Tian, Y. and Yang, Q. (2021) 'An innovative coupled model in view of wavelet transform for predicting short-term PM10 concentration', *Journal of Environmental Management*, Vol. 289, p.112438.

Tao, Q., Liu, F., Li, Y. and Sidorov, D. (2019) 'Air pollution forecasting using a deep learning model based on 1D convnets and bidirectional GRU', *IEEE Access*, Vol. 7, pp.76690–76698.

Wang, J., Du, P., Hao, Y., Ma, X., Niu, T. and Yang, W. (2020a) 'An innovative hybrid model based on outlier detection and correction algorithm and heuristic intelligent optimization algorithm for daily air quality index forecasting', *Journal of Environmental Management*, Vol. 255, p.109855.

Wang, Z., Chen, L., Ding, Z. and Chen, H. (2020b) 'An enhanced interval PM2.5 concentration forecasting model based on BEMD and MLPI with influencing factors', *Atmospheric Environment*, Vol. 223, p.117200.

Wen, C., Liu, S., Yao, X., Peng, L., Li, X., Hu, Y. and Chi, T. (2019) 'A novel spatiotemporal convolutional long short-term neural network for air pollution prediction', *Science of the Total Environment*, Vol. 654, pp.1091–1099.

Wu, Q. and Lin, H. (2019) 'A novel optimal-hybrid model for daily air quality index prediction considering air pollutant factors', *Science of the Total Environment*, Vol. 683, pp.808–821.