Date of publication xxxx 00, 0000, date of current version xxxx 00, 0000.

Digital Object Identifier 10.1109/ACCESS.2017.DOI

Two Stage Feature Engineering to Predict Air Pollutants in Urban Areas

FAREENA NAZ[1], MUHAMMAD FAHIM[1], ADNAN AHMAD CHEEMA[2], (Member, IEEE), NGUYEN TRUNG VIET[3], TUAN-VU CAO[4], RUTH HUNTER[5], AND TRUNG Q. DUONG[1,6], (Fellow, IEEE)

[1]School of Electronics, Electrical Engineering and Computer Science, Queen's University Belfast, BT7 1NN, United Kingdom (e-mail: {fnaz01, m.fahim, trung.q.duong}@qub.ac.uk)
[2]School of Engineering, Ulster University, Belfast, BT15 1AP, United Kingdom (e-mail:a.cheema@ulster.ac.uk)
[3]Thuyloi University, Hanoi, Vietnam (e-mail:nguyentrungviet@tlu.edu.vn)
[4]Norwegian Institute for Air Research, Oslo, Norway (e-mail: tvc@nilu.no)
[5]Centre for Public Health, School of Medicine, Dentistry and Biomedical Sciences, Queen's University Belfast, BT7 1NN, United Kingdom (e-mail: ruth.hunter@qub.ac.uk)
[6]Faculty of Engineering and Applied Science, Memorial University of Newfoundland, St. John's, NL A1C 5S7, Canada (e-mail: tduong@mun.ca)

Corresponding author: Trung Q. Duong (e-mails: tduong@mun.ca, trung.q.duong@qub.ac.uk).

This work has been accepted in part for a presentation at the 10th International Conference on Industrial Networks and Intelligent Systems (INISCOM 2024), February 2024. The work of Trung Q. Duong was supported in part by the Canada Excellence Research Chair program. The work of Muhammad Fahim, Tuan-Vu Cao, and Trung Q. Duong was supported in part by UKRI and European Commission under MISO Project, "Autonomous Multi-Format In-Situ Observation Platform for Atmospheric Carbon Dioxide and Methane Monitoring in Permafrost and Wetlands." The work of Ruth Hunter and Trung Q. Duong was supported by the SPACE (Supportive Environments for Physical and Social Activity for Cognitive Health) Project (https://www.qub.ac.uk/sites/space/) funded by the ESRC Healthy Ageing Challenge under Grant ES/V016075/1.

ABSTRACT Air pollution is a global challenge to human health and the ecological environment. Identifying the relationship among pollutants, their fundamental sources and detrimental effects on health and mental well-being is critical in order to implement appropriate countermeasures. The way forward to address this issue and assess air quality is through accurate air pollution prediction. Such prediction can subsequently assist governing bodies in making prompt, evidence-based decisions and prevent further harm to our urban environment, public health, and climate, all of which co-benefit our economy. In this study, the main objective is to explore the strength of features and proposed a two stage feature engineering approach, which fuses the advantage of influential factors along with the decomposition approach and generates an optimum feature combination for five major pollutants including Nitrogen Dioxide (NO_2), Ozone (O_3), Sulphur Dioxide (SO_2), and Particulate Matter (PM2.5, and PM10). The experiments are conducted using a dataset from 2015 to 2020 which is publicly available and is collected from Belfast-based air quality monitoring stations in Northern Ireland, UK. In stage-1, using the dataset new features such as trigonometric and statistical features are created to capture their dependency on the target pollutant and generated correlation-inspired best feature combinations to improve forecasting model performance. This is further enhanced in stage-2 by an optimum feature combination which is an integration of stage-1 and Variational Mode Decomposition (VMD) based features. This study employed a simplified Long Short Term Memory (LSTM) neural network and proposed a single-step forecasting model to predict multivariate time series data. Three performance indicators are used to evaluate the effectiveness of forecasting model: (a) root mean square error (RMSE), (b) mean absolute error (MAE), and (c) R-squared (R^2). The results demonstrate the effectiveness of proposed approach with 13% improvement in performance (in terms of R^2) and the lowest error scores for both RMSE and MAE.

INDEX TERMS Air quality; feature engineering; variational mode decomposition; machine learning; predictive model.

I. INTRODUCTION

AIR pollution is one of the major global environmental health issues caused by the rapid rise in urbanisation and industrialisation. It has become the biggest threat to our health and the environment we live in. Around 99% of our global population breathes air that contains high levels

of pollutants and leads to increased morbidity and mortality [1], [2]. From neurological, respiratory, cardiovascular, and metabolic to reproductive, every system in the body is affected by air pollution. Each year 6.7 million premature deaths are recorded worldwide, with low and middle income nations accounting for 95% of these deaths [3]. However, to mitigate the effects of pollution on health, environment, economy, and climate, the United Nations (UN) has established sustainable development goals (SDGs) such as 3, 7, and 11. These goals set targets for 2030 with the aim to reduce deaths, illness, and adverse environmental effects in cities by facilitating access to clean and sustainable energy, transportation, and urbanisation with green and blue spaces. The WHO recently released air quality guidelines to establish evidence based global targets to protect public health by enhancing air quality [1]. Likewise, the government of the United Kingdom (UK) has set a goal to curtail 35% of air pollution by 2040 [4].

Generally, air quality is influenced by numerous factors involving local geography, weather, and sources of emissions. In Northern Ireland (NI), major sources of pollutant emission mostly revolve around the combustion of fossil fuels at domestic, transportation, and industrial levels [5]. Pollutants like Nitrogen Dioxide (NO_2) and Sulphur Dioxide (SO_2) are directly released into the air because of combustion processes involving fossil fuels (e.g. coal and oil) in transportation, industrial, commercial, power refineries, and electrical supply sectors. In various regions of the UK, particularly NI, coal is regarded as a significant domestic energy source which explains why these gases are found predominant in emissions. Exposure to these gases irritates the respiratory tract which increases the likelihood of cough, infection, mucus formation, and chronic lung disease. Additionally, also causes damage to our ecosystem with acid rain, reduced photosynthesis, chlorophyll degradation, damage to foliage, acidification of water, and soil which subsequently leads to a decline in biodiversity. In terms of Ozone (O_3) at the ground level, unlike other man-made sourced emissions, it is indirectly emitted in the air because of a photochemical reaction formed between Nitrogen dioxide and volatile organic compounds in the presence of sunlight. It takes hours or days to form and rural areas are the ones most affected due to its long range movement far from its original site of emission. High exposure to O_3 damage airways, irritate eyes and nose, and necessitates hospitalisation. In addition, also cause damage to forest, plant species, and biodiversity. Particulate Matter (PM) which includes PM2.5 and PM10, is typically classified based on the particle size. For instance, a particle less than 2.5 μm diameter is referred to as PM2.5, and a particle smaller than 10 μm in diameter is referred to as PM10. Particulate matter particularly PM2.5 is considered one of the primarily focused pollutants which pose the greatest threat to human health and the environment. In the UK, primary PM is emitted directly into the air because of man-made sources primarily by fuel combustion, engine emission from road transportation, tyre, and brake wear, and other non-exhaust emissions from industries. Whereas, secondary PM is formed by chemical reactions in air from the emission of certain pollutants such as SO_2, Nitrogen Oxide (NO_X), Ammonia (NH_3), and organic compounds sourced from either vegetation or combustion. Both short and long term exposure cause cardiovascular and respiratory diseases along with cognitive decline, other ill health effects, and mortality [6]. In addition, the WHO and Committee on the Medical Effects of Air Pollutants (COMEAP) recently reviewed that exposure to PM2.5 is strongly associated with adverse health impacts [7]. In addition, few recent studies also look into monitoring and modelling of aerosol and air pollution [8]–[12].

Identification of pollutants, their sources of emission, and accurate prediction of their concentration is vital and facilitates the authorities and governing bodies in making evidence-based decisions. They can further put policies and controls in place, where needed to prevent further loss, help public demand, and build healthier communities to improve air quality. Interdisciplinary collaboration of experts with other stakeholders is fundamental to tackling such challenges and helps in educating and public awareness [13]. As we know, all of this is a result of utilising different energy sources to facilitate our lives, in return giving rise to pollution and deteriorating air quality, health, environment, ecosystem, and climate. At this point, experts and measures alone are never enough to resolve this challenge until the public accepts responsibility for their actions and adopts healthy lifestyle modifications, such as walking, cycling, or taking public transport whenever possible instead of driving a car. Additionally, switching to electric cars, using renewable power sources (combustion-free), authorised low emission fuels and installing exempted fireplaces to control smoke, ensuring the boiler is up to date, and having adequate home insulation can all help. Positive incremental improvement is seen in NI air quality compared of what it was even before the industrial revolution via strict successful policies implementations on the emission of certain pollutants from major sources [14]. A few of the successful policies adhered to in NI include the introduction of smoke control areas with the strict usage of only authorised fuels for appliances and exempted fireplaces. Similarly in London (the capital of England) after a successful trial of the Low Emission Zone (LEZ) scheme, the Ultra Low Emission Zone (ULEZ) is recently expanded now and resulted in a significant i.e., 50% reduction in NO_2 emission and 5 times fastest pollution reduction compared to other parts of UK since 2016-2020 [15].

The main contributions of this study include:

- We propose a two stage feature engineering approach. In stage-1, we have considered features that are available in the dataset and created new features to capture dependencies of features with target pollutants. A total of 22 features are generated among categories like meteorological, temporal, statistical and air pollutants. In stage-2, we further used variational mode decomposition (VMD) to generate new features to capture

dependencies with respective pollutants in discussion.

- We experimentally evaluated and comprehensively analysed all feature combinations for target pollutants and proposed a unique optimum feature combination that can improve a simplified forecasting model performance in terms of root mean square error (RMSE), mean absolute error (MAE), and R-squared (R^2). We proposed a two stage feature selection method which in stage-1 performs feature selection using correlation and optimises the performance of forecasting model by further integrating VMD based features (i.e. based on selection of optimum number of IMFs) in stage-2.
- We provide a detailed performance evaluation of optimum feature combination for a total of 5 pollutants (NO_2, SO_2, O_3, PM2.5, and PM10) by considering a simplified LSTM based forecasting model and investigated the key features that can influence various pollutants differently.

The remainder of the paper is organised as follows: related work and contributions are provided in Section II. Section III describes the dataset and Section IV provides details about two stage feature engineering. Model training and testing are discussed in Section V. Results and discussion are provided in Section VI and finally, the paper is concluded in Section VII.

II. RELATED WORK

Machine learning (ML) has revolutionised many scientific domains to tackle intricate engineering challenges, particularly ML-based feature engineering and regression models play a pivotal role in air pollution forecasting. It shows notable progression in research because of its accurate prediction, low-cost implementation, and flexible adaptability. To handle high dimensional large-scale data gathered from 35 air quality monitoring stations situated in Beijing, a light gradient boosting machine model is proposed in [16]. In addition to air pollutants, statistical, temporal, and meteorological features, they used the following 24 hours of weather prediction data as predictive data features to predict the PM2.5 concentration for the following 24 hours. Based on the correlation of features, the performance of the model is compared with other models such as Adaptive boosting (Adaboost), gradient boosting decision tree (GBDT), extreme gradient boosting (XGboost), and deep neural network (DNN) and findings revealed that their model outperformed others under indicators such as symmetric mean absolute percentage error (SMAPE), mean square error (MSE) and MAE. In [17], short term forecasting hybrid approach combining convolutional neural network (CNN) and bidirectional gated recurrent unit (GRU) was proposed to predict PM2.5 concentration in Beijing. Several feature combinations were tested based on the correlation analysis of time series data and found that the performance of the proposed model is better when historical data of pollutant and meteorological factors such as temperature, dew point, wind direction, and speed are used. When compared with shallow ML models and GRU, the suggested model demonstrated a notable improvement in terms of error score. An encoder-decoder LSTM model is proposed with Genetic algorithm (GA) feature selection to predict PM2.5 concentration using two datasets collected from Hanoi and Taiwan [18]. The datasets comprised of meteorological and air pollutant features. Several feature combinations were tested and the results showed that the best combination relied on wind, temperature, radiation, PM2.5, and PM10. Their proposed approach enhanced prediction accuracy using MAE as an assessment metric. In a similar study, a multitask learning model using LSTM autoencoder is presented to predict PM2.5 time series across Beijing city at several locations. LSTM was intended for learning spatial-temporal PM2.5 time series features and autoencoder for encoding meteorological parameters. The dataset was collected from 18 air quality and 13 meteorological monitoring stations. While meteorological time series includes temperature, pressure, humidity, wind speed, and direction, air quality data covers the spatiotemporal aspects of numerous sites [19].

A hybrid deep learning model is proposed which combines multiple one-dimensional (1D) CNN and bi-directional LSTM (BiLSTM) to predict single-step and multi-step (48 hrs) PM2.5 concentration using two datasets [20]. The dataset includes hourly meteorological and air pollutant attributes collected over four years. Local trends and spatial features are extracted using 1D-CNN, while LSTM is used to learn spatial-temporal dependencies. The results indicated good forecasting ability when compared with SVR, LSTM variants, CNN, and RNN. Similarly, in another study, a hybrid ML approach is proposed to predict PM2.5 concentration following the consequences of conflict in the city of Kiev [21]. They proposed optimising multilayer perceptron neural network (MLPNN) using electromagnetic field optimisation (EFO) algorithm to get better prediction. Three distinct sources of data were gathered, and the data included temporal, air pollutants, and meteorological aspects. In addition, principal component analysis (PCA) is used to reduce the data and choose the helpful factors. Though the study had certain limitations, however, the results demonstrated the competency of the models for use in real-world scenarios in Kiev. In [22], Dual LSTM is proposed which combines the single and multi-factor prediction models. Both models used sequence-to-sequence (seq2seq) technology which contains an encoder and decoder while an extreme gradient boosting (XGBoost) regression tree is opted for integration. The dataset contains hourly based spatiotemporal features collected from multiple stations in Beijing. Experimental results indicated improvement in error under five assessment indices in comparison to other models. Moreover, another study proposed a dual-stage attention based on conversion-gated LSTM (DA-CG-LSTM) to predict air quality and traffic flow. To improve the ability of the model to capture short term mutation information, a hyperbolic tangent function is introduced in input and forget gates. Additionally, dual staged attention is added in terms of input and temporal attention. Experimental results show a 50% lower error rate in compar-

ison to dual-staged attention recurrent neural network (DA-RNN) and transformation-gated LSTM (TG-LSTM) [23].

Recent research shows the superiority of hybrid models based on decomposition and ensemble over the single forecasting model. For instance, a recent study proposed a dual layer decomposition and the feedback of the model learning effect for the prediction of PM2.5 concentration [24]. Initially, ensemble empirical mode decomposition (EEMD) is used for decomposing PM2.5 time series followed by sample entropy (SE) methodology and then VMD is employed where SE is higher than the average value. A wavelet neural network (WNN) model is established for each sub-series prediction which is later combined to get the final prediction. Additionally, the network frame structure and prediction ability of the model are improved using feedback of the learning effect. In another study, a VMD based BiLSTM model is proposed for single-step prediction of PM2.5 concentration in various cities of China [25]. In this work, BiLSTM is employed separately for all sub-series decomposed by VMD and concatenated all at last to get the final prediction. Results based on comprehensive analysis with other EMD and VMD based models show improvement in prediction accuracy and error. This study recommends VMD over other signal processing techniques in combination with BiLSTM. A novel hybrid model is proposed for AQI prediction using three datasets collected from Beijing, Tianjin, and Shijiazhuang [26]. Here, a secondary decomposition is proposed which is based on empirical wavelet transform (EWT) for the initial decomposition of AQI time series and VMD for the second decomposition of the sub-series with larger entropy values. In addition, optimal features are extracted using an imperialist competitive algorithm (ICA), and the echo state network (ESN) model is used for the prediction of each sub-series and obtaining the final prediction by integration.

In [27], the parameters of VMD and LSTM models are optimised based on enhanced versions of sparrow search algorithms (SSA) for a single-step AQI prediction. The dataset is used from three different locations in China and the proposed model performance is evaluated on test data and validation data for generalisation ability. In the proposed model, LSTM is used for each IMF (intrinsic mode function), also known as sub-series, and it is found that SSA based VMD-LSTM model has better prediction and generalisation performance. In [28], SE is introduced to reduce the total number of IMFs and AQI from two cities in China is predicted using LSTM models. The AQI prediction is obtained by summing the prediction from each LSTM model. An optimal hybrid model based on secondary decomposition and air pollutant factors for forecasting AQI is proposed in [29]. Primarily, wavelet decomposition (WD) is used to decompose the AQI series into high and low frequency sub-series. High frequency series are further decomposed by VMD-SE to smooth the series and later LSTM is adopted for modelling each decomposed sub-series. While for low frequency series, the least squares support vector machine (LS-SVM) along with bat optimisation algorithm is employed and also considered the effect of air pollutant factors such as NO_2, SO_2, CO, PM2.5 and PM10. The final result is attained by aggregating the predictions of forecasting models for each sub-series. In [30], Dung Beetle Optimisation is used to optimise the VMD decomposition and XGBoost model for PM2.5 single-step prediction. By using correlation, feature filtering is performed and further features are categorised based on the frequency of IMFs and a combination of XGBoost and informer models are used for prediction. Although aforementioned studies have investigated different aspects of feature engineering, feature selection focuses on a single pollutant only (e.g. mostly PM2.5) and complex forecasting models. However, still requires careful consideration to understand the relationship between the target and features and how this information can be used to define a set of optimum features which can improve the performance of a simplified forecasting model. In addition, it is important to find such optimum features for a wide range of pollutants which can allow forecasting using a very simplified common model.

III. DATASET

In this study, the dataset used is comprised of over 50,000 samples measured by an air quality monitoring station situated in Belfast city center, Northern Ireland from 2015 to 2020 [31], [32]. This dataset includes hourly concentration levels of meteorological data and air quality parameters. Meteorological data involves temperature (°C), wind horizontal and wind vertical whereas air quality parameters include NO_2, O_3, SO_2, PM2.5, PM10, Nitric Oxide (NO), NO_X and Carbon Monoxide (CO).

Table.1 provides statistical information such as total count, mean, standard deviation, minimum and maximum value of meteorological data. The mean and standard deviation of all parameters ranges from -2.15 to 8.27 and 3.66 to 4.50, respectively. In addition, minimum and maximum values of all parameters fall between -19.75 to 0 and 15.85 to 24, respectively. The statistical descriptions of the pollutants being predicted in this work are listed in Table. 2, albeit the dataset contains more than these. The NO_2 concentration has a mean of 26.12 with a standard deviation of 17.87 and the values vary from 1 to 203. However, SO_2 has the lowest mean and standard deviation among all which is 1.55 and 1.6, respectively.

TABLE 1. Statistical description of meteorological data.

	Count	Moment (mean, std)	Boundary (min, max)	Percentile (25, 50, 75)
Temperature	52564	(8.27, 4.43)	(0, 24)	(4.9, 8, 11.5)
Wind Horizontal	52564	(-1.06, 3.66)	(-18.53, 15.85)	(-3.4, -1.1, 1.4)
Wind Vertical	52564	(-2.15, 4.50)	(-19.75, 16.68)	(-5.2, -2.6, 1)

IV. FEATURE ENGINEERING

There is widespread agreement that models attempt to reach the limit determined by data and features in ML. Therefore,

TABLE 2. Statistical description of air pollutants in $\mu g/m^3$.

	Count	Moment (mean, std)	Boundary (min, max)	Percentile (25, 50, 75)
NO_2	52564	(26.12, 17.87)	(1, 203)	(13, 22, 35)
O_3	52567	(43.24, 20.65)	(0, 150)	(29, 44, 58)
SO_2	52514	(1.55, 1.6)	(0, 20)	(1, 1, 2)
PM2.5	52545	(9, 7.94)	(0, 104)	(4, 7, 11)
PM10	52510	(14.15, 10.6)	(0, 143)	(8, 12, 17)

the goal is to find the optimum set of features by exploring their respective strength from the given time series data with expectations to have significant improvement in model prediction, training time, and complexity. In this study, we grouped features into four types based on characteristics which include meteorological, temporal, statistical, and air pollutants. In meteorological features, we have considered temperature, wind horizontal and wind vertical since high temperature affects the airflow and strong winds modify the concentration of various pollutants, thus both impact the air quality [16], [33]. In terms of temporal features, datetime index contained in the dataset is utilised to create nine additional features. Initially, datetime index is split into hour, day, and month features. Since day, month, and hour are cyclical variables, trigonometric functions such as sine and cosine are applied to them to create six additional features including month_sin, month_cos, day_sin, day_cos, hour_sin, and hour_cos. With this encoding, the model is better able to capture the cyclic temporal relationships which further enhance the model performance [34]. Table. 3 provides list of notations along with description used in this study. For a given feature $z(t)$, trigonometric features can be generated using (1)-(2):

$$z_{sin}(t) = sin(2\pi z(t)/P), \tag{1}$$

$$z_{cos}(t) = cos(2\pi z(t)/P), \tag{2}$$

where P is the period which is 12, 24, and 31 for month, hour and day data, respectively.

In the statistical feature, we only considered the mean of the previous two hours of the pollutant being predicted. Air pollutant features include eight pollutants NO_2, O_3, SO_2, PM2.5, PM10, NO, NO_X, and CO. In addition, a lag feature is created which is based on the previous hour concentration value of the pollutant being predicted. In summary after feature engineering, a total of 22 features are introduced including 3 meteorological, 9 temporal, 1 statistical, and 9 air pollutants to be used as an input to the model as per relevance with the targeted pollutant as listed in Table. 4.

A. CORRELATION BASED FEATURE SELECTION

Given the fact that irrelevant features not only increase the training time but also add to computational cost, appropriate feature selection is critical for better prediction. For this reason, a filter mechanism is required that performs feature selection independent of the chosen forecasting model. In this work, we have considered a Pearson correlation-based feature selection which is recommended when dealing with numerical features and to confirm collinearity between features and target [35]. The Pearson correlation coefficient r between feature $z(t)$ and target output $x(t)$ is defined as:

$$r = \frac{\sum_{i=1}^{M}(z_i - \bar{z})(x_i - \bar{x})}{\sqrt{\sum_{i=1}^{M}(z_i - \bar{z})^2 \sum_{i=1}^{M}(x_i - \bar{x})^2}}, \tag{3}$$

where z_i and x_i are i^{th} data samples, $\bar{z}$ and $\bar{x}$ are the mean and M is a total number of samples.

All the features with positive correlation are selected in this work. Table. 4 shows the correlation of the features, all positively correlated features are tinted green (dark and light), while negatively correlated features are represented in lime tint. In, addition, dark green represent the best feature combination found for each pollutant (more detail is discussed in Section V(2). For instance, in case of NO_2, all positively correlated features include air pollutants with lag feature, mean from statistical, month_sin, month_cos, day, hour from temporal, and wind vertical from meteorological are considered. Whereas, negatively correlated features such as temperature, wind horizontal, day_sin, day_cos, month, hour_sin, hour_cos, and O_3 are eliminated and not considered in the prediction of NO_2.

TABLE 3. List of notations with description.

Notations	Description
z	Feature from dataset
$\bar{z}$	Mean of a feature
x	Target output from dataset
$\bar{x}$	Mean of target output
q	Target output from test data
$\bar{q}$	Mean of target output from test data
$\hat{q}$	Estimated target output
P	Period (i.e, hours, days or months)
M	Total number of samples in dataset
T	Total number of samples in test data
U	Real-valued signal
u_k	Narrowband sub-signal or IMF
R	Residual signal
K	Total number of IMFs
s	Input of the LSTM cell
e	Output of forget gate
l	Output of the input gate
y	Output of the output gate
d	Current state of a LSTM cell
w	Weights of a gate in LSTM cell
b	Bias factor of each gate of LSTM cell
σ	Sigmoid activation function
g	Final output of the LSTM cell

B. VARIATIONAL MODE DECOMPOSITION BASED FEATURE GENERATION

VMD is a signal decomposition method which decomposes a real-valued signal $U(t)$ into a finite number of narrowband

TABLE 4. Correlation of all features w.r.t target pollutants.

Type	Feature	NO_2	O_3	SO_2	PM2.5	PM10
Meteorological	Temperature	-0.25	0.16	-0.02	-0.19	-0.12
	Wind Horizontal	-0.03	0.03	**0.01**	-0.08	-0.05
	Wind Vertical	**0.13**	-0.02	**0.1**	0.25	0.25
Temporal	Day	**0.003**	-0.001	-0.01	**0.02**	0.02
	Day_Sin	-0.01	0.01	0.01	-0.03	-0.03
	Day_Cos	-0.01	0.02	-0.02	-0.03	-0.05
	Month	-0.01	-0.21	0.01	-0.07	-0.07
	Month_Sin	**0.1**	0.26	0.01	**0.17**	0.15
	Month_Cos	**0.25**	-0.27	0.07	**0.15**	0.05
	Hour	**0.13**	0.02	0.11	**0.07**	0.07
	Hour_Sin	-0.04	-0.1	-0.07	-0.06	-0.05
	Hour_Cos	-0.21	-0.01	-0.1	**0.02**	-0.07
Statistical	Mean (previous 2 hrs)	0.92	0.95	**0.8**	0.96	0.92
Air Pollutant	NO_2	1	-0.57	**0.4**	0.49	**0.42**
	O_3	-0.56	1	-0.2	-0.28	-0.2
	SO_2	0.35	-0.15	1	0.38	**0.33**
	PM2.5	0.43	-0.25	**0.35**	1	**0.73**
	PM10	0.31	-0.14	**0.26**	0.62	1
	Nitric Oxide	0.55	-0.35	**0.5**	0.43	**0.39**
	Nitrogen Oxide	0.7	-0.45	**0.51**	0.49	**0.43**
	Carbon Monoxide	0.42	-0.35	**0.39**	0.51	**0.41**
	Lag (previous 1 hr)	**0.86**	**0.92**	**0.83**	**0.92**	**0.84**

sub-signals $u_k(t)$ (also known as IMFs or sub-series) [36]. In this method, each IMF is represented by amplitude-frequency modulated signal as:

$$u_k(t) = A_k(t)cos(\phi_k(t)), \tag{4}$$

$$U(t) = \sum_{k=1}^{K} u_k(t) + R, \tag{5}$$

where $A_k(t)$ is envelope, $\phi_k(t)$ is phase, K is the total number of IMFs, R is a residual signal and instantaneous frequency can be found as $\phi_k^{'}(t)$ which is non decreasing and varies around the central frequency of the respective mode. In recent years, VMD method has gained much attention as a new feature engineering method in different applications [37]–[39]. In this work, we aim to use VMD method to generate additional new features based on the hourly lag of the pollutant being predicted and investigate an optimum number of IMFs required which can further improve forecasting model performance. Fig. 1 shows an example of lag NO_2 decomposition using 3 IMFs and residual data however careful consideration is required to select parameter K so that such features can improve forecasting model performance.

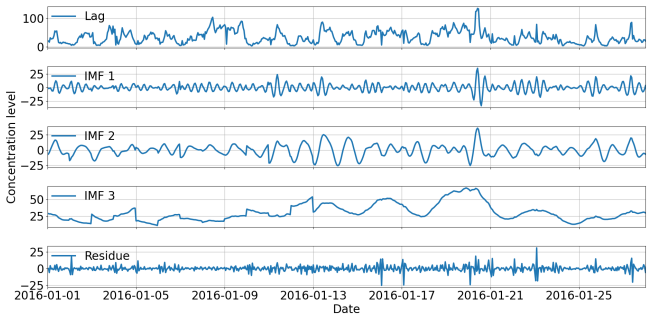

FIGURE 1. Decomposition of lag NO_2 into IMFs and Residual plot.

V. MODEL TRAINING AND TESTING

This section provides details about the data preparation, two stage feature engineering with selection and model training and testing of the single-step forecasting model. Fig. 2 shows the workflow of model training and testing with key components.

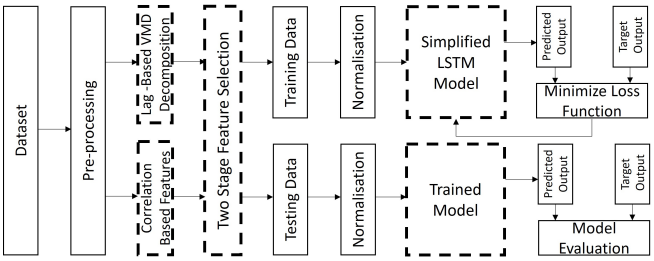

FIGURE 2. Workflow of model training and testing with two stage feature engineering and selection approach.

1) Data Pre-processing

In general, datasets may have outliers, missing or recurring values known as invalid values. Outliers are extreme or unusual values that differ significantly from the rest of the dataset. As outliers can affect the overall distribution of data, outliers may need investigation and must be treated carefully to enhance the model performance. Likewise, it is also important to remove or replace any invalid or missing values with some estimated values prior to modelling. In this study, the interquartile range method (IQR) is employed to pre-process the outliers and invalid values are removed from the dataset [40]. However, for the missing values data is pre-processed by grouping them into day, month, and hour. Missing values are then filled in by taking an average of the available concentration values on the same month, day, and hour across all years of the dataset. Following this approach, a greater spread of values is reached for the missing data. In addition to already existing features in the dataset, the lagged feature is also created by taking the pollutant's previous hour concentration into account alongside splitting datetime index into day, month, and hour for additional features. Fig. 3 depicts the pre-processing workflow of the dataset. A sample of NO_2 is depicted in Fig. 4 before and after pre-processing data, representing the inclusion of missing values. After the pre-processing, additional features are generated as discussed in the section IV. Fig. 5 provide full data of NO_2 after pre-processing.

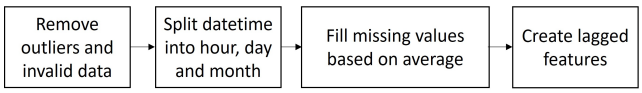

FIGURE 3. Pre-processing of dataset.

2) Two Stage Feature Selection

This study emphasis on the prediction of five major pollutants i.e. NO_2, O_3, SO_2, PM2.5 and PM10. The factors affecting these pollutants may differ or be identical in some cases. Interestingly, only a few of these predictors can be relied

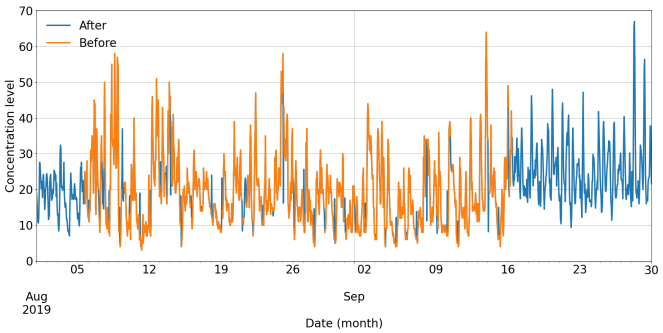

FIGURE 4. NO_2 data sample (over 2 months) representing addition of missing values.

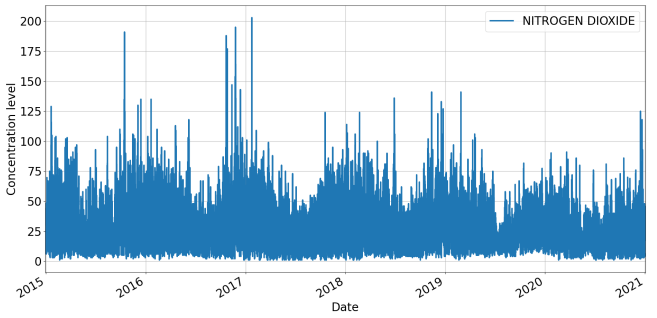

FIGURE 5. NO_2 data from 2015-2021 after addition of missing values.

upon, and those that are effective for one pollutant might not be for another. Since the effectiveness of the model depends on the inputs, it is imperative to identify optimal features for each of these pollutant predictions. Fig. 6 shows our proposed two stage feature selection approach to attain the optimum combination of features. In stage-1, several experiments are conducted with all possible combinations among all features (positively correlated only) categories to determine the best stage-1 combination and their effectiveness is analysed and evaluated using a simplified LSTM model based on RMSE, MAE, and R^2 scores. We kept the same model parameters during all the experiments to ensure that performance can be evaluated due to changes in all combinations of features. We are only reporting here the best stage-1 combinations for each target pollutant after all the experiments. The best stage-1 combination for NO_2 is made up of features from lag, temporal and meteorological categories. It achieves the highest R^2 score in comparison to other combinations and the least error in terms of MAE and RMSE.

In stage-2, the best combination of stage-1 is integrated with IMFs and residual from VMD method. We have performed experiments in seek of an optimum number of IMFs i.e. K for each pollutant. We have considered IMFs up to ten along with residual and the efficacy of each combination is evaluated using ML model (same model as in stage-1 for performance evaluation) using the performance indicators (i.e., R^2, MAE, and RMSE). In the experiments for each pollutant, we have combined the best stage-1 combination and increased the value of K to obtain performance indicators. The value of K is selected where model has obtained the optimum performance. Table. 5 provides summary of performance evaluation on the selection of the optimum value of K for NO_2. It indicates that when NO_2 is decomposed into three IMFs and integrated with the best stage-1 combination, it achieved the highest performance across all indicators and is therefore considered to be the optimum combination for NO_2. The effectiveness of our proposed approach is also examined by comparing how well it performed when using just only lag as a feature, the best stage-1 combination or features based on VMD. A performance comparison for NO_2 is shown in Fig. 7, where it can be observed that the optimum combination outperforms the other combinations based on lag, best stage-1 combination, and VMD features (i.e., $K = 3$). In terms of R^2, only lag or VMD features are not sufficient. However, the stage-1 combination improved the performance by 5% (w.r.t to lag feature performance), which can be further enhanced using optimum combination in stage-2 up to 86%.

A summary of optimum combinations for all pollutants taken into consideration is provided in Table. 6. This demonstrates exactly which stage-1 combination and IMF count work best for each pollutant. It can be observed that lag contributed to the prediction of all pollutants; aside from this meteorological feature is found to be effective for NO_2 and SO_2. Temporal features, on the other hand, are shown to be helpful for NO_2 and PM2.5, air pollutants for SO_2 and PM10, while the statistical feature is for SO_2. Furthermore, the best stage-1 combination for all pollutants is highlighted using dark green tint in Table. 4 and is also summarised in Table. 6. In addition, Table. 6, list the optimum IMF found for each pollutant to be combined with its the best stage-1 combination to produce optimum combination for each respective pollutant. The best stage-1 combination for NO_2 is based on lag in conjunction with temporal (day, month_sin, month_cos, hour) and meteorological (wind vertical) features. In contrast, just lag functioned for O_3. For SO_2, the best combination includes lag along with meteorological (wind horizontal and vertical), statistical (mean of the preceding two hours), and air pollutants (NO_2, PM2.5, PM10, NO, NO_X, and CO). For PM2.5, combination of lag and temporal (day, month_sin, month_cos, hour, hour_cos) is found best stage-1. Lastly, for PM10, lag and air pollutant (NO_2, SO_2, PM2.5, NO, NO_X, and CO) proved to be the best of stage-1 combination. From the foregoing insight, it is evident that all features except lag require careful selection, and that the value of K may vary according to the pollutant being anticipated.

3) Model Parameters and Tuning

Prior to training the model, the dataset is split into training, validation, and testing sets with ratios of 70%, 20%, and 10%, respectively. In each split, the indices are kept higher than the previous set, which will avoid shuffling (i.e., inappropriate in time series). The input features are normalised using Min-

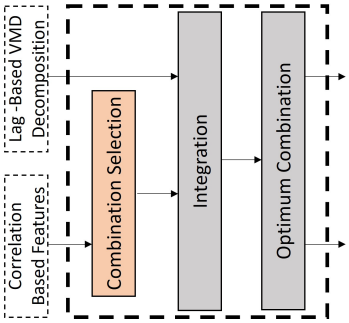

FIGURE 6. Two Stage Feature Selection.

TABLE 5. Selection of optimum number of IMFs for VMD decomposition (NO_2).

K	RMSE	R^2	MAE
2	6.0358	0.8253	4.1309
3	**5.4806**	**0.8559**	**3.6894**
4	5.7264	0.8427	3.9261
5	5.7548	0.8411	3.9875
6	5.4917	0.8553	3.7140
7	5.7630	0.8407	3.9223
8	5.8206	0.8375	3.9869
9	5.6037	0.8494	3.8382
10	5.6535	0.8467	3.9226

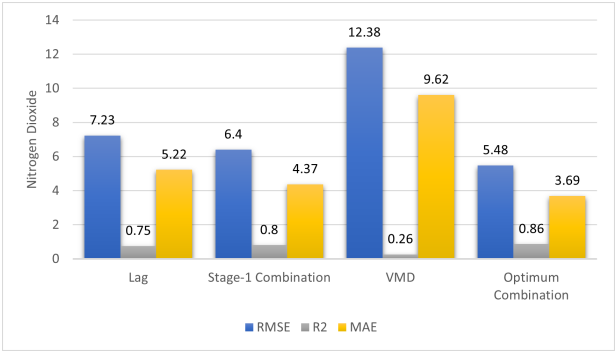

FIGURE 7. Comparison of different combinations to produce optimum combination for NO_2.

TABLE 6. Summary of Stage-1 combinations and IMFs to produce optimum combinations.

Pollutants	Stage-1 Combination	K
NO_2	Lag + Meteorological + Temporal	3
O_3	Lag	4
SO_2	Lag + Meteorological + Statistical + Air Pollutant	4
PM2.5	Lag + Temporal	4
PM10	Lag + Air Pollutant	3

Max normalisation and is defined as:

$$z_{norm} = \frac{z - z_{min}}{z_{max} - z_{min}}, \tag{6}$$

where z_{min} and z_{max} are the minimum and maximum values.

In this work, we are considering a simplified LSTM forecasting model as shown in Fig. 8. The input layer passes features to the model and we have used a LSTM layer with 25 cells, followed by a dropout layer which randomly drops out the number of cells to handle overfitting with the rate of 0.1. A fully connected dense layer with a linear activation function is used to produce an output. Adam optimiser is used during the training of the model and the optimal parameters of the simplified LSTM model is found after several trials to achieve better prediction accuracy on the given training dataset. The summary of the parameters with architectural details is given in Table. 7.

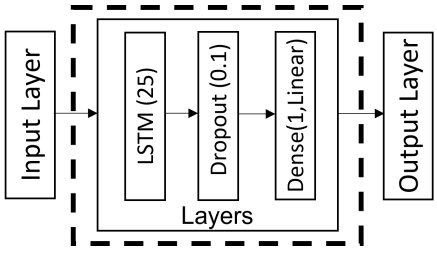

FIGURE 8. Architecture of simplified LSTM model.

TABLE 7. Summary of model parameters.

Parameters	Value
No. of layer	1
No. of cells in each layer	25
Dropout layer	0.1
Dense layer	1, Linear
Optimiser	Adam

In recent years, LSTM has been effectively used in various fields for predicting time series data such as energy demand [41], economics [42], wireless communications [43], and road safety [44]. A functional diagram of the LSTM cell is illustrated in Fig. 9 which is composed of three gates: a) input gate, b) forget gate and c) output gate [45]. The previous cell state $d(t-1)$ can influence the current state of the cell $d(t)$ and the amount of influence is controlled by the forget gate output $e(t)$. Similarly, the amount of influence by the new information $s(t)$ on $d(t)$ is managed by the output of input gate $l(t)$. The final output $g(t)$ of the cell is produced by combining $d(t)$, $s(t)$ and the past hidden state of the cell $g(t-1)$. Eq. (7)-(12) provide a mathematical representation of a LSTM cell as follows:

$$e(t) = \sigma(w_s^e s(t) + w_g^e g(t-1) + b_e), \tag{7}$$

$$l(t) = \sigma(w_s^l s(t) + w_g^l g(t-1) + b_l), \tag{8}$$

$$\tilde{d}(t) = tanh(w_s^{\tilde{d}} s(t) + w_g^{\tilde{d}} g(t-1) + b_{\tilde{d}}), \tag{9}$$

$$y(t) = \sigma(w_s^y s(t) + w_g^y g(t-1) + b_y), \tag{10}$$

$$d(t) = e(t)d(t-1) + l(t)\tilde{d}(t), \tag{11}$$

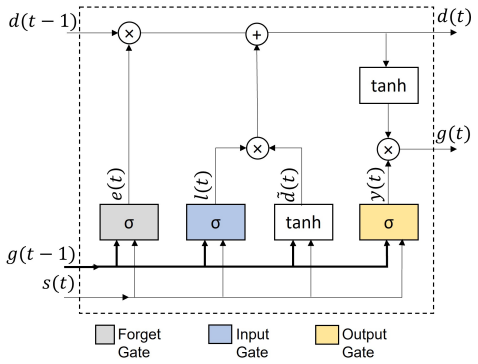

FIGURE 9. LSTM Cell.

$$g(t) = y(t)tanh(d(t)), \tag{12}$$

where b is the bias factor, w is the weight and activation functions are σ and $tanh$ to produce respective gate output.

4) Performance and Error Indicators

The efficacy of the ML forecasting model is assessed in this study using three statistical evaluation indicators namely R^2, MAE and RMSE and mathematically expressed in Eq. (13) , (14) and (15) as follows:

$$R^2 = 1 - \frac{\sum_{i=1}^{T}(q_i - \hat{q}_i)^2}{\sum_{i=1}^{T}(q_i - \bar{q})^2}, \tag{13}$$

$$MAE = \frac{1}{T}\sum_{i=1}^{T}|q_i - \hat{q}_i|, \tag{14}$$

$$RMSE = \sqrt{\frac{1}{T}\sum_{i=1}^{T}(q_i - \hat{q}_i)^2}, \tag{15}$$

where T is the total number of samples in test data, q_i, $\bar{q}$ and $\hat{q}_i$ are the target output at the i^{th} sample, mean derived from target output samples and predicted output at the i^{th} sample, respectively. Both MAE and RMSE are used to measure the prediction error of the forecasting model and indicates the extent to which model match target output in its predictions. Meanwhile, R^2 is another standard statistical indicator used to represent the goodness fit of forecasting model. Generally, models with higher R^2 score (nearly 1) and lower MAE and RMSE values indicates better prediction performance.

VI. RESULTS AND DISCUSSION

This section includes findings along with the related discussions over experiments on the proposed two stage feature engineering and selection method using a simplified forecasting model for the considered pollutants. The effectiveness of the proposed approach is examined and verified by making a comparison of the optimum features generated using the proposed approach with the best stage-1 combination of features and lag of pollutant being predicted. To improve clarity and better understanding, we are showcasing the forecasting model prediction from the testing data spanning only a week. Using test data for NO_2, Fig. 10 illustrates the forecasting model performance over a week when using optimum features as an input to the model.

The results show that for NO_2, our proposed optimum combination via two stage feature engineering and selection method outperform over all performance indicators in comparison to stage-1 combination and lag. In two stage feature selection method, the optimum combination of features for NO_2 is based on the lag of the pollutant being predicted along with the added benefits of the best stage-1 combination of features which includes meteorological and temporal features as well as VMD features ($K = 3$). After experimenting with several feature combinations across the four types depicted in Table. 4, the best combination for stage-1 is chosen. However, the optimum combination significantly improved the performance by 11% compared to the 5% improvement attained by stage-1 combination with respect to lag (using target pollutant) in terms of R^2. Thereby, achieving the highest R^2 score of 86% in comparison to 80% and 75% attained by stage-1 combination and lag respectively. In addition, the RMSE and MAE evaluation scores attained by optimum combination indicate the least error values comparatively others. Under the RMSE indicator, the evaluation score is dropped by 0.92 and 1.75 compared to using stage-1 combination and lag respectively. However, in terms of MAE, the error is reduced by 0.68 and 1.53 with respect to stage-1 combination and lag. Fig. 15, 16, 17 illustrates comparison of the proposed methodology for NO_2.

The forecasting model predictions over testing data for O_3, SO_2, PM2.5, and PM10 are shown in Fig. 11, 12, 13, 14, respectively. In case of O_3, our proposed method yet performed best under all evaluation indicators. We experienced O_3 a singular case wherein no combination of features was found helpful in improving the prediction in comparison to lag. Since the accuracy achieved from all feature combinations of stage-1 was equal to lag, we only integrated lag with 4 IMF from VMD decomposition to get the optimum combination, leaving stage-1 null. This resulted in 87% accuracy in terms of R^2, while the error scores for RMSE and MAE were 6.85 and 4.88, respectively. For SO_2 time series data, the recommended methodology attained 69% accuracy in terms of R^2 with corresponding RMSE and MAE error values of 0.51 and 0.36. The optimum combination includes lag, meteorological, statistical, and air pollutants features in addition to 4 IMFs. These sub-series incorporated lag, meteorological, statistical, and air pollutant features to determine the optimal combination. Similarly, for PM2.5 and PM10, our proposed method consistently performed better than stage-1 combination and lag and resulted in securing 86% and 76% accuracy respectively, in terms of R^2. Furthermore, error score is found 2.19 and 1.17 for PM2.5 and 4.88 and 2.38 for PM10 under the RMSE and MAE assessment indicators, respectively. The summary of the performance comparison of the single-step

forecasting model for all pollutants obtained by different feature combinations methods in terms of RMSE, R^2 and MAE is presented in Fig. 15, 16, 17, respectively.

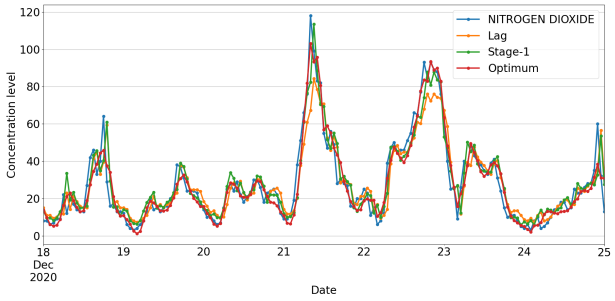

FIGURE 10. Comparison between actual and predicted data of NO_2 over a week.

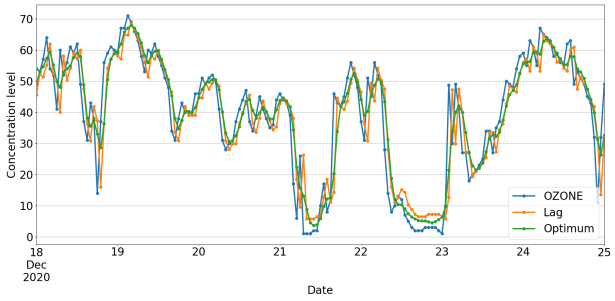

FIGURE 11. Comparison between actual and predicted data of O_3 over a week.

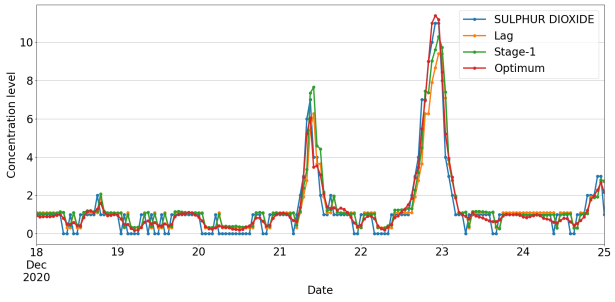

FIGURE 12. Comparison between actual and predicted data of SO_2 over a week.

Table. 8 summarises the proposed approach so that its effectiveness can be evaluated. In case of NO_2, using an optimal combination, the forecasting model achieved 11% more improvement with respect to R^2, in comparison to 5% using stage-1 combination. However, For O_3, a 3% improvement resulted from the usage of the optimum combination of features. Furthermore, among all the pollutants, SO_2 attained the maximum performance improvement using optimal combination which is 13% more than 2% gain by stage-1 combination of features. Lastly, for PM2.5 and PM10, the stage-1 combination could only enhance the performance by 1%, whereas the optimal combination of features made

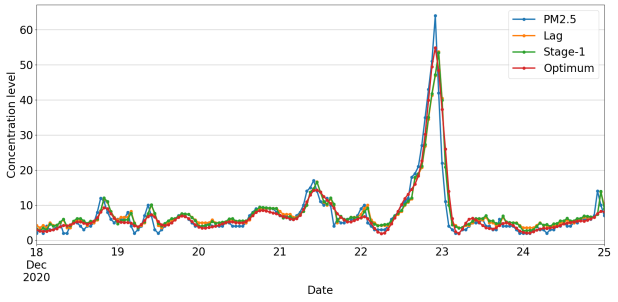

FIGURE 13. Comparison between actual and predicted data of PM2.5 over a week.

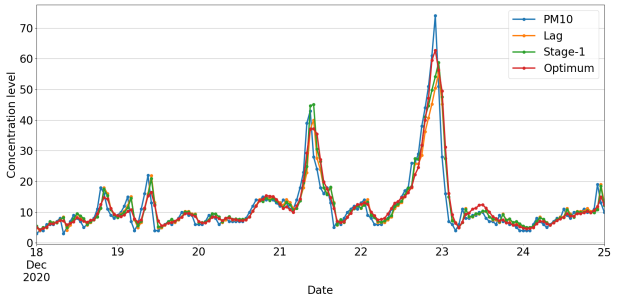

FIGURE 14. Comparison between actual and predicted data of PM10 over a week.

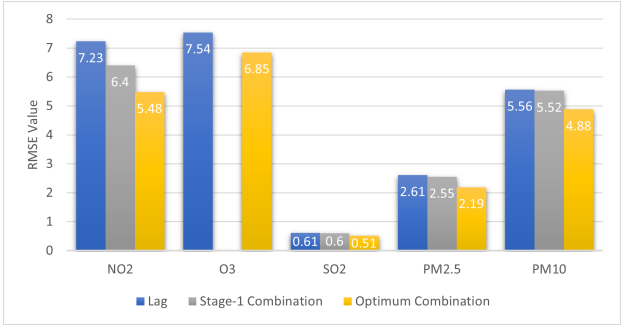

FIGURE 15. Comparison of features in terms of RMSE.

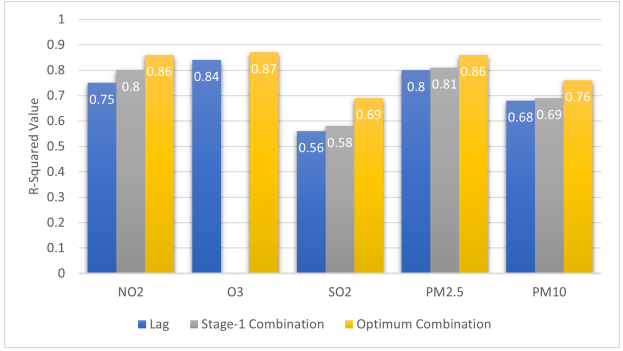

FIGURE 16. Comparison of features in terms of R^2.

6% and 8% more improvement respectively. To summarise the findings based on evaluation indicators, it can be easily

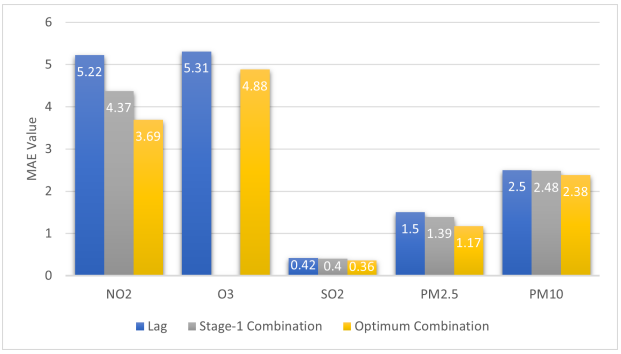

FIGURE 17. Comparison of features in terms of MAE.

concluded that performance attained by the proposed two stage feature engineering and selection approach is consistently improved along with the lowest error scores for all the pollutants in comparison to stage-1 combination and lag.

TABLE 8. Overall performance improvement based on proposed two stage feature engineering and selection w.r.t R^2.

Pollutants	Improvement (%)	
	Stage-1	Optimum
NO_2	5	11
O_3	-	3
SO_2	2	13
PM2.5	1	6
PM10	1	8

VII. CONCLUSION

Feature engineering is a fundamental step towards effective modelling, particularly in the domain of time series prediction and has a substantial effect on the performance of the model. This study provides a comprehensive investigation of the effectiveness of the proposed two stage feature engineering and selection inspired by their correlation and VMD approach for accurate prediction of 5 major air pollutants, which are beneficial in assessing air quality. Given the fact that there is no standard set of known features for a specific pollutant prediction. Optimum feature combinations may work differently for different pollutants and require customisation. In this work, we created new features and categorised them all into four major types (meteorological, temporal, statistical, and air pollutants) and generated 22 features in total. For stage-1, positively correlated features are selected and it is found that different pollutants require different feature combinations and such features can improve model performance by 1-5% compared to lag-based prediction. Moreover, performance is further enhanced by integrating stage-1 features with features of VMD (only for the optimum value of K) to form the optimum feature for a respective pollutant. It is observed that such an optimum combination can bring an overall performance improvement of 3 to 13%. Our findings through results demonstrated that with the optimum selection of features, a simplified forecasting model is sufficient and has shown significant improvement in terms of RMSE, MAE, and R^2 scores.

The demonstrated two stage approach can play a critical and important role in the urban planning such as traffic management, establishment of new industrial or residential areas and public health such as disease management, hotspot identification and reliable forecasting can leads to evidence based decision and policy making. However, further investigation is required to develop better integrated approach where new feature engineering approaches can be developed to improve performance over the longer time horizon such as over next 24 hours or even longer. Another possible direction could be to investigate hybrid decomposition approach by taking benefit of different signal decomposition methods.

ACKNOWLEDGMENT

The work of Trung Q. Duong was supported in part by the Canada Excellence Research Chair program. The work of Muhammad Fahim, Tuan-Vu Cao, and Trung Q. Duong was supported in part by UKRI and European Commission under MISO Project, "Autonomous Multi-Format In-Situ Observation Platform for Atmospheric Carbon Dioxide and Methane Monitoring in Permafrost and Wetlands." The work of Ruth Hunter and Trung Q. Duong was supported by the SPACE (Supportive Environments for Physical and Social Activity for Cognitive Health) Project (https://www.qub.ac.uk/sites/space/) funded by the ESRC Healthy Ageing Challenge under Grant ES/V016075/1.

REFERENCES

[1] WHO, "WHO global air quality guidelines: particulate matter (PM2.5 and PM10), ozone, nitrogen dioxide, sulfur dioxide and carbon monoxide," accessed: Nov. 14, 2023. [Online]. Available: https://www.who.int/publications/i/item/9789240034228

[2] F. H. Dominski, J. H. Lorenzetti Branco, G. Buonanno, L. Stabile, M. Gameiro da Silva, and A. Andrade, "Effects of air pollution on health: A mapping review of systematic reviews and meta-analyses," *Environmental Research*, vol. 201, p. 111487, 2021.

[3] WHO, "Ambient (outdoor) air pollution," accessed: Nov. 14, 2023. [Online]. Available: https://www.who.int/news-room/fact-sheets/detail/ambient-(outdoor)-air-quality-and-health

[4] WHO., "Air quality and health," accessed: Nov. 14, 2023. [Online]. Available: https://www.who.int/teams/environment-climate-change-and-health/air-quality-and-health/policy-progress/sustainable-development-goals-air-pollution

[5] DAERA, "Air pollution and smoke controll in Northern Ireland," accessed: Nov. 14, 2023. [Online]. Available: https://www.daera-ni.gov.uk/articles/air-pollution

[6] R. Thompson, R. B. Smith, Y. B. Karim, C. Shen, K. Drummond, C. Teng, and M. B. Toledano, "Air pollution and human cognition: A systematic review and meta-analysis," *Science of The Total Environment*, vol. 859, p. 160234, 2023.

[7] COMEAP, "Medical effects of air pollutants," accessed: Dec. 10, 2023. [Online]. Available: https://committees.parliament.uk/writtenevidence/121530/pdf/

[8] X. Su, L. Wang, M. Cao, L. Yang, M. Zhang, W. Qin, Q. Cao, Y. Yang, and L. Li, "Fengyun 4A land aerosol retrieval: Algorithm development, validation, and comparison with other datasets," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 61, pp. 1–16, 2023.

[9] X. Su, L. Wang, X. Gui, L. Yang, L. Li, M. Zhang, W. Qin, M. Tao, S. Wang, and L. Wang, "Retrieval of total and fine mode aerosol optical depth by an improved MODIS Dark Target algorithm," *Environment International*, vol. 166, p. 107343, 2022.

[10] X. Su, L. Wang, M. Zhang, W. Qin, and M. Bilal, "A high-precision aerosol retrieval algorithm (HiPARA) for Advanced Himawari Imager (AHI) data: Development and verification," *Remote Sensing of Environment*, vol. 253, p. 112221, 2021.

[11] X. Deng, Q. Cao, L. Wang, W. Wang, S. Wang, S. Wang, and L. Wang, "Characterizing urban densification and quantifying its effects on urban thermal environments and human thermal comfort," *Landscape and Urban Planning*, vol. 237, p. 104803, 2023.

[12] M. Cao, M. Zhang, X. Su, and L. Wang, "A two-stage machine learning algorithm for retrieving multiple aerosol properties over land: Development and validation," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 61, pp. 1–17, 2023.

[13] U. Nations, "COP26: Together for our planet," accessed: Nov. 14, 2023. [Online]. Available: https://www.un.org/en/climatechange/cop26

[14] N. I. Air, "Air quality in Northern Ireland," accessed: Nov. 14, 2023. [Online]. Available: https://www.airqualityni.co.uk/air-quality

[15] ULEZ, "The ultra low emission zone (ULEZ) for London," accessed: Nov. 14, 2023. [Online]. Available: https://www.london.gov.uk/programmes-strategies/environment-and-climate-change/pollution-and-air-quality/ultra-low-emission-zone-ulez-london

[16] Y. Zhang, Y. Wang, M. Gao, Q. Ma, J. Zhao, R. Zhang, Q. Wang, and L. Huang, "A predictive data feature exploration-based air quality prediction approach," *IEEE Access*, vol. 7, pp. 30 732–30 743, 2019.

[17] Q. Tao, F. Liu, Y. Li, and D. Sidorov, "Air pollution forecasting using a deep learning model based on 1D convnets and bidirectional GRU," *IEEE Access*, vol. 7, pp. 76 690–76 698, 2019.

[18] M. H. Nguyen, P. Le Nguyen, K. Nguyen, V. A. Le, T.-H. Nguyen, and Y. Ji, "PM2.5 prediction using genetic algorithm-based feature selection and encoder-decoder model," *IEEE Access*, vol. 9, pp. 57 338–57 350, 2021.

[19] X. Xu and M. Yoneda, "Multitask air-quality prediction based on LSTM-Autoencoder model," *IEEE Transactions on Cybernetics*, vol. 51, no. 5, pp. 2577–2586, 2021.

[20] S. Du, T. Li, Y. Yang, and S.-J. Horng, "Deep air quality forecasting using hybrid deep learning framework," *IEEE Transactions on Knowledge and Data Engineering*, vol. 33, no. 6, pp. 2412–2424, 2021.

[21] M. Mehrabi, M. Scaioni, and M. Previtali, "Forecasting air quality in Kiev during 2022 military conflict using sentinel 5p and optimized machine learning," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 61, pp. 1–10, 2023.

[22] H. Chen, M. Guan, and H. Li, "Air quality prediction based on integrated dual LSTM model," *IEEE Access*, vol. 9, pp. 93 285–93 297, 2021.

[23] S. Feng and Y. Feng, "A dual-staged attention based conversion-gated long short term memory for multivariable time series prediction," *IEEE Access*, vol. 10, pp. 368–379, 2022.

[24] H. Wang and H. Chen, "A novel particulate matter 2.5 concentration prediction model based on double-layer decomposition and feedback of model learning effect," *IEEE Access*, vol. 10, pp. 12 164–12 178, 2022.

[25] Z. Zhang, Y. Zeng, and K. Yan, "A hybrid deep learning technology for PM2.5 air quality forecasting," *Environmental Science and Pollution Research*, vol. 28, pp. 39 409–39 422, 8 2021.

[26] H. Liu and X. Zhang, "Aqi time series prediction based on a hybrid data decomposition and echo state networks," *Environmental Science and Pollution Research*, vol. 28, pp. 51 160–51 182, 10 2021.

[27] K. Wang, X. Fan, X. Yang, and Z. Zhou, "An AQI decomposition ensemble model based on SSA-LSTM using improved AMSSA-VMD decomposition reconstruction technique," *Environmental Research*, vol. 232, p. 116365, 2023.

[28] Q. Wu and H. Lin., "Daily urban air quality index forecasting based on variational mode decomposition, sample entropy and LSTM neural network," *Sustainable Cities and Society*, vol. 50, p. 101657, 2019.

[29] Q. Wu and H. Lin, "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, 2019.

[30] Q. Shao, J. Chen, and T. Jiang, "A novel coupled optimization prediction model for air quality," *IEEE Access*, vol. 11, pp. 69 667–69 685, 2023.

[31] N. I. Air, "Download air quality data - Northern Ireland," accessed: Dec. 1, 2022. [Online]. Available: https://www.airqualityni.co.uk/data

[32] F. Naz, C. Mccann, M. Fahim, T.-V. Cao, R. Hunter, N. T. Viet, L. D. Nguyen, and T. Q. Duong, "Comparative analysis of deep learning and statistical models for air pollutants prediction in urban areas," *IEEE Access*, vol. 11, pp. 64 016–64 025, 2023.

[33] N. Sarkar, R. Gupta, P. K. Keserwani, and M. C. Govil, "Air quality index prediction using an effective hybrid deep learning model," *Environmental Pollution*, vol. 315, p. 120404, 2022.

[34] L. Babu Saheer, A. Bhasy, M. Maktabdar, and J. Zarrin, "Data-driven framework for understanding and predicting air quality in urban areas," *Frontiers in Big Data*, vol. 5, 2022.

[35] J. Ma, Z. Li, J. C. Cheng, Y. Ding, C. Lin, and Z. Xu, "Air quality prediction at new stations using spatially transferred bi-directional long short-term memory network," *Science of The Total Environment*, vol. 705, p. 135771, 2020.

[36] K. Dragomiretskiy and D. Zosso, "Variational mode decomposition," *IEEE Transactions on Signal Processing*, vol. 62, no. 3, pp. 531–544, 2014.

[37] W. Fu, D. Yi, Z. Huang, C. Huang, Y. Geng, and X. Li, "Multiple event recognition scheme using variational mode decomposition-based hybrid feature extraction in fiber optic DAS system," *IEEE Sensors Journal*, pp. 1–1, 2023.

[38] H. Zhuo, X. Wu, Q. Zhong, and H. Zhang, "Position-free breath detection during sleep via commodity WiFi," *IEEE Sensors Journal*, vol. 23, no. 20, pp. 24 874–24 884, 2023.

[39] Y. Cao, Y. Sun, P. Li, and S. Su, "Vibration-based fault diagnosis for railway point machines using multi-domain features, ensemble feature selection and SVM," *IEEE Transactions on Vehicular Technology*, pp. 1–9, 2023.

[40] F. M. Dekking, C. Kraaikamp, H. P. Lopuhaä, and L. E. Meester, *A Modern Introduction to Probability and Statistics: Understanding why and how*. Springer, 2005, vol. 488.

[41] J.-Y. Kim and S.-B. Cho, "Predicting residential energy consumption by explainable deep learning with long-term and short-term latent variables," *Cybernetics and Systems*, vol. 54, no. 3, pp. 270–285, 2023.

[42] S. Park and J.-S. Yang, "Interpretable deep learning LSTM model for intelligent economic decision-making," *Knowledge-Based Systems*, vol. 248, p. 108907, 2022.

[43] B. Tian, G. Wang, Z. Xu, Y. Zhang, and X. Zhao, "Communication delay compensation for string stability of CACC system using LSTM prediction," *Vehicular Communications*, vol. 29, p. 100333, 2021.

[44] M. K. Nesrine Kadri, Ameni Ellouze and S. H. Turki, "New LSTM deep learning algorithm for driving behavior classification," *Cybernetics and Systems*, vol. 54, no. 4, pp. 387–405, 2023.

[45] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Computation*, vol. 9, pp. 1735–1780, 11 1997.

FAREENA NAZ received her B.S. degree in computer science from The University of Agriculture, Pakistan, in 2008, and the M.S. degree in computer science from COMSATS University, Islamabad, Pakistan, in 2010. She is currently a Ph.D. student at the Queen's University Belfast, UK, and part of the EU SPACE project. Her research interests include machine learning, data analytics, air pollution modelling, and natural language processing.

MUHAMMAD FAHIM received his B.S. with distinction from Gomal University, Pakistan in 2007. He got M.S. from the National University of Computer and Emerging Sciences (NUCES), Pakistan in 2009. He got his Ph.D. degree from Kyung Hee University, South Korea in February 2014. He worked as Post-Doctoral Fellow at the Department of Computer Engineering, Kyung Hee University, South Korea. He served as an Assistant Professor in the Department of Computer and Software Engineering, Faculty of Engineering and Natural Sciences, Istanbul Sabahattin Zaim University, Istanbul, Turkey for 3 years. He also led the Machine Learning research laboratory in Istanbul Sabahattin Zaim University. He worked as an Assistant Professor at the Institute of Information Systems, Innopolis University, Innopolis, Russia for four years. Currently, he is lecturer in School of Electronics, Electrical Engineering and Computer Science, Queen's University Belfast, UK. His research interests include machine learning models, wearable computing, digital signal processing, and behavior recognition in intelligent environments.

ADNAN AHMAD CHEEMA (Member, IEEE) received the B.Sc. degree from the COMSATS University, Pakistan., M.Sc. degree from the King's College London, U.K., and the Ph.D. degree from the Durham University, U.K., in 2006, 2008 and 2015, respectively. From 2015 to 2017, he was a postdoctoral Research Associate at Durham University and was involved in 5G channel measurements and modelling (sub-6 GHz and 24-90 GHz). In 2017, he joined Ulster University, U.K., as a Lecturer in electronics engineering and currently leading the Advanced Wireless Technologies Lab. His research interests include wireless communications, channel modelling, reconfigurable intelligent surface (RIS), non-orthogonal multiple access (NOMA), and machine learning.

NGUYEN TRUNG VIET obtained his PhD from Tohoku University, Japan in 2007. He is currently a Professor at Thuyloi University, Vietnam. He has strong expertise in nearshore hydromorphodynamics using both field measurements and numerical modeling, remote sensing video techniques. He was a project leader of numerous MOST projects on Nha Trang Coast (2013-2019) and Cua Dai Beach (2015-2018) in very close collaborations with IRD/EPOC/AFD-France, Japan and UK. He published more than 100 papers in journals and international conferences. He has been promoted to Vice President of Thuyloi University since 2014. He is appointed as a distinguished member of IAHR-APD and the executive member of Asian and Pacific Coasts (APAC) council and a member of the State Council for Professorship, Vietnam in the field of hydraulic engineering since 2019.

TUAN-VU CAO is a senior scientist at NILU Norwegian Institute for Air Research. He received PhD in electrical engineering from University of Oslo in 2012. He has a long experience in both academic and industry. From 2012 to 2018, he worked as Adjunct Associate Professor at University of Tromso (Norway), a postdoctoral fellow at Norwegian University of Science and Technology and senior engineers at WINS Instrumentation AS and Prediktor AS for various projects such as monitoring nutrient food content; wireless DST real time instrumentation; micropower sensor interface in nanometer CMOS Technology. He joined NILU from 2018. His research areas and interests are Enabling technologies and Autonomous Systems for Environmental monitoring and management. He is the PI of Horizon Europe project- MISO (2023-2026) – Autonomous Multi-Format In-Situ Observation Platform for Atmospheric Carbon Dioxide and Methane Monitoring in Permafrost & Wetlands; WP leader of NFR IKTPLUSS project AirQMan (2021-2025) - Low Latency Air Quality Management; co-PI of EEA grant project– HAPADS (2020-2023)-a novel air mobile monitoring system enables end-users to make information-driven decisions to mitigate air pollution exposure; key designer and technical coordinator for NFR 273394-"Leopard-Wearable particle detector enabling safer working environments" (2018-2020).

RUTH HUNTER is a Professor of Public Health and Planetary Health at the Centre for Public Health, Queen's University Belfast, and Director of the WHO Collaborating Centre for research and training in systems thinking and complexity science for NCD prevention and control. Her work primarily involves investigating how we can improve where we live (i.e. our built, social and natural environments) for better population health and planetary health. She is particularly interested in research at the intersection of public health and planetary health. Her work involves the application of systems-thinking and complexity science methods.

TRUNG Q. DUONG (Fellow) is a Canada Excellence Research Chair (CERC) and a Full Professor at Memorial University of Newfoundland, Canada. He is also an adjunct Chair Professor in Telecommunications at Queen's University Belfast, UK. He has received the two prestigious Research Chair of the Royal Academy of Engineering (2021-2025) and the Royal Academy of Engineering Research Fellowship (2015-2020). He was a Distinguished Advisory Professor at Inje University, South Korea (2017-2019). He is an Adjunct Professor and the Director of Institute for AI and Big Data at Duy Tan University, Vietnam (2012-present), a Distinguished Professor at Thuyloi University, Vietnam (2023-2028) and a Visiting Professor (under Eminent Scholar program) at Kyung Hee University, South Korea (2023-2024). His current research interests include quantum communications, wireless communications, signal processing, machine learning, and realtime optimisation.

Dr. Duong has served as an Editor/Guest Editor for the IEEE Transactions on Wireless Communications, IEEE Transactions on Communications, IEEE Transactions on Vehicular Technology, IEEE Communications Letters, IEEE Wireless Communications Letters, IEEE Wireless Communications, IEEE Communications Magazines, and IEEE Journal on Selected Areas in Communications. He received the Best Paper Award at the IEEE VTC-Spring 2013, IEEE ICC 2014, IEEE GLOBECOM 2016, 2019, 2022, IEEE DSP 2017, IWCMC 2019, 2023, and IEEE CAMAD 2023. He is the recipient of the prestigious Newton Prize 2017. He is a Fellow of Asia-Pacific Artificial Intelligence Association (AAIA).