# Features Inspired PM2.5 Prediction: A Belfast City Case Study

Fareena Naz[1], Muhammad Fahim[1], Adnan Ahmad Cheema[2], Nguyen Trung Viet[3], Tuan-Vu Cao[4], and Trung Q. Duong[1,5(✉)]

[1] School of Electronics, Electrical Engineering and Computer Science, Queen's University Belfast, Belfast BT7 1NN, UK
{fnaz01,m.fahim,trung.q.duong}@qub.ac.uk
[2] SenComm Research Lab, School of Engineering, Ulster University, Belfast BT15 1AP, UK
a.cheema@ulster.ac.uk
[3] Thuyloi University, Hanoi, Vietnam
nguyentrungviet@tlu.edu.vn
[4] Norwegian Institute for Air Research, Oslo, Norway
tvc@nilu.no
[5] Faculty of Engineering and Applied Science, Memorial University of Newfoundland, St. John's A1C 5S7, Canada
tduong@mun.ca

**Abstract.** Air pollution is one of the key challenges to both human health and our environment, and managing it requires collective systematic efforts to prevent and mitigate future effects. Fundamentally, this required a better understanding of sources that generate pollution and forecasting models to predict current and future air pollution levels. In this work, we investigated features inspired PM2.5 prediction based on a dataset collected in Northern Ireland, UK. We analysed the influence of different features available in the dataset and newly generated with approaches such as Variational Mode Decomposition (VMD) and evaluated single-step forecasting model performance. We found that a single Long Short Term Memory (LSTM) layer model with a small number of cells and integrated features are sufficient to achieve a good forecasting performance. The combination of VMD integrated features enabled the forecasting model to achieve $R^2$ score over 85% and achieve a gain of 6% when compared with lag based prediction only.

**Keywords:** Feature generation · Signal decomposition · PM2.5 · Machine learning · Forecasting models · Long Short Term Memory

## 1 Introduction

Air pollution is one of the major global environmental health challenges caused by the rapid rise in urbanisation and industrialisation. Around 99% of our global population breathes air that contains high levels of pollutants and leads to


N.-S. Vo et al. (Eds.): INISCOM 2024, LNICST 595, pp. 204–212, 2024.
https://doi.org/10.1007/978-3-031-67357-3_15

increased morbidity and mortality. Each year 6.7 million premature deaths are recorded worldwide, with low and middle income nations accounting for 95% of these deaths [1]. 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 [2]. 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. Both short and long term exposure to PM2.5 cause cardiovascular and respiratory diseases along with other ill health effects and mortality.

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. 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. 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 [3]. In addition to air pollutants, statistical, temporal, and meteorological features, following 24 h of weather prediction data is used as predictive data features to predict the PM2.5 concentration for the following 24 h. Based on the correlation of features, the performance of the model is compared with other models and findings revealed that their model outperformed others under indicators such as symmetric mean absolute percentage error (SMAPE), mean square error (MSE), and mean absolute error (MAE). In [4], proposed a short term forecasting hybrid approach combining convolutional neural network (CNN) and bidirectional gated recurrent unit (GRU) 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 are used. An encoder-decoder Long Short Term Memory (LSTM) model is proposed with Genetic algorithm (GA) feature selection to predict PM2.5 concentration using two datasets collected from Hanoi and Taiwan [5]. 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.

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 [6]. Initially, ensemble empirical mode decomposition (EEMD) is used for decomposing PM2.5 time series followed by sample entropy (SE) methodology, and then Variational Mode Decomposition (VMD) is employed where SE is higher than the average value. In another study, a VMD based BiLSTM model is proposed for single-step prediction of PM2.5 concentration in various cities of China [7]. In this work, BiLSTM is employed separately for all sub-series decomposed by VMD and concatenated

all at last to get the final prediction. In [8], 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 [9], SE is introduced to reduce the total number of Intrinsic Mode Functions (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. Although aforementioned studies have investigated different aspects of feature engineering, filtration approaches 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 features that can improve the performance of a forecasting model. The main contributions of this study include:

- We investigated different categories of features and influences on PM2.5. In addition, VMD decomposition is used to create new features based on the lag of PM2.5.
- We analysed features for target pollutant PM2.5 and show the performance improvement based on a limited number of LSTM cells in terms of root mean square error (RMSE) and R-squared ($R^2$).

The remainder of the paper is organised as follows: Sect. 2 describes the dataset and details on feature engineering, model training and testing are discussed in Sect. 3. Results and discussion are provided in Sect. 4 and finally, the paper is concluded in Sect. 5.

## 2 Dataset and Feature Generation

### 2.1 Dataset Description

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 [10,11]. 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 PM2.5, PM10, Nitrogen Dioxide ($NO_2$), Ozone ($O_3$), Sulphur Dioxide ($SO_2$), Nitric Oxide (NO), $NO_X$ and Carbon Monoxide (CO).

Statistical information such as total count, mean, standard deviation, minimum and maximum value of meteorological data and target pollutant is as follows: The total count for each meteorological parameter is 52564. The mean and standard deviation range 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 PM2.5 concentration has a mean of 9 with a standard deviation of 7.94 and the values vary from 0 to 104 with a total count of 52545.

### 2.2 Feature Generation

In this study, we grouped features into three types based on characteristics which include meteorological, temporal, and air pollutants. In meteorological features, we have considered temperature, wind horizontal, and wind vertical. In terms of temporal features, the datetime index contained in the dataset is utilised to create nine additional features. Initially, the datetime index is split into hour, day, and month features. Further to this, trigonometric functions are applied to them to create six additional features including month_sin, month_cos, day_sin, day_cos, hour_sin, and hour_cos. 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 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 21 features are introduced including 3 meteorological, 9 temporal, and 9 air pollutants to be used as an input to the model as listed in Table 1. All the features with positive Pearson correlation are selected in this work. Table 1 shows the correlation of the features, all positively correlated features are tinted blue (dark and light), while negatively correlated features are represented in grey tint. For instance, in case of PM2.5, all positively correlated features include air pollutants (except ozone) with a lag of target pollutant, month_sin, month_cos, hour_cos, hour, day 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, and $O_3$ are eliminated and not considered in the prediction of PM2.5.

In this work, we aim to use the 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. Figure 1 shows an example of lag PM2.5 decomposition using 4 IMFs and residual data, however careful consideration is required to select a number of IMFs so that such features can improve forecasting model performance.

## 3 Model Training and Testing

The workflow of model training and testing is shown in Fig. 2. The interquartile range method (IQR) is employed to pre-process the outliers and invalid values are removed from the dataset. 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 [11]. We have considered only positively correlated (i.e.

**Table 1.** Correlation of all features w.r.t target pollutant

| Type | Feature | PM2.5 |
|---|---|---|
| **Meteorological** | Temperature | -0.19 |
| | Wind Horizontal | -0.08 |
| | Wind Vertical | 0.25 |
| **Temporal** | Day | **0.02** |
| | Day_Sin | -0.03 |
| | Day_Cos | -0.03 |
| | Month | -0.07 |
| | Month_Sin | **0.17** |
| | Month_Cos | **0.15** |
| | Hour | **0.07** |
| | Hour_Sin | -0.06 |
| | Hour_Cos | **0.02** |
| **Air Pollutant** | $NO_2$ | 0.49 |
| | $O_3$ | -0.28 |
| | $SO_2$ | 0.38 |
| | PM2.5 | 1 |
| | PM10 | 0.62 |
| | Nitric Oxide | 0.43 |
| | Nitrogen Oxide | 0.49 |
| | Carbon Monoxide | 0.51 |
| | Lag (previous 1 hr) | **0.92** |

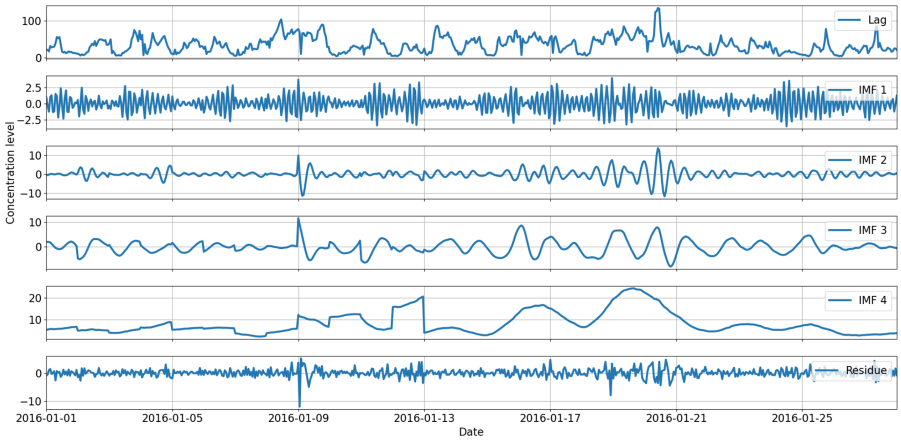


**Fig. 1.** VMD decomposition of PM2.5

Pearson correlation) features and evaluated the forecasting model performance over all categories based combinations and IMFs ranging from 2 to 10 with residual. We found that features like lag and temporal with an IMF value of four are suitable features to achieve superior performance.

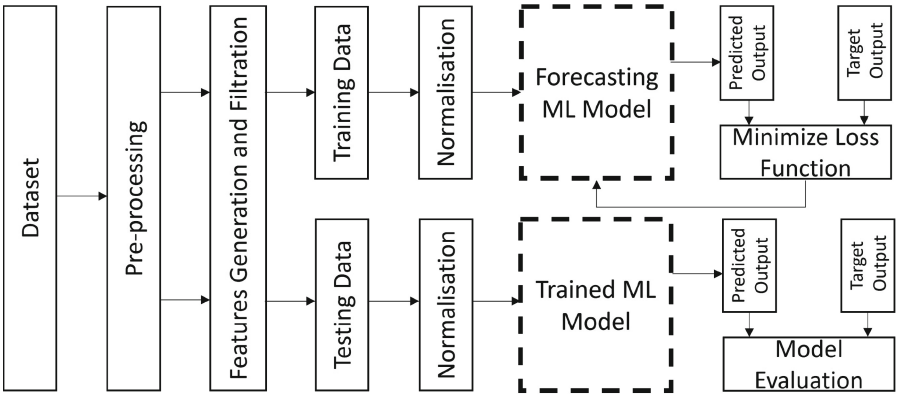


**Fig. 2.** Workflow of model training and testing

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). Each input feature $z$ is normalised using Min-Max normalisation and is defined as

$$z_{norm} = \frac{z - z_{min}}{z_{max} - z_{min}} \tag{3}$$

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

In this work, we are considering a simple LSTM forecasting model for single-step prediction. The input layer passes features to the model and we have used an 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 forecasting model are found after several trials to achieve better prediction accuracy on the given training dataset.

The efficacy of the ML forecasting model is assessed in this study using two statistical evaluation indicators namely $R^2$ and RMSE and mathematically expressed in Eq. (4) and (5) as follows:

$$R^2 = 1 - \frac{\sum_{i=1}^{T}(q_i - \hat{q}_i)^2}{\sum_{i=1}^{T}(q_i - \bar{q})^2} \tag{4}$$

$$RMSE = \sqrt{\frac{1}{T}\sum_{i=1}^{T}(q_i - \hat{q}_i)^2} \tag{5}$$

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. RMSE is used to measure the prediction error of the forecasting model and indicates the extent to which the model matches 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 a higher $R^2$ score (nearly 1) and lower RMSE value indicate better prediction performance.

## 4 Results and Discussion

In this section, we have evaluated performance of the single-step PM2.5 forecasting model based on a correlation-driven feature combination and then later integrated with VMD based features. As discussed in Sect. 3, we found features like lag and temporal (i.e. day, hour, month, hour_cos, month_cos, month_sin) provide better performance which can be further enhanced by integrating lag based four VMD features. The forecasting model predictions over testing data (a sample over 4 weeks only) for PM2.5 is shown in Fig. 3. The feature combination of lag, temporal, and IMFs improved the performance by 6% (86% in total) compared to the 1% improvement attained without VMD based features when compared to lag (using target pollutant) in terms of $R^2$. In addition, the RMSE evaluation score attained by VMD integrated features indicates the least error value compared to others. Under the RMSE indicator for VMD integrated features, the error is dropped by 0.42 and 0.36 compared to lag and without VMD features, respectively. Figure 4 illustrates a comparison of the feature combination for PM2.5 single-step forecasting based on performance gain (based on relative $R^2$), $R^2$ and RMSE.

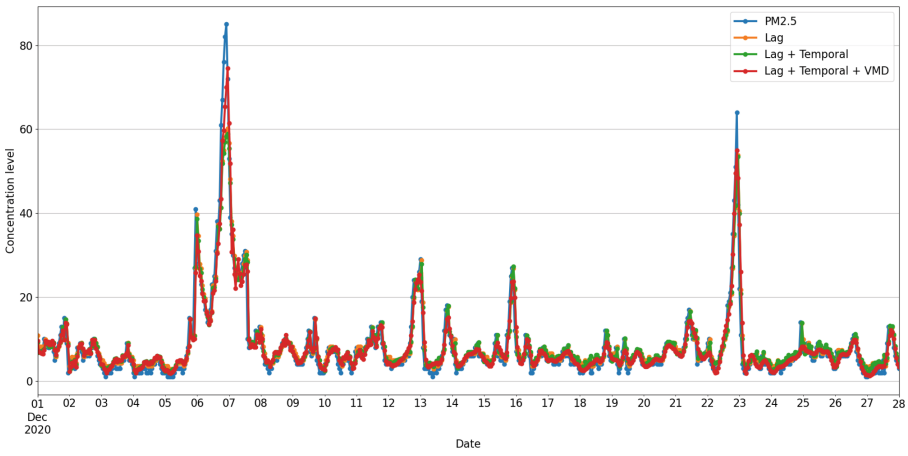


**Fig. 3.** Comparison between prediction and actual PM2.5 data over four weeks

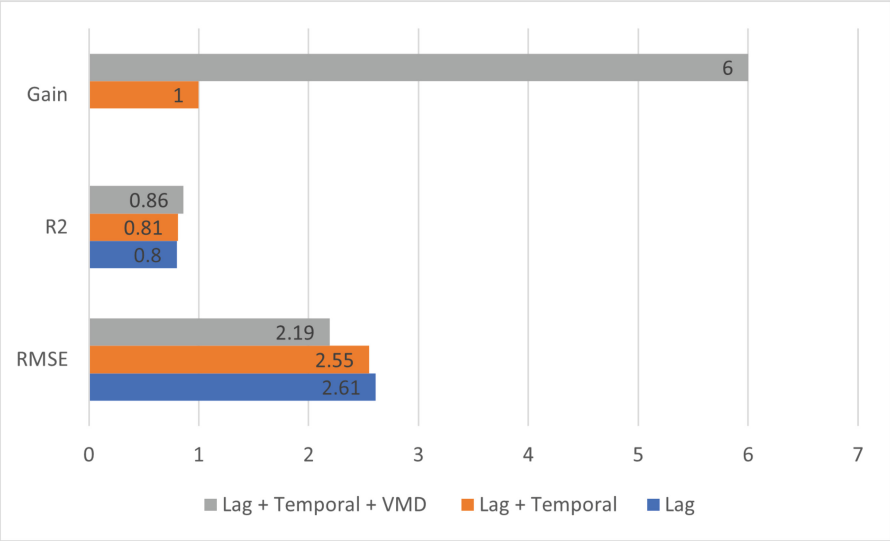


**Fig. 4.** Comparison of model prediction performance using evaluation indicators and overall gain

## 5 Conclusion

Features play an important role in forecasting models to learn better from data and achieve desired performance. This study investigates single-step forecasting model performance to predict PM2.5 and comprehensively analyses features role in enhancing model performance. We found that features like lag and temporal are catalysts to predict PM2.5. In addition to this, VMD features can be integrated to maximise forecasting model performance. We discovered that a single LSTM layer model can predict PM2.5 with a $R^2$ score of 86% and RMSE of 2.19. Moreover, our approach can obtain a gain of 6% and improve RMSE by 0.42 when compared with lag based prediction.

## References

1. WHO: Ambient (outdoor) air pollution. https://www.who.int/news-room/fact-sheets/detail/ambient-(outdoor)-air-quality-and-health. Accessed 14 Nov 2023
2. DAERA: Air pollution and smoke control in Northern Ireland. https://www.daera-ni.gov.uk/articles/air-pollution
3. Zhang, Y., et al.: A predictive data feature exploration-based air quality prediction approach. IEEE Access **7**, 30 732–30 743 (2019)
4. Tao, Q., Liu, F., Li, Y., Sidorov, D.: Air pollution forecasting using a deep learning model based on 1d convents and bidirectional GRU. IEEE Access **7**, 76 690–76 698 (2019)

5. Nguyen, M.H., Le Nguyen, P., Nguyen, K., Le, V.A., Nguyen, T.-H., Ji, Y.: Pm2.5 prediction using genetic algorithm-based feature selection and encoder-decoder model. IEEE Access **9**, 57 338–57 350 (2021)
6. Wang, H., Chen, H.: A novel particulate matter 2.5 concentration prediction model based on double-layer decomposition and feedback of model learning effect. IEEE Access **10**, 12 164–12 178 (2022)
7. Zhang, Z., Zeng, Y., Yan, K.: A hybrid deep learning technology for pm2.5 air quality forecasting. Environ. Sci. Pollut. Res. **28**, 39 409–39 422 (2021)
8. Wang, K., Fan, X., Yang, X., Zhou, Z.: An AQI decomposition ensemble model based on SSA-LSTM using improved AMSSA-VMD decomposition reconstruction technique. Environ. Res. **232**, 116365 (2023)
9. Wu, Q., Lin, H.: Daily urban air quality index forecasting based on variational mode decomposition, sample entropy, and LSTM neural network. Sustain. Urban Areas **50**, 101657 (2019)
10. N. I. Air: Download air quality data - Northern Ireland. https://www.airqualityni.co.uk/data. Accessed 1 Dec 2022
11. Naz, F., et al.: Comparative analysis of deep learning and statistical models for air pollutants prediction in urban areas. IEEE Access **11**, 64 016–64 025 (2023)