SPACE
SUPPORTIVE ENVIRONMENTS FOR
PHYSICAL & SOCIAL ACTIVITY,
HEALTHY AGEING & COGNITIVE HEALTH

Two Stage Feature Engineering to Predict Air pollutants in Urban Areas: A Belfast City Case Study

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

[1]Centre School of Electronics, Electrical Engineering and Computer Science, Queen's University Belfast, Belfast, NI, UK; [2]SenComm Research Lab, School of Engineering, Ulster University, Belfast, UK; [3] Thuyloi University, Hanoi, Vietnam; [4] Norwegian Institute for Air Research, Oslo, Norway; [5] 5Centre for Public Health, School of Medicine, Dentistry and Biomedical Sciences, Queen's University Belfast, UK; [1,6] Faculty of Engineering and Applied Science, Memorial University of Newfoundland, St. John's, Canada

***Contact information: fnaz01@qub.ac.uk**

BACKGROUND

- ❑ Air pollution is the global environmental health challenge.
- ❑ **99%** of global population breath air that contains high level of pollutants and is estimated to cause **6.7 million** premature deaths worldwide each year, with low- and middle-income nations accounting for **95%** of these deaths.
- ❑ UK govt. has set a goal to curtail **35% of air pollution by 2040**.
- ❑ 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.
- ❑ AIM: *To build **features** based **simplified** Machine Leaning prediction. Model.*

MACHINE LEARNING MODEL

- ❑ We proposed two stage feature selection method which is based on correlation and selection of an optimum number of intrinsic mode functions (IMFs) to achieve optimum performance using a simplified LSTM model.

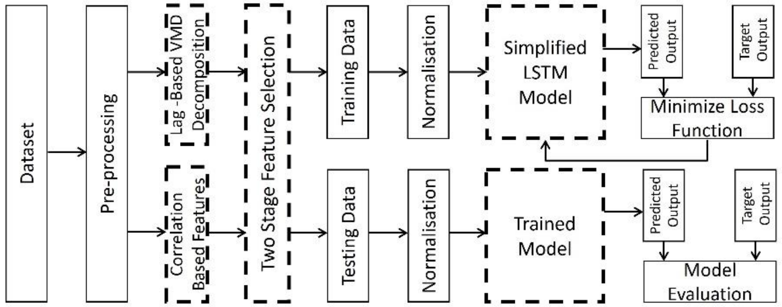

Figure 1. Workflow of model training and testing with two stage feature engineering and selection approach

METHODOLOGY AND RESULTS

- ❑ In this study, we explored 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 NO_2, O_3, SO_2, PM2.5 and PM10.
- ❑ In stage-1, using the dataset we created new features 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.
- ❑ We employed a simplified Long Short-Term Memory (LSTM) neural network and proposed a single-step forecasting model to predict multivariate time series data.

Table 1. Summary of Stage-1 combinations and IMFs to produce optimum combinations with respective Gains

Pollutants	**Stage-1 Combination**	***K***	***Stage-1 Gain***	***Optimum Gain***
NO_2	Lag + Meteorological + Temporal	3	5	11
O_3	Lag	4	-	3
SO_2	Lag + Meteorological + Statistical + Air Pollutant	4	2	13
PM2.5	Lag + Temporal	4	1	6
PM10	Lag + Air Pollutant	3	1	8

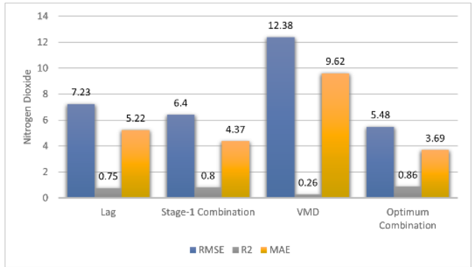

Figure 2. Comparison of different combinations (NO_2)

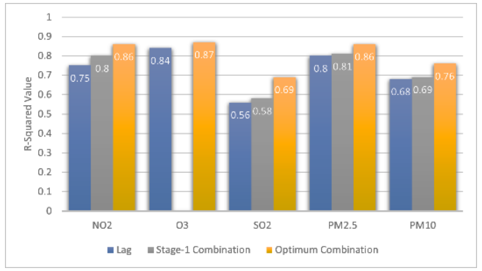

Figure 3. Proposed approach and evaluation based on R^2

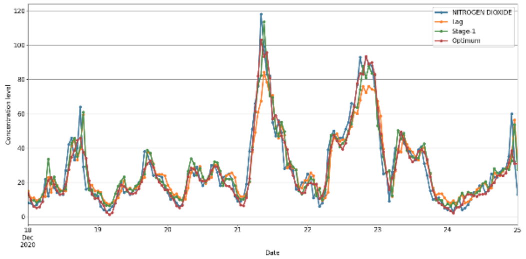

Figure 4. Actual and predicted data of NO_2 over a week

CONCLUSIONS

- ❑ 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.
- ❑ It is observed that such an optimum combination can bring an overall performance improvement up to 13%.

QUEEN'S UNIVERSITY BELFAST

qub.ac.uk/sites/space/
@spacequb

This work was supported by **UK Research and Innovation** [ES/V016075/1]

UKRI Economic and Social Research Council

UKRI Healthy Ageing Challenge Social, Behavioural and Design Research