A Meta-graph Deep Learning Framework for Forecasting Air Pollutants in Stockholm

Zhiguo Zhang*, Xiaoliang Ma*, Christer Johansson†, Junchen Jin*‡, Magnuz Engardt†

* ITS Lab, Department of Civil and Architectural Engineering, KTH Royal Institute of Technology,
Brinellvägen 23, 10044, Stockholm Sweden
†Environmental and Health Administration, City of Stockholm,
Fleminggatan 4, 10420, Stockholm
‡College of Electrical Engineering, Zhejiang University, Hangzhou, China
Corresponding author: X. Ma (email: liang@kth.se)

***Abstract*—Forecasting air pollution is an important activity for developing sustainable and smart cities. Generated by various sources, air pollutants distribute in the atmospheric environment due to the complex dispersion processes. The emerging sensor and data technologies have promoted the development of data-driven approaches to replace conventional physical models in urban air pollution forecasting. Nevertheless, it is still challenging to capture the intricate spatial and temporal patterns of air pollutant concentrations measured by heterogeneous sensors, especially for long-term prediction of the multi-variate time series data. This paper proposes a deep learning framework for longer-term forecast of air pollutants concentrations using air pollution sensing data, based on a conceptual framework of meta-graph deep learning. The key modules in the framework include meta-graph units and fusion layers, which are designed to learn temporal and spatial correlations respectively. A detailed case was formulated for forecasting air pollutants in Stockholm using air quality sensing data, meteorological data and so on. Experiments were conducted to evaluate the performance of the proposed modelling framework. The computational results show that it outperforms the baseline models and conventional deterministic dispersion models, demonstrating the potential of the framework to be deployed for the real air quality information systems in Stockholm.**

I. INTRODUCTION

According to the World Health Organisation (WHO), air pollution is one of the leading causes of human mortality worldwide [1]. Public information, regarding the expected health risk associated with air pollutants concentrations of the current day or future days, becomes very important to disclose in European countries, but the challenge is to accurately infer pollutant concentrations over a long-term horizon i.e. forecasting hourly for future 24 hours and even several days [2]. Spatial variability of air pollution concentration is influenced not only by pollutants from local areas but also due to regional transport [3]. The temporal fluctuation of air pollutant concentrations is due to its complex non-linear relationship with various exogenous factors, such as meteorological and road traffic [4].

This work is supported by the Richterska Stiftelsen (2019-00498) and KTH Centre for Digital Futures (iHorse 2021). The computations were enabled by the supercomputing resource Berzelius provided by National Supercomputer Centre at Linköping University and the Knut and Alice Wallenberg foundation.

Different approaches are established for modeling the variation of air pollutant concentrations, and can be categorized into deterministic dispersion models and data-driven statistical approaches. Deterministic models take multivariate inputs of topography and meteorology, and simulate the transport and dispersion of pollutants in the atmosphere[5]. Gaussian plume models are widely applied in urban areas for estimating impacts on atmospheric concentrations from different emission sources [6]. Eulerian chemical transport models describe the emission, transport, mixing, and chemical transformation of trace gases and aerosols, and are part of the Copernicus Atmosphere Monitoring Service (CAMS) model, which has been applied to predict air pollution in Europe [7]. The uncertainties of deterministic models come from uncertainties of inputs, the validity of model forms etc.

On the other hand, data-driven approaches provided a new way to carry out statistical prediction without explicitly representing the internal physical processes. Many recent studies have applied statistical machine learning (ML) methods to predict hourly or daily average pollutant concentrations using meteorological and other inputs e.g. [8]. In addition, integration of ML, dispersion modeling and satellite measurements have been used to capture the temporal and spatial distribution of pollutant concentrations [9]. Most of these ML models focused on air pollutant concentration data itself and modelling of a single air pollutant. More recently, deep learning has gained momentum in the application of air pollution forecasts. Recurrent Neural Networks (RNNs) were applied for prediction of air pollutants e.g. [10], [11]. CNN and Bi-LSTM [12] were used for extraction of the spatial-temporal correlation of the air pollutant $PM_{2.5}$. Meanwhile, attention mechanisms were also applied for air pollution forecast with RNN [13] and Bayesian RNN [14]. None of these models consider other exogenous factors such as meteorology, road traffic etc.

In parallel with the advancement of data-driven modeling methods, small and cheaper sensors become increasingly deployed for measuring air pollutant concentrations in large spatial areas due to the development of Internet of Things (IoT) technologies. They are promising to complement traditional precision sensing stations that can mainly capture the temporal variation of local pollutant concentration values in air quality

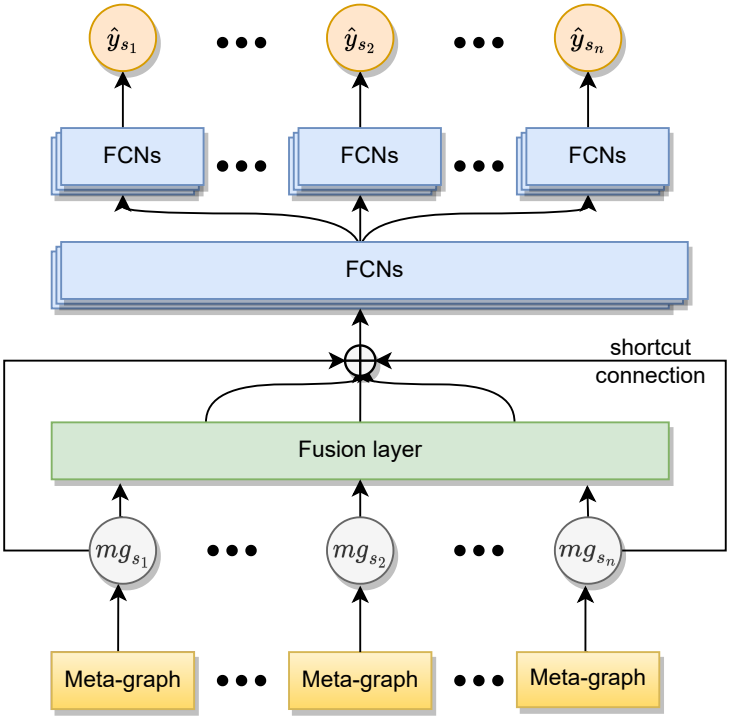

Fig. 1: Deep learning framework for the proposed approach.

surveillance system. Following the technical trend, emerging IoT air quality sensors have been evaluated for air pollutant measurement near motorways in Stockholm [15]. It is believed that such type of sensors will be increasing deployed for future smart cities. The main objective of this study is to cater for the trend, and develop a data-driven modeling framework that may predict air pollutant concentrations using data collected by heterogeneous types of sensors installed for a large spatial area. Meanwhile. the framework should be general enough to include not only air pollutant concentration values measured hourly but also exogenous factors such as meteorological condition.

II. METHODOLOGY

In this section, we introduce a novel deep-learning framework to model air pollutant concentrations based on sensor data from observation stations. The framework employs meta-graph units and fusion layers to effectively extract dynamic temporal and spatial correlations. The meta-graph units integrate diverse data sources to establish granular knowledge bases whereas the fusion layers aggregate the abstracted temporal features and capture station-to-station correlations.

A. Problem formulation

Air quality measurements are collected from air quality sensors at each monitoring station. $a_t^{s,p}$ represents the concentration value of an pollutant p (for NOx, $PM_{2.5}$, PM_{10} etc.) at station s and time t,

Meteorological forecasts are generated by existing weather forecasting systems. $\mathbf{m}_t^s$ denote meteorological forecasts for station s and time t, and it is a vector of temperature, wind speed, humidity, etc. *Deterministic forecasts* are predicted using deterministic models. $\mathbf{det}_t^{s,p}$ denote deterministic forecasts of an pollutant p at monitoring station s and time t. *Time features* refer to the temporal information associated with each data sample. $\mathbf{t}_t$ is a vector of tranformed timestamp information.

Additional features are also constructed through statistical analysis in the feature engineering process to expand the feature space and facilitate the targeted extraction of relevant features. *Aggregate engineering features* $\mathbf{e}_t^{s,p}$ are computed using the historical data of an air pollutant concentration p at station s and before time t, such as descriptive statistical features in a rolling horizon and autoregressive features based on correlation analysis.

Based on the notations above, the forecasting problem is formulated to train a deep learning model $\mathbf{F}$:

$$[\hat{\mathbf{a}}_{t+1:t+h}^{s_1}, \hat{\mathbf{a}}_{t+1:t+h}^{s_2}, \cdots, \hat{\mathbf{a}}_{t+1:t+h}^{s_s}] = \mathbf{F}(\mathbf{X}_t^{s_1}, \mathbf{X}_t^{s_2}, \cdots, \mathbf{X}_t^{s_s}) \quad (1)$$

where $\hat{\mathbf{a}}_{t+1:t+h}^{s}$ denotes a vector of predicted pollutant concentration values at station s and h is the length of prediction horizon. $\mathbf{X}_t = [\mathbf{x}_{t-d+1}, \mathbf{x}_{t-d+2}, \cdots, \mathbf{x}_t]^\top \in \mathbb{R}^{D \times N}$ is the input tensor of each station, D is the historical memory size of input tensor and N is the number of input features. $\mathbf{x}_t$ is the input data at time t, a concatenation of all input features including historical pollutant data, meteorological forecasts, deterministic forecasts, and other engineering features. In principle, the model can be general to have different features for different stations. But we simplify the notation here by assuming consistent feature space for each station.

B. Deep learning framework

The proposed deep learning framework is shown in Figure 1, and is developed to carry out multi-step ahead forecast for multivariate input data of air pollutant concentrations and exogenous factors. The general model can be considered as a combination of local meta-graph units and the fusion layer part. Based on the specified features of each monitoring station, the bottom meta-graph unit processes the input features, and represents the temporal properties in the input sequence. Furthermore, the abstracted temporal features are aggregated into the fusion layer, which incorporates station-to-station spatial correlations. In the end, the predicted values for all stations are output separately after downscaling through multiple fully connected networks and a fully connected network for each station.

1) Meta-graph units: The meta-graph unit is designed to extract local temporal context based on raw data. Considering the time-series nature of the air pollutants, fully connected layers and LSTM layers are used to extract temporal features and reduce the dimension for a specific monitoring station s. The detailed structure of the meta-graph unit is shown in Figure 2.

For the meta-graph unit for station s, the local model can be represented by

$$\mathbf{mg}_t^s = f(\mathbf{X}_t^s) = f([\mathbf{x}_{t-d+1}^s, \mathbf{x}_{t-d+2}^s, \cdots, \mathbf{x}_t^s]^\top) \tag{2}$$

where $\mathbf{x}_t = [a_{t-h+1}, \mathbf{m}_t^\top, \mathbf{det}_t^\top, \mathbf{t}_t^\top, \mathbf{e}_{t-h+1}^\top] \in \mathbb{R}^N$. It is worth of mentioning that the aggregate features were calculated using available measurement data before time $t-h$, whereas meteorological, deterministic and time features are predicted values at time t already available for training and inference.

2) Fusion layers: The fusion layer is designed to capture the spatial correlation among different monitoring stations using context information extracted from meta-graph units. A multi-head attention (MHA) layer [16] and residual connection are used to aggregate the spatial-temporal features among stations. The detailed structure of the fusion layer is described by

$$\mathbf{Z}_{att} = MHA(\mathbf{mg}_t^{s_1}, \mathbf{mg}_t^{s_2}, \cdots, \mathbf{mg}_t^{s_s}) \tag{3}$$
$$\mathbf{M}_t^{tot} = Concat(\mathbf{mg}_t^{s_1}, \mathbf{mg}_t^{s_2}, \cdots, \mathbf{mg}_t^{s_s}) \tag{4}$$
$$\mathbf{Z}_{fus} = LayerNorm(\mathbf{Z}_{att} + \mathbf{M}_t^{tot}) \tag{5}$$

where $\mathbf{Z}_{att} \in \mathbb{R}^{S\times H}$ is the output of the multi-head attention layer, $\mathbf{M}_t^{tot} \in \mathbb{R}^{S\times H}$ is the concatenation of all meta-graph units, $\mathbf{Z}_{fus} \in \mathbb{R}^{S\times H}$ is the output of the fusion layer, S is the number of monitoring stations, H is the number of hidden units in the meta-graph unit.

3) Prediction layers: A fully-connected network layer is used to reduce the dimension of the processed information generated by previous layers, and to output the corresponding predicted values for each monitoring station utilizing an individual fully-connected network layer for each station.

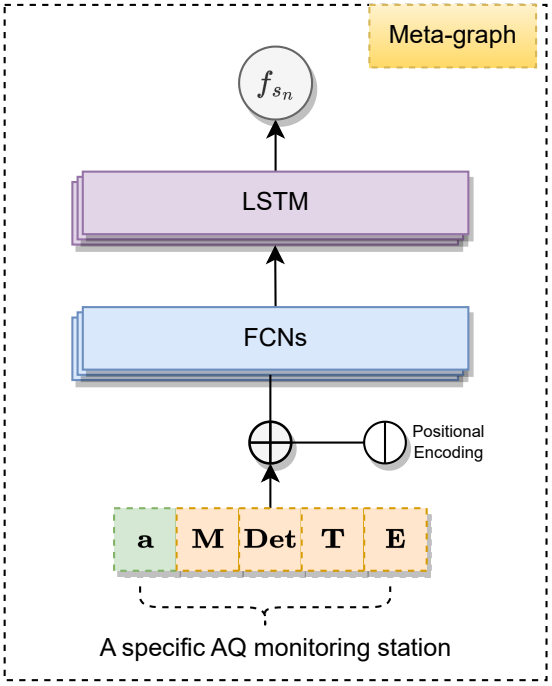

Fig. 2: The meta-graph unit $\mathbf{mg}^s$ is designed to model the local temporal relation at the station s.

The detailed structure of the prediction layers is analytically presented by

$$\mathbf{z}_{fc} = FC\big(flat(\mathbf{Z}_{fus})\big) \tag{6}$$
$$\hat{a}_{t+h}^s = FC^s(\mathbf{z}_{fc}) \qquad \forall s \in \{s_1, s_2, \cdots, s_s\} \tag{7}$$

where $\mathbf{z}_{fc} \in \mathbb{R}^{S\times H'}$ is the output of the fully-connected layer, which takes the flatten vector of $\mathbf{Z}_{fus}$ as input. H' is the number of hidden units in the fully-connected layer, $FC^s(\mathbf{z}_{fc}) \in \mathbb{R}$ is the output of the fully-connected layer for a specific monitoring station s.

III. Case Study

This section describes our experimental study for evaluating the proposed deep learning model. In the experiment case, the model was developed and trained for forecasting a single air pollutant, NOx concentration value per hour, for the future 24 hours at three urban air quality surveillance stations in Stockholm. Details of feature selection and baseline models are illustrated in the study of [2].

A. Data

1) AQ measurements: First, NOx concentrations(ug/m^3) are collected from three monitoring stations of urban street canyon sites in central Stockholm, including Hornsgatan(HG), Folkungagatan(FG) and Sveavägen(SV). They are all located in central Stockholm in Fig. 3.

The measurement data covers 500 days (from 5 August 2020 to 31 December 2021), with 12335 samples in total. All the data was collected at an hour intervals.

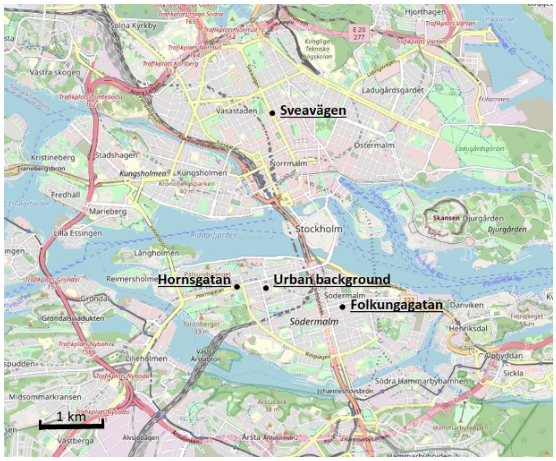

Fig. 3: The locations of three air pollutant monitoring stations in Stockholm. Base map credits: © OpenStreetMap contributors.

2) Meteorological forecasts: As an essential part of the Stockholm Air Quality system, meteorological data and forecasts are downloaded every day from the Swedish Meteorological and Hydrological Institute(SMHI). The meteorological forecasts extend over 10 days and are a combination of outputs from several regional and global numerical weather prediction models, including pressure, temperature, precipitation, cloudiness, wind speed, wind direction, relative humidity, boundary layer height and so on.

3) Deterministic forecasts: Considering emissions and dispersion at the scales of country, region and street, three different deterministic physical models are integrated to predict AQ concentrations, including the CAMS ensemble model, the Gaussian dispersion model and the Operational Street Pollution Model (OSPM). This prediction has been used as an essential part of the Stockholm Air Quality Information system at the beginning. The integrated results are published once a day, and the predicted values cover the future 3 days (72 hours).

4) Engineering features: The engineering features are the processed information from the raw data to serve the forecasting model. The features include mainly two categories:

- Autocorrelation features: 24-hour lagged air pollutant concentrations based on autocorrelation analysis.
- Aggregate statistical features: The mean, median, minimum and maximum values of the rolling 24 hour period.

B. Baseline Models

Three models, already implemented in the Stockholm Air Quality System [2], are applied in this study as baseline models for comparison purposes. These models are:

1) Random Forest: Random Forest [17] is an ensemble method for regression analysis that combines multiple independently trained trees and undertakes a majority vote of all these trees to determine the final prediction result.

TABLE I: Results of hyperparameter tuning

Hyperparameters	**Best**	**Search Range**
Lengths of input sequence	12	[12, 24, 36, 48, 60, 72]
Hidden size of LSTM	96	[32, 64, 96, 128, 160]
Batch size	96	[16, 32, 48, 64, 72, 96]
Learning rate	5e-4	[5e-5, 1e-4, 5e-4, 8e-4, 1e-3]
α of Loss function	0.9	[0.1, 0.3, 0.5, 0.7, 0.9]

2) XGBoost: XGBoost [18] is a sequential ensemble of tree models, which creates a shallow tree to correct the errors of the previous tree, resulting in many weak classifiers that are combined to form a strong model.

3) Long Short-Term Memory: Long Short-Term Memory (LSTM) is a type of recurrent neural network (RNN) architecture that is popular for many deep learning applications. A vanilla LSTM model, containing two hidden layers with 96 units for each layer, is applied in the study.

For all three baseline models, the prediction schemem is setup with longer term prediction horizon and same size of memory data [2]. These models were trained using the training data of three streets in the Stockholm city, and the predictions are made for the three streets separately.

C. Experimental Setup

1) Data Pre-processing: The measurement data with a missing rate of less than 5% and missing values are replaced with mean values of available data in the neighborhood according to the respective autocorrelation properties.

To preserve temporal properties, the data is divided into training, validation and test sets by the ratio of 6:2:2 along the time axis.

2) Module Selection: The meta-learner unit and fusion layers are employed as follows:

- The meta-learner unit: one fully connected layer, two LSTM layers(with the same number of hidden units), and one LayerNorm layer.
- The fusion layer: one multi-head attention layer, one LayerNorm layer and residual connection.

3) Model Tuning: Hyperparameter tuning is implemented on all models by grid search, where the vanilla LSTM model and our model use the same hyperparameter search space. The best parameters of our model are summarised in Table I.

A combined loss function is chosen to weight the sum of $L1$ norm and $L2$ norm according to the selected weighting parameter α to reduce the problem of conservative prediction results caused by using $L2$ norm alone.

$$\begin{aligned}\mathscr{L}(\mathbf{a}_{t+l}, \hat{\mathbf{a}}_{t+l}) = &\alpha \sum_{s=1}^{S} \sum_{p=1}^{P} \left| a_{t+l}^{s,p} - \hat{a}_{t+l}^{s,p} \right| + \\ &(1-\alpha) \sum_{s=1}^{S} \sum_{p=1}^{P} \left(a_{t+l}^{s,p} - \hat{a}_{t+l}^{s,p} \right)^2 \end{aligned} \tag{8}$$

In addition, the model is trained using the Adam optimizer and the strategy of learning rate decay is ReduceLROnPlateau,

TABLE II: Comparison of prediction performance

Model	MSE	MAE	RMSE
RandomForest	1082.726±2.241	19.870±0.011	32.905±0.034
XGBoost	1156.869±26.197	20.294±0.177	34.009±0.384
Vanilla LSTM	1020.068±33.515	18.239±0.273	31.931±0.520
Proposed model	**920.338**±56.050	**18.162**±1.003	**29.967**±0.898

where patience is 5 and the reducing factor is 0.5. The number of heads in the attention mechanism is 4 and the dropout ratio is 0.1.

4) Evaluation metrics: Several common performance metrics have been selected for comparing the prediction results of different machine learning models including mean average error (MAE), mean absolute percentage error (MAPE) and root mean squared error (RMSE). The formulas are shown in Equations 9, 10 and 11 respectively.

MAE is estimated by

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i| \tag{9}$$

MAPE is estimated by

$$MAPE = \frac{1}{n}\sum_{i=1}^{n}\frac{|y_i - \hat{y}_i|}{y_i} \tag{10}$$

RMSE is calculated by

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2} \tag{11}$$

D. Result analysis

1) Prediction performance: The performance of the competitive baselines and proposed model are shown in Table II. All models are evaluated 10 times using different random seeds, and the 95% confidence interval is used to compare the model performance with the assumption of the t-distribution.

Among the models, the tree-based models, i.e., XGBoost and RandomForest, show similar performance. Although XGBoost is slightly worse than RandomForest in terms of MSE, the training speed is much faster due to the consideration of gradient information for each weak classifier. The performance of the vanilla LSTM model is better than the tree-based models, which demonstrates the effectiveness of the RNN model in capturing temporal correlation.

In comparison, the proposed model not only extracts temporal correlation through LSTM but also captures the spatial correlation of different monitoring stations through Attention operation, which significantly improves the model performance by 20.4% and 9.8% on MSE, compared to XGBoost and vanilla LSTM, respectively. When comparing the models, the full feature space is used for each individual model.

2) Feature analysis: Feature analysis facilitates the understanding of the importance of input features, and helps determination of best model form. For the RNN model, gradient-based methods are used to estimate feature ranking,

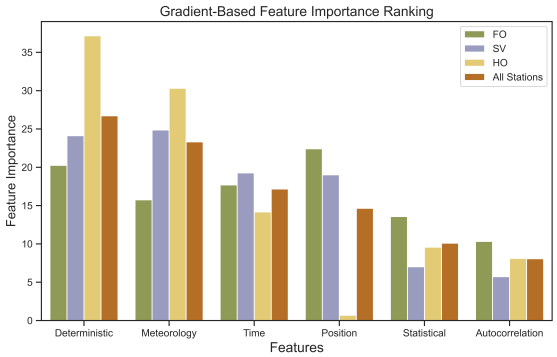

Fig. 4: Gradient-Based Feature Importance Ranking

which depends on both input and output data. Therefore, the feature importance was computed as the average of the gradient obtained from all samples in the test dataset [19]. The results are normalized proportionally and are shown in Figure 4.

The importance of different features may change from sensor station to sensor station, while the deterministic and meteorological forecasts are the most influential factors. This indicates that the deterministic model captures a certain amount of temporal-spatial context, and weather condition has a strong impact on the dispersion of pollutants, which is in alignment with our subjective reasoning. The importance rankings of calendar features indicate the latent temporal cyclic pattern. Positional encoding has little effect on HG street, which is consistent with the relatively accurate prediction of the deterministic model for this site.

3) Effects of Input Features: To further evaluate the proposed model, different input features are applied to train different models according to the feature rankings shown in Figure 4. Table III summarizes the model performance, in terms of three metrics values, of the proposed model with different input features. It is easy to see that the more features incorporated in the proposed model the better performance the proposed model can achieve. The first model with all input features leads to the best prediction results. The trend of model performance and corresponding confidence intervals are demonstrated in Figure 5.

While the model containing all features leads to the best performance, removing the deterministic forecasts results in a slight decrease in the prediction accuracy. This indicates that it is possible to remove the input factor in the air quality information system, given the high complexity and computational cost of the diffusion model. Nevertheless, removing both deterministic and meteorological forecasts leads to a significant reduction in the model performance.

The evaluation results of the models with different input features in Fig. 5 are consistent with the feature importance analysis in Fig. 4. This reflects the influence of multiple external variables on air pollutants forecasting, and meanwhile supports the effectiveness of the gradient-based feature impor-

TABLE III: Effects of different features

No.	Features	MSE	MAE	RMSE
All	**a + E + T + M + Det**	**920.338**	**18.162**	**29.967**
Model 1	**a + E + T + M**	1031.356	20.082	31.586
Model 2	**a + E + T**	1660.692	25.533	39.933
Model 3	**a + E**	2105.533	28.191	44.818
Model 4	**a**	1858.296	26.515	42.320

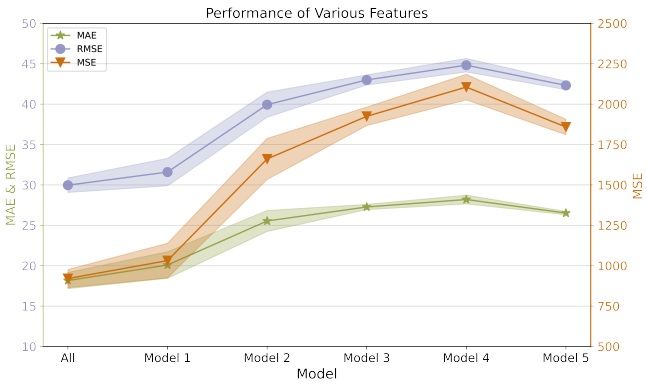

Fig. 5: Performance of Different Features

tance ranking method.

IV. Conclusions

In this paper, a meta-graph based deep learning framework is proposed in the context of forecasting air pollutants concentrations. The framework is designed to accommodate heterogeneous sensor data, also capable to capture temporal and spatial correlations effectively. Each meta-graph unit extracts the temporal correlation of input features and represents the local context of a specific sensor station. The multi-head attention mechanism is applied in the fusion layer to aggregate the spatial correlations across different sensor stations.

Comparing to the tree-like and LSTM models, the prediction performance of the proposed model achieves significant improvement, 20.4% and 9.8% with respect to MSE. The promising results suggest the effectiveness of the proposed deep learning framework. Meanwhile, the results of gradient-based feature importance analysis and ablation experiments demonstrate the crucial influence of deterministic and meteorological forecasts on air pollutant predictions and also lay the foundation for subsequent model distillation.

The computational experiment of the model is still simplified. In principle, the inputs and outputs for each station can be different when applying this framework. The scalability of the model for large sensor network needs further investigation since the current evaluation is still based on a number of sensor stations. Meanwhile, since road traffic is the major source of air pollution in Stockholm, bringing the road traffic data into the model may achieve potential improvement in the model prediction.

References

[1] R. Fuller, P. J. Landrigan, K. Balakrishnan, G. Bathan, S. Bose-O'Reilly, M. Brauer, J. Caravanos, T. Chiles, A. Cohen, L. Corra, *et al.*, "Pollution and health: a progress update," *The Lancet Planetary Health*, 2022.

[2] Z. Zhang, C. Johansson, M. Engardt, M. Stafoggia, and X. Ma, "Improving 3-day deterministic air pollution forecasts using machine learning algorithms," *Atmospheric Chemistry and Physics Discussions*, pp. 1–52, 2023.

[3] H. Fan, Y. Wang, C. Zhao, Y. Yang, X. Yang, Y. Sun, and S. Jiang, "The role of primary emission and transboundary transport in the air quality changes during and after the covid-19 lockdown in china," *Geophysical Research Letters*, vol. 48, no. 7, p. e2020GL091065, 2021.

[4] M. H. Askariyeh, M. Venugopal, H. Khreis, A. Birt, and J. Zietsman, "Near-road traffic-related air pollution: Resuspended pm2. 5 from highways and arterials," *International journal of environmental research and public health*, vol. 17, no. 8, p. 2851, 2020.

[5] L. Gidhagen, C. Johansson, J. Langner, and V. Foltescu, "Urban scale modeling of particle number concentration in stockholm," *Atmospheric Environment*, vol. 39, no. 9, pp. 1711–1725, 2005.

[6] S. Munir, M. Mayfield, D. Coca, L. S. Mihaylova, and O. Osammor, "Analysis of air pollution in urban areas with airviro dispersion model—a case study in the city of sheffield, united kingdom," *Atmosphere*, vol. 11, no. 3, p. 285, 2020.

[7] C. Granier, S. Darras, H. D. van Der Gon, D. Jana, N. Elguindi, G. Bo, G. Michael, G. Marc, J.-P. Jalkanen, J. Kuenen, *et al.*, "The copernicus atmosphere monitoring service global and regional emissions (april 2019 version)," 2019.

[8] M. Stafoggia, C. Johansson, P. Glantz, M. Renzi, A. Shtein, K. de Hoogh, I. Kloog, M. Davoli, P. Michelozzi, and T. Bellander, "A random forest approach to estimate daily particulate matter, nitrogen dioxide, and ozone at fine spatial resolution in sweden," *Atmosphere*, vol. 11, no. 3, p. 239, 2020.

[9] A. Shtein, I. Kloog, J. Schwartz, C. Silibello, P. Michelozzi, C. Gariazzo, G. Viegi, F. Forastiere, A. Karnieli, A. C. Just, *et al.*, "Estimating daily pm2. 5 and pm10 over italy using an ensemble model," *Environmental science & technology*, vol. 54, no. 1, pp. 120–128, 2019.

[10] R. Navares and J. L. Aznarte, "Predicting air quality with deep learning lstm: Towards comprehensive models," *Ecological Informatics*, vol. 55, p. 101019, 2020.

[11] 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, June 2021.

[12] ——, "Multivariate time series forecasting via attention-based encoder-decoder framework," *Neurocomputing*, vol. 388, pp. 269–279, 2020.

[13] H. Bi, L. Lu, and Y. Meng, "Hierarchical attention network for multivariate time series long-term forecasting," *Applied Intelligence*, vol. 53, no. 5, pp. 5060–5071, 2023.

[14] Y. Han, J. C. Lam, V. O. Li, and Q. Zhang, "A domain-specific bayesian deep-learning approach for air pollution forecast," *IEEE Transactions on Big Data*, vol. 8, no. 4, pp. 1034–1046, 2020.

[15] X. Ma, J. Xu, M. Nordenvaad, and T. Julner, "Digiways: A digitalisation testbed for sustainable traffic management on Swedish motorways," in *The 9th IEEE World Forum on Internet of Things (IEEE WFIoT 2023)*. IEEE, Oct. 2023.

[16] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," *Advances in neural information processing systems*, vol. 30, 2017.

[17] L. Breiman, "Random forests," *Machine learning*, vol. 45, pp. 5–32, 2001.

[18] T. Chen, T. He, M. Benesty, V. Khotilovich, Y. Tang, H. Cho, K. Chen, R. Mitchell, I. Cano, T. Zhou, *et al.*, "Xgboost: extreme gradient boosting," *R package version 0.4-2*, vol. 1, no. 4, pp. 1–4, 2015.

[19] D. Baehrens, T. Schroeter, S. Harmeling, M. Kawanabe, K. Hansen, and K.-R. Müller, "How to explain individual classification decisions," *The Journal of Machine Learning Research*, vol. 11, pp. 1803–1831, 2010.