AccuAir: Winning Solution to Air Quality Prediction for KDD Cup 2018

Zhipeng Luo
DeepBlue Technology
luozp@deepblueai.com

Jianqiang Huang
Peking University
jqhuang@pku.edu.cn

Ke Hu
Alibaba Group
ke.hu@alibaba-inc.com

Xue Li
Microsoft
lixue421@gmail.com

Peng Zhang
Tianjin University
pzhang@tju.edu.cn

ABSTRACT

Since air pollution seriously affects human heath and daily life, the air quality prediction has attracted increasing attention and become an active and important research topic. In this paper, we present AccuAir, our winning solution to the KDD Cup 2018 of Fresh Air, where the proposed solution has won the 1st place in two tracks, and the 2nd place in the other one. Our solution got the best accuracy on average in all the evaluation days. The task is to accurately predict the air quality (as indicated by the concentration of PM2.5, PM10 or O3) of the next 48 hours for each monitoring station in Beijing and London. Aiming at a cutting-edge solution, we first presents an analysis of the air quality data, identifying the fundamental challenges, such as the long-term but suddenly changing air quality, and complex spatial-temporal correlations in different stations. To address the challenges, we carefully design both global and local air quality features, and develop three prediction models including LightGBM, Gated-DNN and Seq2Seq, each with novel ingredients developed for better solving the problem. Specifically, a spatial-temporal gate is proposed in our Gated-DNN model, to effectively capture the spatial-temporal correlations as well as temporal relatedness, making the prediction more sensitive to spatial and temporal signals. In addition, the Seq2Seq model is adapted in such a way that the encoder summarizes useful historical features while the decoder concatenate weather forecast as input, which significantly improves prediction accuracy. Assembling all these components together, the ensemble of three models outperforms all competing methods in terms of the prediction accuracy of 31 days average, 10 days average and 24-48 hours.

CCS CONCEPTS

• **Applied computing** → *Environmental sciences*;

KEYWORDS

Air quality prediction; KDD CUP 2018; Seq2Seq; Deep Learning

ACM Reference Format:
Zhipeng Luo, Jianqiang Huang, Ke Hu, Xue Li, and Peng Zhang. 2019. AccuAir: Winning Solution to Air Quality Prediction for KDD Cup 2018. In *The 25th ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD ’19), August 4–8, 2019, Anchorage, AK, USA.* ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/3292500.3330787

1 INTRODUCTION

Air pollution, as one of the most serious side-effects of rapid urbanization, has become an environmental and social issue in many developing countries, endangering the health of billions of people [22]. To monitor air pollution in real time, Beijing has established 35 air monitoring stations and 13 air monitoring stations have been established in London. In addition to monitoring, there is a growing demand for accurately predicting future air quality, which can influence government policy making (such as traffic control when air pollution is severe) and people’s decisions (such as whether it is suitable for exercising outdoors). Therefore, it is very necessary to develop accurate air quality prediction algorithms. Indeed, an accurate air quality prediction, especially for long-term and sudden changing phenomenon, is extremely challenging.

First, the concentration of air pollutants can be affected by many factors, including natural meteorologic conditions such as weather and wind, human factors such as traffic pollution and industrial pollution, as well as the surrounding air quality. The complex interactions of all the aforementioned factors make it a very difficult prediction task. To intuitively show this, Figure 1 depicts the fluctuations of PM2.5, a key measurement for air quality, measured at Beijing Olympic Sports Center Station, during March and May in 2018. According to Figure 1, this station suffers from air pollution for most the time, and the values of PM2.5 change wildly and quickly. For instance during the period highlighted as red, the values of PM2.5 form two sharp peaks in less than 6 days. The sudden change of air quality poses challenges to its accurate prediction.

Second, due to these complex factors, the change of air quality are significant over time and location, with temporal-spatial correlations. As shown in Figure 2, by comparing the air quality of three stations, the PM2.5 values are very close when air quality appears good (highlighted as green), whereas huge discrepancy is observed when air quality gets worse (highlighted as red). In addition, PM2.5 at three stations varies from time to time. As shown in Figure 3, on May 5-6, station S1 and station S2 showed a tendency to rapidly

KDD ’19, August 4–8, 2019, Anchorage, AK, USA
© 2019 Association for Computing Machinery.
ACM ISBN 978-1-4503-6201-6/19/08…$15.00
https://doi.org/10.1145/3292500.3330787

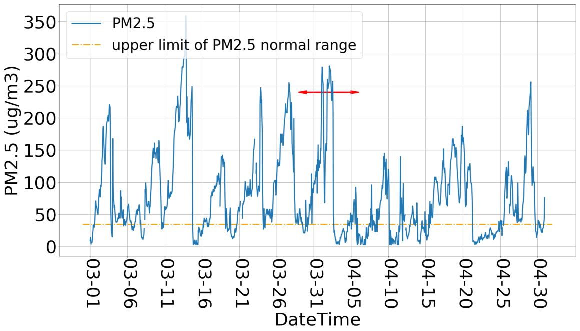

Figure 1: PM2.5 vs. date time at Beijing Olympic Sports Center Station, where the metric values change suddenly without any obvious patterns.

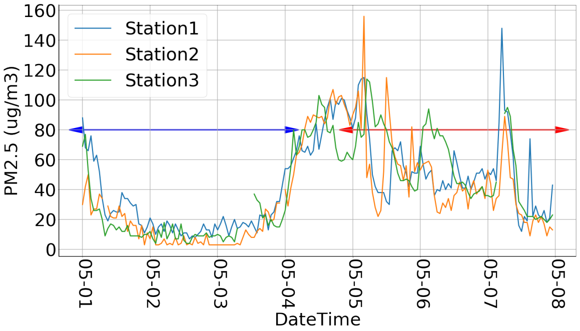

Figure 2: PM2.5 vs. date time at 3 different Beijing stations.the PM2.5 values are very close when air quality appears good (highlighted as green), whereas huge discrepancy is observed when air quality gets worse (highlighted as red).

decline and then rise again, while station S3 maintained this 20 declining trend, monotonously. This implies that it is important to model the spatial-temporal correlation in the air prediction.

Third, in addition to normal fluctuations, the air quality can be dramatically changed due to certain factors. As shown in Figure 3, the PM2.5 values in this station show a large mutation during May 13 to 14, soaring to 190 from 40 and then fall back to 40 within 12 hours. The sudden changes usually result from unusual weather conditions, such as rainstorm and typhoon, or from the surrounding environment of the city such as sandstorms. The above observations and analyses on sudden changes pose greater challenges for long-term predictions for the air quality, while the human decision-making often depends on such sudden changes. This implies that it is challenging and important to predict the sudden changes of air quality.

Fourth, due to the influence of some unknown factors, it is extremely difficult to predict the air quality for a longer time in the future. In people's daily lives, many people want to know the air quality in the future, so that it is easy to arrange future travel plans. Usually, the current air quality will have a certain impact on the next day. The longer the predicted range of time, the more unknown factors, and the more difficult these factors to be captured. Thus, it is difficult to predict future air quality for a longer period of time.

To address the above challenges, we propose in this paper an ensemble model named AccuAir, which is capable of Accurate prediction of the Air quality. We address the challenges in air quality prediction from both the feature and model perspective, as shown in Figure 4. For the former, to better characterize the interactions among all the influencing factors, we design several groups of feature, including both global features and local ones.

For the model perspective, we design an ensemble model comprising a LightGBM, a spatial-temporal gated DNN, and a Seq2Seq model. Specifically, LightGBM is used as a fast and robust feature selector, which allows us to identify the most important features more efficiently. Taking LightBGM as a baseline, we further design a spatial-temporal gated DNN model to reflect the complex influence of spatial-temporal correlations. Compared with conventional DNN models which concatenate all features indiscriminately, our gated network enables us to leverage such correlations to impose direct impact on the final prediction results, in a simple and effective way. Finally, the Seq2Seq model is adapted as a sequential predictor, with an encoder to summarize all the recent historic features, and a decoder to consider the future air quality data with weather forecast features as input. Combining their respective advantages, our proposed ensemble model won the first prize in 2018 KDD Cup of Fresh Air, outperforming all the competing methods from 4206 teams[1].

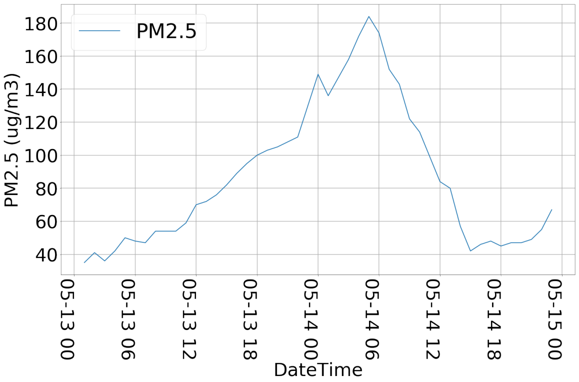

Figure 3: PM2.5 changes sharply at Beijing Olympic Sports Center Station in two days.The sudden changes usually result from unusual weather conditions, such as rainstorm and typhoon, or from the surrounding environment of the city such as sandstorms.

Our main contributions are summarized as below:

- We present a brief analysis on the influencing factors of air quality prediction, identify several major challenges in this problem, and propose to address such challenges from both the feature and model perspective.

[1] https://biendata.com/competition/kdd_2018/

- We propose an ensemble framework for air quality prediction, including a LightGBM baseline for efficient feature selection, a spatial-temporal gated network to effectively capture spatial-temporal correlations, as well as an adapted Seq-2Seq model to improve long-term prediction accuracy by incorporating weather forecast features.
- We conduct extensive experiments on the data offered by 2018 KDD Cup of Fresh Air, demonstrating the effectiveness of the proposed method.

The remainder of this paper is organized as follows. Section 2 provides a brief introduction of existing air quality prediction methods. Section 3 explains our proposed method in detail, followed by Section 4 which summarizes the main experimental results. Finally, Section 5 concludes this paper.

2 RELATED WORK

With the increasing attention and impact of the air quality prediction task, many methods and systems have been proposed [5, 6, 12, 16, 17]. Roughly speaking, in this paper, we will review those works in three categories, namely deterministic methods [2, 9], statistical learning methods [1, 8, 21], and the recently deep learning methods [12, 16, 17]. The deterministic models, often with meteorological emissions and chemical models, are employed to build numerical functions for air quality predictions [4]. However, these numerical functions are built from many factors, which are either incomplete or inaccurate, leading to a relatively poor prediction accuracy [8]. In addition, these models have high complexity and thus rely on lots of computation powers [17].

The statistical learning methods, which can be divided to parametric methods [3, 7, 14, 18] and non-parametric ones [1, 8], are adopted for air quality predication problem. For instance, some parametric models were proposed, based on Classification And Regress Trees (CART) [3] and fuzzy logic ones [14]. Still, a problem is the tradeoff between the model complexity and computational feasibility [8]. In this paper, we are trying to adopt LightGBM [10], a highly efficient gradient boosting decision tree method. LightGBM can enable us to not only incorporate and select various features for air quality prediction, but also with satisfied computational efficiency.

Some earlier models only predict the air quality for each station, but many ignore the spatial correlation between different stations [6, 12], as now emphasized by the latest tasks, e.g., KDD 2018 Cup of Fresh Air. In addition, since the air prediction involves a relatively long-term prediction, i.e., 48 hours predication versus the previous 8-24 hours. This also poses great challenges towards modeling spatial-temporal correlation for air quality prediction. Therefore, we need a deep learning structure that is able to extract complex spatial-temporal correlation features [12].

Recently, some deep learning algorithms have been proposed, to extract spatial correlation and temporal dependency, respectively. For instance, Convolutional Neural Networks (CNNs) are adopted to capture spatial correlation [19, 20], while Recurrent Neural Networks (RNNs) are utilized to model the temporal dependency [15]. However, a direct use of CNNs and RNNs (or LSTMs) is not practical for air quality prediction task, since this task involves sparse data and low-quality temporal data for training. Hence, more recently, a deep distributed fusion network is proposed, which can fuse heterogene-ous urban data [17].

Different from the fusion strategy [17], for modeling spatial-temporal correlation, we propose a spatial-temporal gate in DNN. Our method can effectively capture the spatial topological correlations as well as temporal relatedness, making the prediction more sensitive to spatial and temporal signals. In addition, we effectively utilize the RNN structure in the air quality prediction task, by incorporating the historical air quality and meteorology data in the encoder, and the concatenate weather forecast data in the decoder. In such a way, together with the feature engineering well designed, our architecture achieves the best result in KDD 2018 Cup of Fresh Air.

3 PROPOSED METHOD

In this section, we first introduce the overall structure of our ensemble models, and then explain the three component models in detail.

3.1 An Overview

Figure 4 shows the schematic illustration of the proposed model, including both the major feature groups as well as the component models used in the final ensemble.

To better characterize the influencing factors, we construct five feature groups.

As mentioned previously, we design an ensemble model comprising three component models, each plays a different role in the final prediction. Specifically, a LightGBM model is taken as baseline due to its efficiency and robustness, and is designed to perform prediction by seeking the optimal partition strategy among different features. The feature importance obtained by LightGBM also sheds lights on the training of other models. In the meanwhile, a spatial-temporal gated DNN model is designed to construct high-order features via combining multiple linear and nonlinear operations, and encourage sensitivity to spatial-temporal relatedness by inserting a spatial-temporal gate. Finally, a Seq2Seq model is incorporated to enhance sequence prediction capacity by leveraging weather forecast feature in the decoder side, which greatly improves prediction accuracy on sharp changes.

We deliberately increase discrepancy among component models in order to achieve better ensemble results. Details about the three component models will be explained in the following sections.

3.2 LightGBM for Feature Selection

As a highly efficient gradient boosting decision tree, LightGBM has remarkable capabilities in handling noisy data, making it a popular choice for a wide spectrum of applications. In our proposed method, we use LightBGM not only as a baseline, but also an efficient feature selector to help us iteratively improve our model and validate the importance of features.

Figure 5 shows the procedure of how we improve the model step by step using LightGBM. We starting from training the LightGBM with the meteorology feature in the last 72 hours, as well as the time and station id. As can be seen from the top-left sub-figure of Figure 5, this simple baseline takes the same feature to make

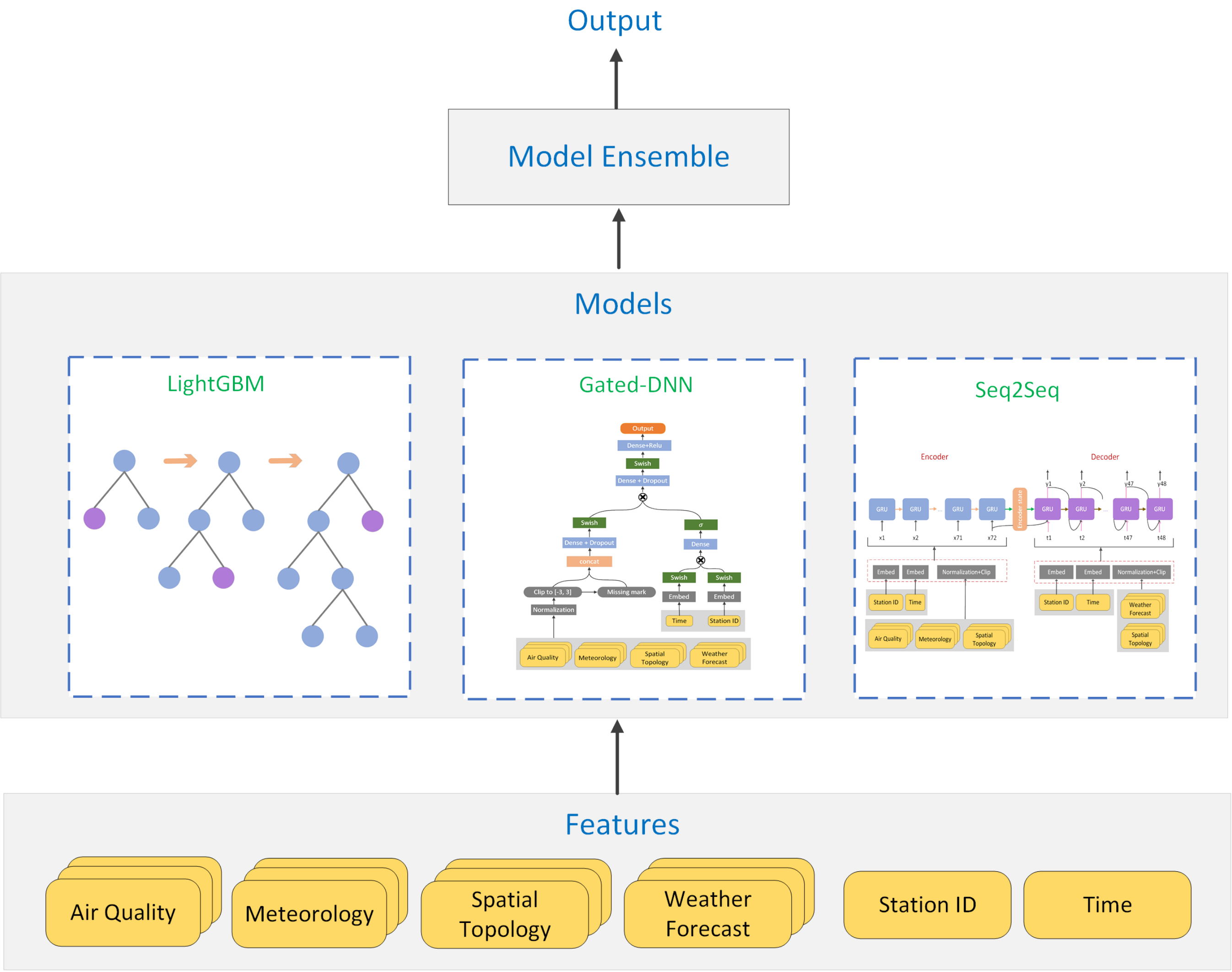

Figure 4: A schematic illustration of the proposed model, which shows the major feature groups including time & station id embedding, air quality feature, meteorology feature, weather forecast feature, and spatial-temporal feature, as well as three model blocks including LightGBM, spatial-temporal gated DNN, and a Seq2Seq model.

predictions about future air quality, without special manipulation to help it distinguish future time steps clearly, and thus fails to make accurate long-term predictions, especially for the dramatic changes shown in Figure 5.

Having observed such problems, we attempt to increase the difference among time steps by adding weather forecast feature directly into the corresponding time step, enabling the model to capture the relatedness between air quality and instant weather conditions. As shown in the top-right sub-figure of Figure 5, our model now could basically predict the air quality change to some degree.

Next, we try to add the history data, meteorology feature and weather forecast feature from the 8-neighboring stations into the prediction of current station, since the air quality of adjacent locations are intrinsically correlated. As shown in the bottom-left sub-figure of Figure 5, the prediction accuracy is further improved.

Finally, we evenly sample 12 stations and 12 grids within the entire city, to obtain the air quality and weather feature of different directions for better long-term prediction. This significantly improves the prediction accuracy, as shown in the bottom-right sub-figure of Figure 5.

3.3 Spatial-temporal Gated DNN

This section introduces our spatial-temporal gated DNN model, which is chosen as a component model due to its distinctions from decision tree models. Specifically, DNN models are better at automatically learning salient representations via its deep structures,

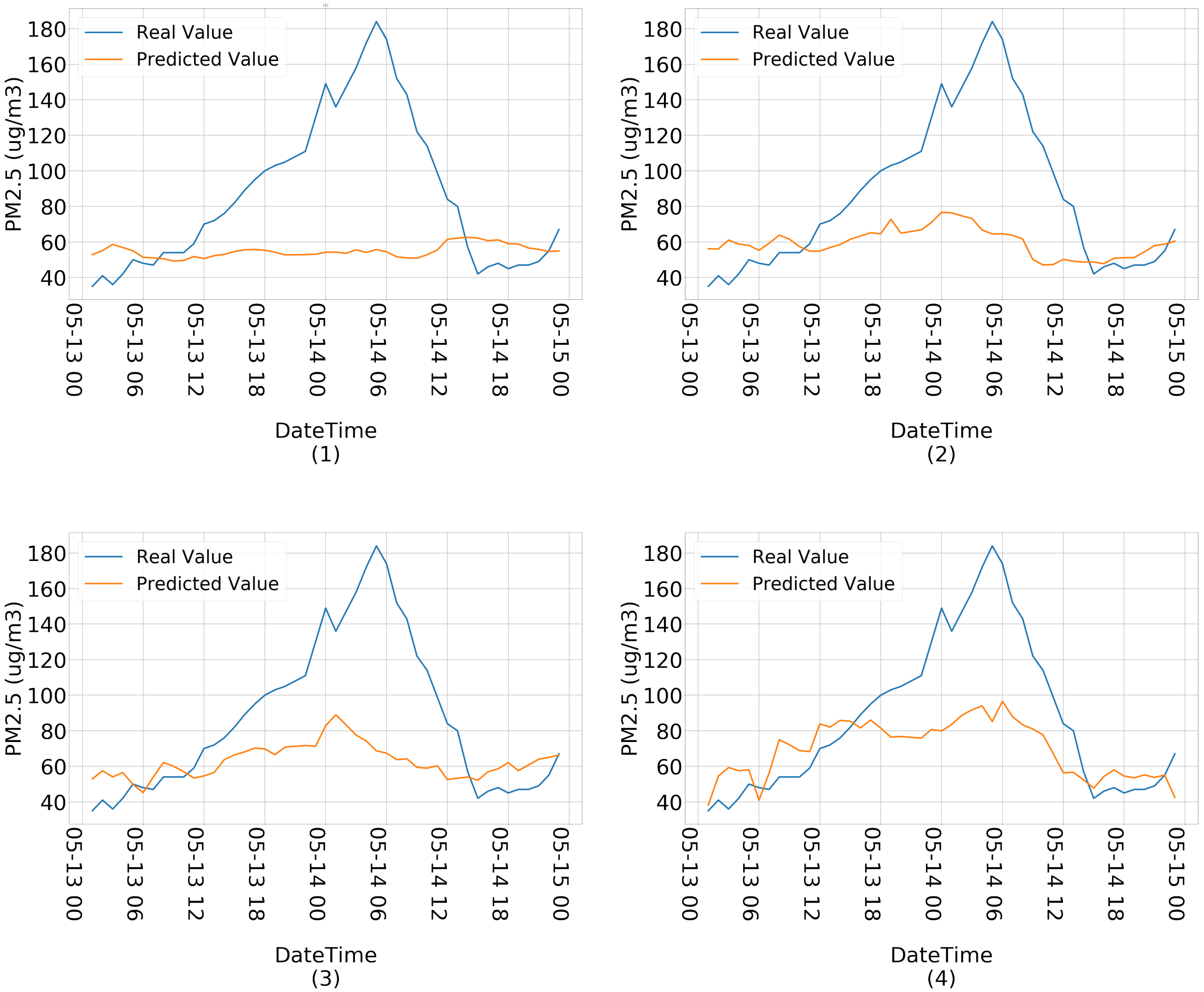

Figure 5: An illustration of feature iteration process.(1)The meteorology feature in the last 72 hours.(2)Adding weather forecast feature directly into the corresponding time step.(3)Add the history data, meteorology feature and weather forecast feature from the 8-neighboring stations into the prediction of current station.(4)Sampling 12 stations and 12 grids within the entire city, to obtain the air quality and weather feature of different directions for better long-term prediction.

and are capable of embedding the complex spatial-temporal relatedness into different dimensions. Such distinctions make it a suitable choice as a complement of LightGBM.

However, conventional fully connected DNN models does not suit our task. To see this, consider the prediction task at a certain station. When making predictions for next 3 days, a conventional deep regression model will tend to make very close predictions for all three days. This is easy to understand since when most of the available information are identical, the model can hardly find any clues to reflect spatial and temporal difference. To fix this issue, we propose a spatial-temporal gate, which is able to generate gating signal from spatial-temporal features, and to impose direct influence on prediction results by controlling the strength of that gating signal.

The structure of the proposed spatial-temporal gated DNN model is shown in Figure 6. This network takes as input the history air quality data, meteorology feature, spatial-temporal feature as well as weather forecast feature, and performs some basic pre-processing such as normalization, clipping, and masking. The masking operation is designed to reduce noise brought by missing data samples. Then, the clipped and masked data are concatenated and denoted by x, and is fed into a dense layer with Dropout:

$$h_1 = f(W_1 x + b_1) \tag{1}$$

where $f(\cdot)$ represents Swish activation function proposed in [13], defined as below:

$$f(x) = x \cdot \sigma(\beta x) \tag{2}$$

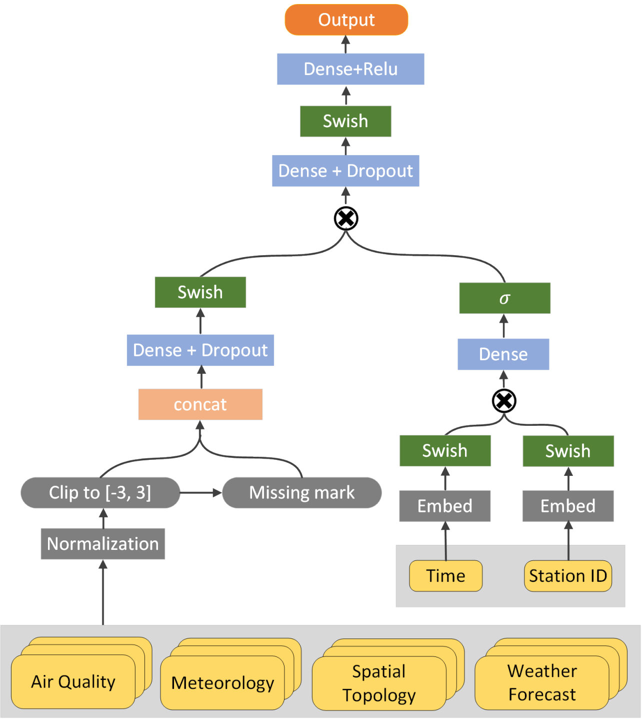

Figure 6: Structure of the proposed spatial-temporal gated DNN model.

where

$$\sigma(z) = (1 + exp(-z))^{-1} \tag{3}$$

The bottom-right part in Figure 6 shows the proposed spatial and temporal gate, which is constructed from time feature and station ID. For this purpose, the time feature and station ID are first embedded into t and s with same dimension, and the gating signal is then generated as follows:

$$h_2 = \sigma(W_2(f(t) \cdot f(s)) + b_2) \tag{4}$$

Equation 4 generates gating signal by passing the element-wise product of $f(t)$ and $f(s)$ through a dense layer. This gating signal is then used for controlling the output of Equation 1:

$$h_3 = f(W_3(h_1 \cdot h_2) + b_3) \tag{5}$$

Finally, the output is obtained as

$$y = max(W_4 h_3 + b_4, 0) \tag{6}$$

As defined in Equation 5, we perform element-wise product on h_1 and h_2, allowing fine-grained control in specific stations and certain periods. Since air quality measurements are always positive, we use ReLU as activation function in final prediction result. As can be seen later in Section 4, the proposed gating mechanism greatly improves the sensitivity of model towards spatial-temporal relatedness.

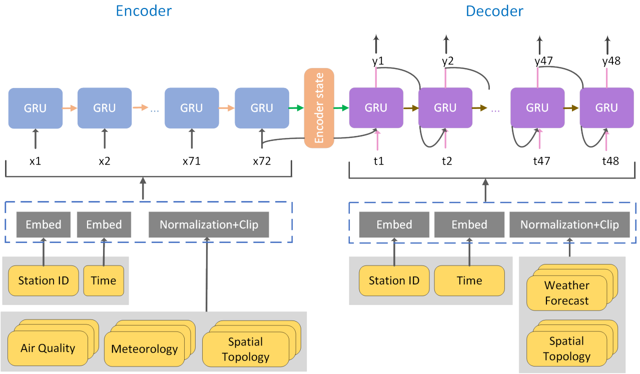

Figure 7: Structure of the adapted Seq2Seq model.

3.4 Sequence to Sequence Model

Long-term air quality prediction is a typical scenario of sequence modeling, which inspires us to include sequence to sequence models into our ensemble. In a typical encoder-decoder architecture, the encoder takes source language as input, encoding it into an intermediate state, and feeds it to the decoder. The decoder then decode the received intermediate state into a target language, using RNN to output a word sequence step by step.

Our sequence to sequence model, as shown in Figure 7, inherits this encoder-decoder structure, with the encoder takes as input feature in the last 72 hours, including station ID embedding, time embedding, history air quality, meteorology feature, spatial topology, etc., forming a 72-dimensional vector represented as x_1 to $x_7 2$, each characterizing a single hour. Each x_i is mapped into a hidden state as below:

$$h_{ie} = f(h_{i-1}, x_i) \tag{7}$$

where $f(\cdot)$ denotes a GRU unit, h_{ie} is the hidden state for the corresponding GRU.

Our decoder keeps the output from the last time step, concatenates it with time embedding, station ID embedding as well as weather forecast feature, forming a 48-dimensional input vector, denoted by t_1 to t_{48}. For each t_i, we mapping it as below:

$$h_{id} = f(h_{i-1}, y_{i-1}, t_i) \tag{8}$$

where h_{id} denotes the hidden state for decoder GRU.

The final prediction result is obtained by a fully connected layer with ReLU activation:

$$y_i = max(Wh_i + b, 0) \tag{9}$$

Similar to the DNN model in Section 3.3, we also add masks on missing data samples to reduce noise, but this time we apply the mask on loss calculation instead of feature pre-processing. Specifically, following the idea of [11], an additional cost term is added as below:

$$\beta \frac{1}{T} \sum_{t=1}^{T} (||h_t||_2 - ||h_{t-1}||_2)^2$$

where h_t is the hidden state of GRU. According to the analysis in [11], adding such an addition loss could effectively prevent the exponential growth of RNN's activation outside of their training horizon, allowing them to generalize to much longer sequences.

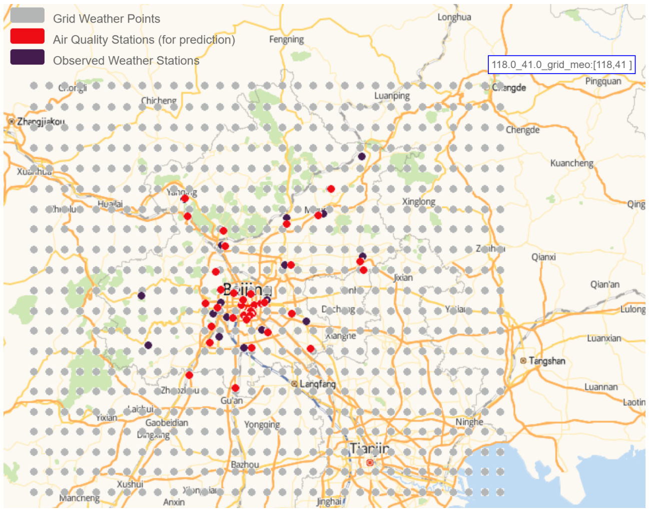
Grid Weather Points
Air Quality Stations (for prediction)
Observed Weather Stations
118.0_41.0_grid_meo:[118,41]
Zhangjiakou
Beijing
Tangshan
Langfang
Tianjin

Figure 8: Illustration of the geographical grids for Beijing.

Our encoder-decoder model is constructed using GRU as recurrent unit, and it predicts air quality in the future 48 hours according to 72-hour history feature. To handle missing samples, we apply a mask to the loss corresponding to the missing node.

4 EXPERIMENTAL RESULTS

In this section, we first explain our experimental setup, including evaluation protocol and data, and then compare our solutions with that of other teams. We also provide an analysis on the features and ensemble strategies.

4.1 Experimental Setup

In this section we evaluate the proposed model using the data provided by KDD Cup 2018 of Fresh Air. This dataset contains three types of features, including 1) history air quality feature such as values of PM2.5, PM10 and O3, 2) meteorology feature for different geographical grids, such as weather, temperature, air pressure, humidity, speed and direction of wind, and 3) weather forecast features for the same set of geographical grids. Figure 8 and Figure 9 illustrate the geographical grids for Beijing and London, respectively.

The data is composed of air quality data of the last year, and weather forecast feature of the last month. The evaluation metrics used for KDD Cup is SMAPE (Symmetric Mean Absolute Percentage Error), which is defined as below:

$$\text{SMAPE} = \frac{1}{n}\sum_{t=1}^{n}\frac{|F_t - A_t|}{(A_t + F_t)/2} \tag{10}$$

where F_t and A_t represent predicted and true values at time step t, respectively. Smaller SMAPE values indicate better performance.

4.2 Performance Comparison

In order to make a comprehensive comparison of the models, Table 1 presents the scores of the top5 teams in 4 major parts. In each part we list the name of the top5 teams (where our team is represented by *getmax*), and their scores in terms of a particular metric. The evaluation scores are based on five air pollutant concentration measurements in two cities, Beijing and London. We first calculate the respective SMAPE of each city, and then use their arithmetic mean across the two cities as the final SMAPE value for a given day. In order to reflect the performance of the model from different aspects, we consider four evaluation methods, namely:

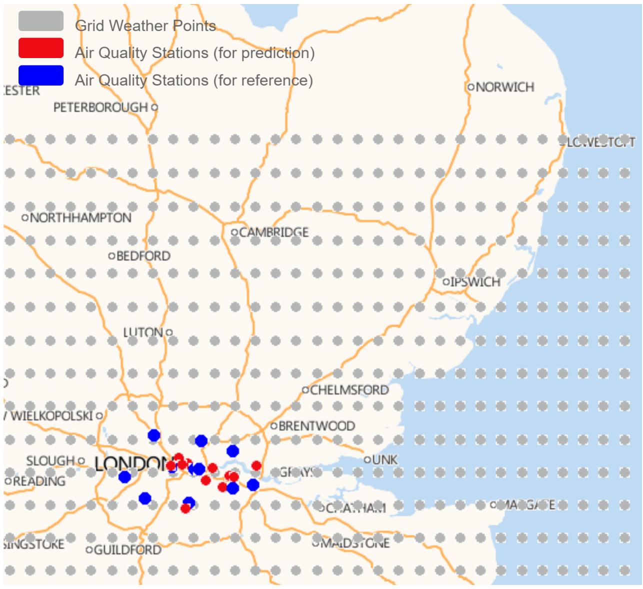

Figure 9: Illustration of the geographical grids for London.

- To reflect the model performance under mild conditions and reduce the impact of sudden change, the most challenging 6 days are excluded from the from the entire 31 days, leading to a prediction task spanning 25 days, as shown in the first part of Table 1.
- The mean score of the last 10 days. Since the model is updated every day, most models gradually reach their optimal performance within the last 10 days. Therefore, this metric could reflect the best performance of each model.
- The mean score of 24-48 hours over the 31 days. It reflects the performance of long-term predictions, which is important for people to consider their choice of travel.
- The mean score of the entire 31 days. Although this metric is not adopted in the original track of the competition, it reflects the true performance of the model under very challenging conditions including sudden change. Note that our model not only achieves good results under mild conditions (see the 25 days case), but also gives the best results in the case of sudden changes in air quality.

The first three evaluation scores are used for the three tracks in the competition. Our team with more stable performance achieved first place in 31 days average, last 10 days average and 31 days 24-48 hours forecast average in the SMAPE evaluation respectively. It was ranked second after removing the most difficult 6-day results. Since the competition requires daily submission of predictions, the model is likely to change every single day. The evaluation of last 10 days indicates that our final model has achieved the best accuracy. Taking a huge advantage, it exceeds the second place with a 0.0097 of SMAPE. The gap between 2-5 is 0.0069, suggesting that our final model is significantly better than any other team.

Table 1: Overall Performance.

Team (Top5)	Mean SMAPE in 25 days
First floor to eat Latiao	**0.3681**
getmax (ours)	0.3696
头号玩家	0.3767
deepx	0.3847
迟到大队	0.3854
Team (Top5)	Mean SMAPE in 31 days
getmax (ours)	**0.3920**
First floor to eat Latiao	0.3935
头号玩家	0.3995
迟到大队	0.4051
oneday	0.4083
Team (Top5)	Mean SMAPE in last 10 days
getmax (ours)	**0.3652**
deepx	0.3749
First floor to eat Latiao	0.3785
迟到大队	0.3806
头号玩家	0.3818
Team (Top5)	Mean SMAPE in last 24-48 hours
getmax (ours)	**0.4317**
头号玩家	0.4321
迟到大队	0.4531
DAWN	0.4537
First floor to eat Latiao	0.4540

4.3 Analysis

In order to verify the effect of the features we use, we divide the features into different groups and add a new group each time for performance validation. Here we take Beijing PM2.5 as an example to conduct evaluations.

We divide the features into the following six groups:

F1. Basic features: the hours to be predicted, the predicted weekday, hour, the station ID and the latitude and longitude of the station, etc.

F2. Air quality features: concentration of pollutants in the past 1, 3, 5, ... 72 hours. Statistics including Average, median, maximum and minimum values of pollutant concentration, etc.

F3. Meteorology features: average wind speed in the past 1, 3, 5...48 hours, wind direction binning, etc.

F4. Weather forecast features: the wind speed and wind direction at the predicted hour. The cumulative value of the wind speed smoothing statistics, the extreme values, etc.

F5. Surrounding topological features: the historical air quality features of eight adjacent stations near the current station, and the historical meteorology features and weather forecasting of the eight grid points adjacent to the current station, etc.

F6. City topological features: the historical air quality features of twelve stations sampled throughout the city, and the historical meteorology and weather forecasting features of twelve grid points obtained form equidistant sampling, etc.

Table 2 summarizes our comparison results with the above feature groups.

Table 2: Comparison of different features.

	SMAPE
F1 + F2	0.5344
F1 + F2 + F3	0.5036
F1 + F2 + F3 + F4	0.4435
F1 + F2 + F3 + F4 + F5	0.4341
F1 + F2 + F3 + F4 + F5 + F6	0.3948

Table 3: SMAPE of top10 teams on 2018-05-27.

Rank	SMAPE	Rank	SMAPE
Rank1 (ours)	0.48	Rank6	0.60
Rank2	0.54	Rank7	0.62
Rank3	0.55	Rank8	0.64
Rank4	0.56	Rank9	0.66
Rank5	0.57	Rank10	0.66

Table 3 shows the model performance on May 27 2018, which witnesses a sudden change of air quality. The performance of our model on this day shows excellent mutation effect. In particular, we are the only team that reports a SMAPE score less than 0.5 among more than 4000 teams.

Model ensemble is a commonly used method to improve model accuracy in algorithmic competitions. Sometimes at the late stage of an intense competition, dozens or even hundreds of models are put into use. To verify the impact of each model in the final result, we design three types of model for accuracy improvement. Besides the accuracy of single model, the diversity of sub models also influences the ensemble accuracy. GBDT model uses the full features engineering sets, while Gated DNN model relies more on the automatic feature combination learning and is able to control the flow of time and spatial information through gated design. To further improve the modeling of time series, we use Air Seq2seq model with lots of adaptations. Finally, we use an additional model to learn the interaction of meta-learning (GBDT, DNN, seq2seq model). Unlike other problems such as Click-Through Rate (CTR) prediction in recommendation, the data sets of air problem is not so large whilst the data is very noisy. Thus, it is rather easy to over fit when non-linear models such as GBDT and DNN are used. We use a linear regression model to learn the interactions of meta-learners. Furthermore, to guarantee that the average prediction score is unchanged, a constraint of weight sum is added in linear model. The weight is constrained to [0, 1].

In this competition, the next 48 hours air quality is predicted. We choose 15 days data as evaluation sets. In our experiments, we evaluate GBDT, Gated DNN, Air Seq2Seq and Constrained Linear Ensemble Model (CLEM). We find that the accuracy of CLEM model is better than each individual model, and it is also better than the ensemble of two individual models.

Comparison of ensemble strategies are shown in Table 4.

Table 4: Comparison of different ensemble strategies.

	SMAPE
LightGBM	0.3948
DNN	0.4019
Gate-DNN	0.3969
Seq2Seq	0.3925
LightGBM + Gate-DNN	0.3866
LightGBM + Seq2Seq	0.3834
Seq2Seq + Gate-DNN	0.3851
LightGBM + Gate-DNN + Seq2Seq	0.3812

5 CONCLUSIONS

In this paper, we present our proposed ensemble model, AccuAir, which is capable of **Accu**rate prediction of the **Air** quality. Three models are integrated to predict air quality. In addition to the conventional GBDT, we develop gated DNN and Seq2Seq models in air quality prediction. Specifically, in order to solve the spatial-temporal problem in the DNN model, we designed a structure of spatial-temporal gates, which enables DNN, with the same input feature but different timing and physical positions, obtain different prediction results and capture the dynamic pattern of air quality. In addition, we also further improve the Seq2Seq model by considering the air quality predication task as a sequence prediction problem. Unlike the common Seq2Seq model, we use the weather forecast data to construct some future information features, by combining the temporal embedding features with the previous prediction results as input to the Decoder. This effectively reduces the difficulty of long-term prediction. In addition, in order to enhance the prediction stability of the model, we add a hidden layer regularization term in the Seq2Seq network, in which Decoder is used to obtain the information of the first 3 days.

In terms of feature design, we combine time space with air quality and weather data, including air quality features, weather forecast features, and topological location related features. Base on these features, more hidden high-level features are utilized.

In the SMAPE evaluation index, among the 4000+ teams in the competition, we achieved first place in the 31-day on average, the first place the last 10 days average and the first place in the long-term (24-48 hour) forecast 31-day on average, as well as the second place when excluding the most difficult 6 days in 31 days.

It is worth pointing out that the competition only provides relevant data for the predictive testing, but the air quality will be affected by the surrounding urban areas. Therefore, such data of the surrounding cities could be added to further augment the air quality prediction. In the meanwhile, historical data of the past few years and more training data could be considered to better train the model. This competition predicts the data in the real physical world, so it has a strong correlation with the actual application. Hope the framework and adaptations proposed in this paper could shed lights upon solving similar problems in real world.

REFERENCES

[1] Phillip Boyle and Marcus Frean. 2005. Multiple output Gaussian process regression. (2005).

[2] L Bruckman. 1993. Overview of the enhanced geocoded emissions modeling and projection (enhanced GEMAP) system. *Regional Photochemical Measurement and Modeling Studies. Volume* 2 (1993), 8–12.

[3] William R Burrows, Mario Benjamin, Stephen Beauchamp, Edward R Lord, Douglas McCollor, and Bruce Thomson. 1995. CART decision-tree statistical analysis and prediction of summer season maximum surface ozone for the Vancouver, Montreal, and Atlantic regions of Canada. *Journal of applied meteorology* 34, 8 (1995), 1848–1862.

[4] Jianjun Chen, Jin Lu, Jeremy C. Avise, John A. DaMassa, Michael J. Kleeman, and Ajith P. Kaduwela. 2014. Seasonal modeling of PM2.5 in California's San Joaquin Valley. *Atmospheric Environment* 92 (2014), 182 – 190.

[5] Cristiana Croitoru and Ilinca Nastase. 2018. A state of the art regarding urban air quality prediction models. In *E3S Web of Conferences*, Vol. 32. EDP Sciences, 01010.

[6] Xiao Feng, Qi Li, Yajie Zhu, Junxiong Hou, Lingyan Jin, and Jingjie Wang. 2015. Artificial neural networks forecasting of PM2. 5 pollution using air mass trajectory based geographic model and wavelet transformation. *Atmospheric Environment* 107 (2015), 118–128.

[7] Xiao Feng, Qi Li, Yajie Zhu, Jingjie Wang, Heming Liang, and Ruofeng Xu. 2014. Formation and dominant factors of haze pollution over Beijing and its peripheral areas in winter. *Atmospheric Pollution Research* 5, 3 (2014), 528–538.

[8] Vitor Campanholo Guizilini and Fabio Tozeto Ramos. 2015. A Nonparametric Online Model for Air Quality Prediction.. In *AAAI*. 651–657.

[9] Jaein I. Jeong, Rokjin J. Park, Jung-Hun Woo, Young-Ji Han, and Seung-Muk Yi. 2011. Source contributions to carbonaceous aerosol concentrations in Korea. *Atmospheric Environment* 45, 5 (2011), 1116 – 1125.

[10] Guolin Ke, Qi Meng, Thomas Finley, Taifeng Wang, Wei Chen, Weidong Ma, Qiwei Ye, and Tie-Yan Liu. 2017. Lightgbm: A highly efficient gradient boosting decision tree. In *Advances in Neural Information Processing Systems*. 3146–3154.

[11] David Krueger and Roland Memisevic. 2015. Regularizing RNNs by Stabilizing Activations. *CoRR* abs/1511.08400 (2015).

[12] Xiang Li, Ling Peng, Yuan Hu, Jing Shao, and Tianhe Chi. 2016. Deep learning architecture for air quality predictions. *Environmental Science and Pollution Research* 23, 22 (2016), 22408–22417.

[13] Prajit Ramachandran, Barret Zoph, and Quoc V Le. 2017. Swish: a self-gated activation function. *arXiv preprint arXiv:1710.05941* (2017).

[14] Rouzbeh Shad, Mohammad Saadi Mesgari, Arefeh Shad, et al. 2009. Predicting air pollution using fuzzy genetic linear membership kriging in GIS. *Computers, environment and urban systems* 33, 6 (2009), 472–481.

[15] Xuan Song, Hiroshi Kanasugi, and Ryosuke Shibasaki. 2016. DeepTransport: Prediction and Simulation of Human Mobility and Transportation Mode at a Citywide Level.. In *IJCAI*, Vol. 16. 2618–2624.

[16] Zheng Yan Jie Lu Guangquan Zhang Wang, Bin and Tianrui Li. 2018. Deep Multi-task Learning for Air Quality Prediction. In *International Conference on Neural Information Processing*. Springer, Cham, 93–103.

[17] Xiuwen Yi, Junbo Zhang, Zhaoyuan Wang, Tianrui Li, and Yu Zheng. 2018. Deep Distributed Fusion Network for Air Quality Prediction. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery &#38; Data Mining (KDD '18)*. 965–973.

[18] Jing Yuan, Yu Zheng, and Xing Xie. 2012. Discovering regions of different functions in a city using human mobility and POIs. In *Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, 186–194.

[19] Junbo Zhang, Yu Zheng, and Dekang Qi. 2017. Deep Spatio-Temporal Residual Networks for Citywide Crowd Flows Prediction. In *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence, February 4-9, 2017, San Francisco, California, USA*. 1655–1661.

[20] Junbo Zhang, Yu Zheng, Dekang Qi, Ruiyuan Li, Xiuwen Yi, and Tianrui Li. 2018. Predicting citywide crowd flows using deep spatio-temporal residual networks. *Artif. Intell.* 259 (2018), 147–166.

[21] Y Zheng, F Liu, and HP Hsieh. 2013. When urban air quality inference meets big data. In *Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining pp.(1436-1444). ACM.*

[22] Julie Yixuan Zhu, Yu Zheng, Xiuwen Yi, and Victor OK Li. 2016. A gaussian bayesian model to identify spatio-temporal causalities for air pollution based on urban big data. In *Computer Communications Workshops (INFOCOM WKSHPS), 2016 IEEE Conference on*. IEEE, 3–8.