PREDICTING AIR QUALITY USING TIME SERIES METEOROLOGICAL DATA IN TORONTO

Greta Skorupska-Ruiz
Student# 1010806718
greta.skorupskaruiz@mail.utoronto.ca

Tingkai Xue
Student# 1010806754
tingkai.xue@mail.utoronto.ca

Vivin Rithik Paul Rajendran
Student# 1010806745
vivin.paul@mail.utoronto.ca

Xinyi Gao
Student# 1005798406
xinyisara.gao@mail.utoronto.ca

ABSTRACT

Due to the worsening air quality in urban environments, prediction of air quality has become increasingly important. In this proposal, we outline a deep learning model that takes time series meteorological data (such as temperature and precipitation rate) as input to predict quantities of air pollutants. —-Total Pages: 6

1 INTRODUCTION

Due to urbanisation and economic growth, air pollution has become a serious issue in cities (Bikis, 2023). This can negatively affect a society on different levels, from individuals' health to the general economic productivity. To mitigate these negative impacts, efforts have been placed into performing air quality forecast, so that air pollution can be anticipated and preemptive measures can be taken.

Weather has been shown to have impact on air quality. For instance, rain and snow can improve air quality by removing particles and pollutants from the atmosphere and bringing aerosols to the ground (Wang et al., 2023). These factors will be called meteorological factors from now on. Various models can be developed to explain relationships between meteorological factors and air quality, however, these relationships are likely to be complex. This makes deep learning the ideal approach to develop a model to predict these relationships.

To gain a richer understanding of the relationships, time series weather data will be used. Meteorological factors such as temperature and precipitation are known to show spatio-temporal patterns (Huang et al., 2022). Uncovering these patterns can be beneficial to air quality prediction. These temporal patterns can also appear in different time spans.

In this project, we propose to use a deep learning approach to perform air quality prediction using time series meteorological data in Toronto. If possible, we will try to interpret the predictions made by the model.

2 PROJECT OUTLINE

The outline of our project can be found in figure 1. With the data obtained, preprocessing will be performed. This will involve data cleaning and also dimension reduction. The data will then be split into the training/validation set as well as test set. The training/validation set will be used to train the neural network model, and the performance of the trained model will be evaluated on the unseen test set. We will use data of different time spans to be used as input. More information on data processing and architecture can be found in section 4 and 5 respectively.

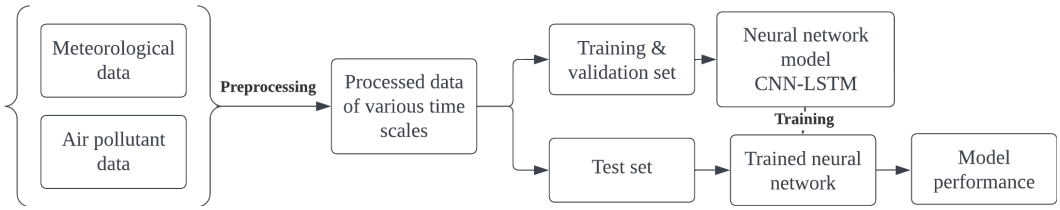

Figure 1: Illustration of project outline

3 Background and Related Work

Traditionally, air quality forecasts have been based on computational models. Recently, the NACC-CMAQ model has become part of the new operational National Air Quality Forecasting Capability (NAQFC) system in the US (Campbell et al., 2022). It involves coupling meteorological data (GFSv16) and chemical transport models (CMAQ) which have improved prediction accuracy. The advantages of these models are that they are deterministic and pay more attention to the underlying mechanisms that take place. However, these models are usually computationally intensive and may not uncover more complex relationships.

Convolutional neural networks (CNN) are often used in image processing due to their ability to extract shift-invariant features. Kow et al. (2022) used a combination of CNN and regression classifier (RC) to estimate air pollution quantities using images taken in Kaohsiung City in Taiwan and achieved accuracies of about 80%.

CNN may also be used to discover sequential features in time series data. Mao & Lee (2019) used one-dimensional CNN to perform air quality prediction by using air pollutant data from a fixed duration in the past to predict future air pollutant levels. These models performed well, while not making use of prior empirical knowledge of the effects of meteorological data on air pollutant concentration.

Another set of models commonly used in meteorology is recurrent neural networks (RNN). RNNs allow the outputs of previous training steps to used as subsequent inputs to the model, and are commonly used to analyze time series data. Some variants that are commonly used in air quality prediction are Long Short Term Memory (LSTM) Gated Recurrent Unit (GRU). Using RNN, LSTM, and GRU, Athira et al. (2018) managed to predict air quality with minimal errors, using meteorological features as inputs and the amount of air pollutants as outputs. These models' success can be expected as pollutants may linger around, therefore the current pollutant quantity may be dependent on its previous values.

Due to the advantages of both CNN and RNN in predicting air quality, some joint models have been developed. Zhang & Li (2022) proposed a CNN-LSTM pipeline to first extract temporal pattern features in air pollutant data using CNN, which are then passed to the LSTM model to learn changes in patterns for future air pollution prediction.

4 Data Processing

In our project, historical meteorological data will be obtained from the Government of Canada website (`https://climate.weather.gc.ca/historical_data/search_historic_data_e.html`) or from Weather Dashboard for Toronto (`https://toronto.weatherstats.ca/`). The data can be used for non-commercial purposes without additional permission and exported as a csv file. The data include:

- Wind speed
- Relative humidity
- Temperature
- Precipitation
- Dew point

- Atmospheric pressure

The available data are of different time scales - daily data over a period of around 5 years, and monthly data over a period of around 60 years. Choosing a time scale (daily/monthly) will be important for optimizing the model. Additionally, for both possible time scales, the number of data points taken into account in the prediction (sequential length) will be an optimizable hyperparameter.

Various pollutant concentrations can be found at the Ministry of the Environment, Conservation and Parks website (`https://www.airqualityontario.com/`). The pollutants include

- O_3
- PM2.5
- NO_2
- SO_2
- CO

While general air pollution data for the City of Toronto will be form the outputs, data from multiple (at least five) weather stations across Toronto will be used for the prediction and the measured parameters from all stations will be input as separate features. Some features are missing from selected datasets (some weather stations do not measure all aforementioned metrics, while others measure more), therefore the missing features will be removed, and singular missing data points (single days or months) will be replaced by a mean computed value. The weather data and the pollutant data will be matched, so that the measurement times match.

The cleaned data will be understood by looking at mean, variance, median and upper/lower quartiles. The change in the data against time can also be plotted to visualize any underlying trends. The correlation of the input features will also be plotted in a correlation matrix.

The data will be then normalized to ensure that extreme values do not affect the training of the models. Due to the inclusion of multiple weather stations, a large number of (correlated) features will be present. This might significantly increase the computational time of the model (especially relevant when optimizing), while also posing a higher risk of overfitting to a specific singular feature. Therefore, to reduce the computational time and improve the extent the model can generalize to unseen data, dimensionality reduction techniques have been considered. Principal Component Analysis will be used to reduce the dimensions of the features.

These features will form the inputs of the model, while chosen pollutant concentrations will be the outputs. Since PM2.5 is the most commonly interpretable metric for air polution, the principal focus will be put on its prediction (ie. it will be the output of the model). If this can be predicted with accuracy, other metrics will be included as outputs.

Finally, the dataset will be split into train, validation and test sets in a non-random manner to preserve the temporal aspect of the data.

5 ARCHITECTURE

The neural networks used will include CNN and (unidirectional) LSTM models. Other RNN models, such as the GRU have been considered, but none of them offer a significant advantage over the LSTM, which is considered a powerful method to model temporal data. Those models will first be developed individually; later, a tandem CNN-LSTM model will be created to ensure that both models' advantages can be combined. The CNN will be used to extract the most important local features of the temporal data and the LSTM will capture the repetitive time-dependencies. A dense layer will be added after the LSTM to obtain an output of the correct dimensions.

6 BASELINE MODEL

For the baseline model the random forest has been chosen, due to the readability and a low level of variation of the weather data. Some of the weather parameters will often return to the same state (ex. precipitation is zero), which can be well-separated using CART models. The random forest

will have a limited ability to model time-dependencies and its input will have to be flattened or averaged. The hyperparameters of the random forest model will be optimized to ensure the best possible performance of the baseline.

7 Ethical Considerations

Inaccurate or low-quality data can lead to unreliable predictions. If the model's output is used for critical applications like public health decisions, data quality is a significant ethical concern. Decisions based on this model's predictions if used for resource allocation for pollution control, may have ethical implications. The limitation of our model is that it is hard to make fair recommendations for society due to the diverse and unique nature of different communities. The limitation of our training data is that while it covers the different townships in Toronto it does not account for the specific features of each township, only providing a general prediction for the township.

8 Project Plan

Our team will be meeting every Monday from 5-7pm in person as a weekly group meeting for the updates and weekly work assignment. Team members could communicate with each other through a WhatsApp group chat for animmediate response or during the group meetings.

Proper coordination of coding tasks will be communicated before coding is done to avoid the situation where we overwrite each other's code. Annotations and code documentation should be added to allow other teammates to understand the rationale behind a piece of code. Github will be used to ensure all previous code can be retrieved.

The separation of tasks and tentative Gantt chart can be found in figure 2 at the end of the document.

9 Risk Register

9.1 Teammate dropping the course

This project is designed to simulate real-world industrial experience; it is important that the remaining teammates adapt to the increased responsibilities of each member. The project plan will be a vital and helpful document in this case as it clearly outlines the responsibilities taken on by the team member who has dropped the course. The remaining teammates should decide how to split these tasks in a fair and efficient way to ensure the project does not fall behind the course deadlines.

9.2 Inconsistency in data used

Different websites might have different metrics to calculate and report the various weather data. It is important to ensure that the instruments, matrices and units used in the data being used in the model are consistent. If data sourcing and cleaning are not done well it will cause the model to provide inaccurate predictions.

9.3 Poor dataset

If either the meteorological or air pollution data has low variability it might cause the model to be biased with low variability datasets. This overfitting on low variability data could lead to poor model performance when input with unexpected data with high variability.

9.4 Model training took too long

While unlikely, there is still a possibility of the model training taking too long. In this scenario, several actions could be taken to improve the training times. Data augmentation techniques to artificially increase the size of the training dataset could be used to allow for faster convergence. Adaptive Moment Estimation could be used where each weight has its own rate which would allow for faster convergence. Alternatively, data batching can be used to prevent the model from processing all available data at once.

9.5 Model is unable to predict unseen test data well

The model could have very low test accuracy. To improve test accuracy we could use Explainable AI (XAI). This would provide insights into how the model is making predictions. By visualizing and explaining the model's decision-making process we would be able to see where the model needs to be improved and make improvements accordingly.

10 Links

Our code will be stored in the following Github repository.

`https://github.com/xuetingkai/APS360AirQuality.git`

REFERENCES

V Athira, P. Geetha, R. Vinayakumar, and K. P. Soman. DeepAirNet: Applying recurrent networks for air quality prediction. *Procedia Computer Science*, 132:1394–1403, 2018.

Addis Bikis. Urban air pollution and greenness in relation to public health. *J Environ Public Health*, 2023:8516622, 2023.

Patrick C. Campbell, Youhua Tang, Pius Lee, Barry Baker, Daniel Tong, Rick Saylor, Ariel Stein, Jianping Huang, Ho-Chun Huang, Edward Strobach, Jeff McQueen, Li Pan, Ivanka Stajner, Jamese Sims, Jose Tirado-Delgado, Youngsun Jung, Fanglin Yang, Tanya L. Spero, and Robert C. Gilliam. Development and evaluation of an advanced national air quality forecasting capability using the NOAA Global Forecast System version 16. *Geoscientific Model Development*, 15: 3281–3313, 2022.

Yufei Huang, Chunyan Lu, Yifan Lei, Yue Su, Yanlin Su, and Zili Wang. Spatio-temporal variations of temperature and precipitation during 1951–2019 in arid and semiarid region, china. *Chinese Geographical Science*, 32:285–301, 2022.

Pu-Yun Kow, I-Wen Hsia Hsia, Li-Chiu Chang, and Fi-John Chang. Real-time image-based air quality estimation by deep learning neural networks. *Journal of Environmental Management*, 307:114560, 2022.

Yushun Mao and Shiejue Lee. Deep convolutional neural network for air quality prediction. *Journal of Physics: Conference Series*, 1302:032046, 2019.

Ruoxin Wang, Kangping Cui, Hwey-Lin Sheu, Lin-Chi Wang, and Xueyan Liu. Effects of precipitation on the air quality index, pm2.5 levels and on the dry deposition of pcdd/fs in the ambient air. *Aerosol and Air Quality Research*, 23:220417, 2023.

Jiaxuan Zhang and Shunyong Li. Air quality index forecast in Beijing based on CNN-LSTM multi-model. *Chemosphere*, 308:136180, 2022.

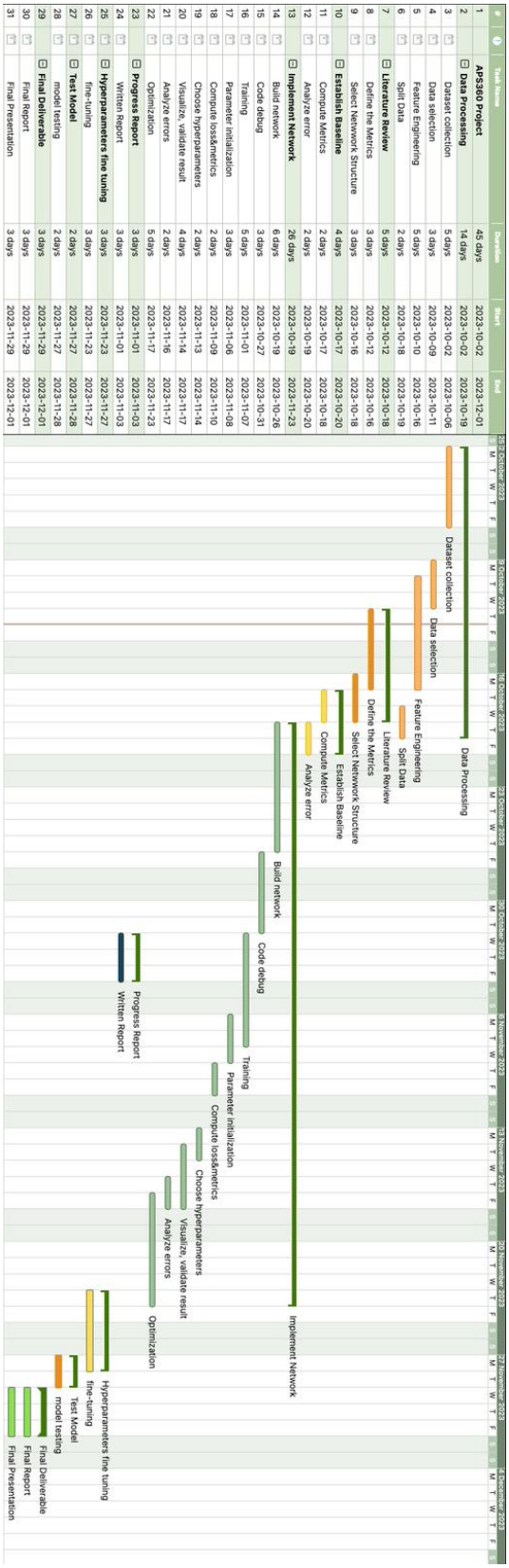

Figure 2: Gantt chart