**Air Quality Project**

# Introduction

Air pollution is a major global problem which has withering effects on human health, environment and economy. The World Health Organization (WHO) approximates close to the whole world inhales air that is polluted by the levels of contamination that surpass the perceived safe levels. It is a large issue and requires a good system of monitoring, learning, and most importantly, forecasting of air quality. Good forecasting of air quality may also be one of the tools that will be most useful in helping the health department authorities issue warnings to the citizens in good time to allow the legislators to take the right measures in regulating pollution and also to ensure that the citizens are ready to keep themselves safe. The non-linear interactions that exist between different pollutants and drivers may not be revealed using the traditional approaches of predicting air quality which rely on statistical or numeric models. When machine learning and deep learning are introduced, more complex and more accurate predictive models can be created. In this project, the objective is to develop a predictive system in such a way that it can also predict the future air quality level; however, it should provide some details about the causal relationship between the different variables.

# Data Source

In this project, we are going to utilize the UCI Air Quality Dataset. This data is especially suitable to our aims because it is a multivariate time-series dataset that includes hourly averaged responses of an array of metal oxide chemical sensors. It contains the ground truth concentrations of many pollutants such as Carbon Monoxide (CO), Non-Methane Hydrocarbons (NMHC), Benzene (C6H6 ), Nitrogen Oxides (NOx) and Nitrogen Dioxide (NO2) as well as meteorological parameters (e.g., Temperature, Relative Humidity, Absolute Humidity). Its time-based structure and the presence of several correlated variables make it an excellent time-series forecasting and causal inference dataset.

# Problem Definition and Literature Review

## Existing Research on Air Pollution Forecasting

The air quality forecasting has been advancing through the years, moving away from the previous statistical and numerical tools to the new machine learning and deep learning tools. Earlier models like ARIMA (Auto Regressive Integrated Moving Average) and other statistical approaches depending on time series used historical data patterns extensively to make predictions. They are good at simple linear changes, but non-linear relationships that are complex and prevalent with air quality data are frequently not well modeled using the method. On the other hand, numerical models are atmospheric physics and chemistry models, but are not only computationally demanding, but also highly sensitive in their input data. However, with the advent of machine learning algorithms, including Support Vector Machines (SVM) and Random Forests, a crucial breakthrough was achieved, as they can handle high-dimensional and non-linear data (Gautam, 2020). However, with sequential data like air quality time series, deep learning models have been the new state-of-the-art. Specifically, Recurrent Neural Networks (RNNs) and their variations, such as Long Short-Term Memory (LSTM), and Gated Recurrent Units (GRUs) can be adapted to this task with ease due to the ability to learn long-term temporal dependencies (S. Li et al., 2017; A. Z. Li et al., 2019). Hopefully, this project will be able to utilize the power of these powerful deep learning models to predict.

Among the most critical flaws of the majority of machine learning models, one should mention the propensity to focus on correlation, as opposed to causation. This is possible since a model might learn that variable A and variable B have a stronger or weaker relationship with each other but it cannot learn that either or both of those variables are the cause or their cause of a third unobserved variable C. This is amenable to new research in the field of causal inference, in particular Structural Causal Models (SCMs) and algorithms like Invariant Causal Prediction. With causal discovery integrated as part of our modeling framework, we will be in a position to surpass a simple forecasting model into a more robust and interpretive system that will have a deeper insight into underlying causes of air quality variations.

## Main Factors Influencing Air Quality

. Air quality is not a simple process because it is related to the interaction of meteorological and anthropogenic and natural factors. The weather conditions are the most common external forces. As an indicator, the speed and direction of the wind are pertinent variables in the distribution or concentration of the pollutants. The rate of chemical reaction in the atmosphere (formation of ground-level ozone) is influenced by temperature and the formation and deposition of pollutants may be influenced by humidity. Others are: solar radiation that is favorable to photochemical reactions. The short-term contributions of pollutants, together with the weather are the source of emissions. These mainly anthropogenic, due to vehicle exhaust, industry and energy generation. More natural sources such as wildfires and dust storms also can play an important role, but they are not as common. The UCI Air Quality dataset is perfectly applicable to our research study since it contains the concentration of pollutants and critical meteorological parameters, which will enable us to explore the complex relations between them.

# Data Cleaning

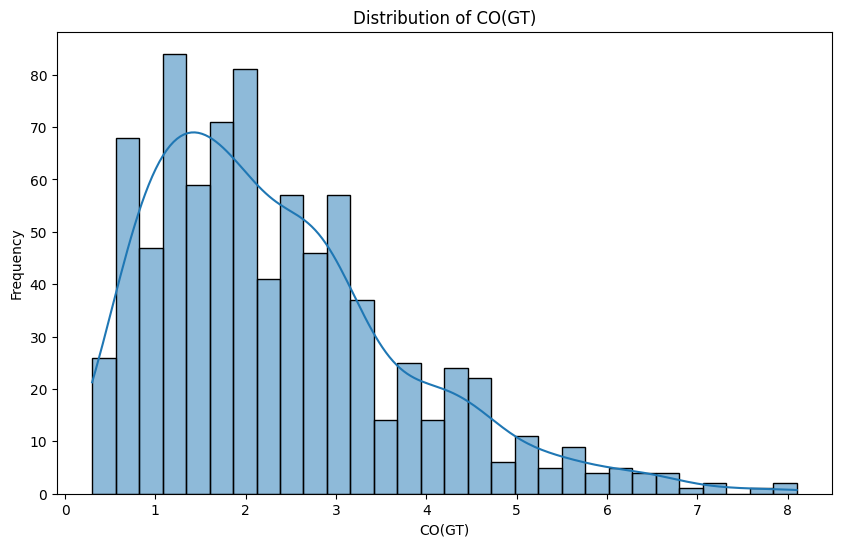
The raw Air Quality data provided by the UCI Machine Learning Repository were subjected to preprocessing steps to enable its use in time series forecasting, machine learning, and causal analysis. Firstly, the delimiter was used with the correct value to ensure correct separation of the variables and the Date and Time fields were combined to create one single datetime field to create a chronological order. Pollutant concentrations, sensor readings, and environmental variables were converted into numbers, and NaN was used to treat missing data as -200, which is turned into NaN in the course of subsequent manipulations. To minimize bias, duplicate records and extreme outliers were removed and cleaned the data were saved in a standard format.

# Exploratory Data Analysis

Exploratory Data Analysis was conducted to better understand the distribution, correlations, and temporal patterns of the air quality variables.

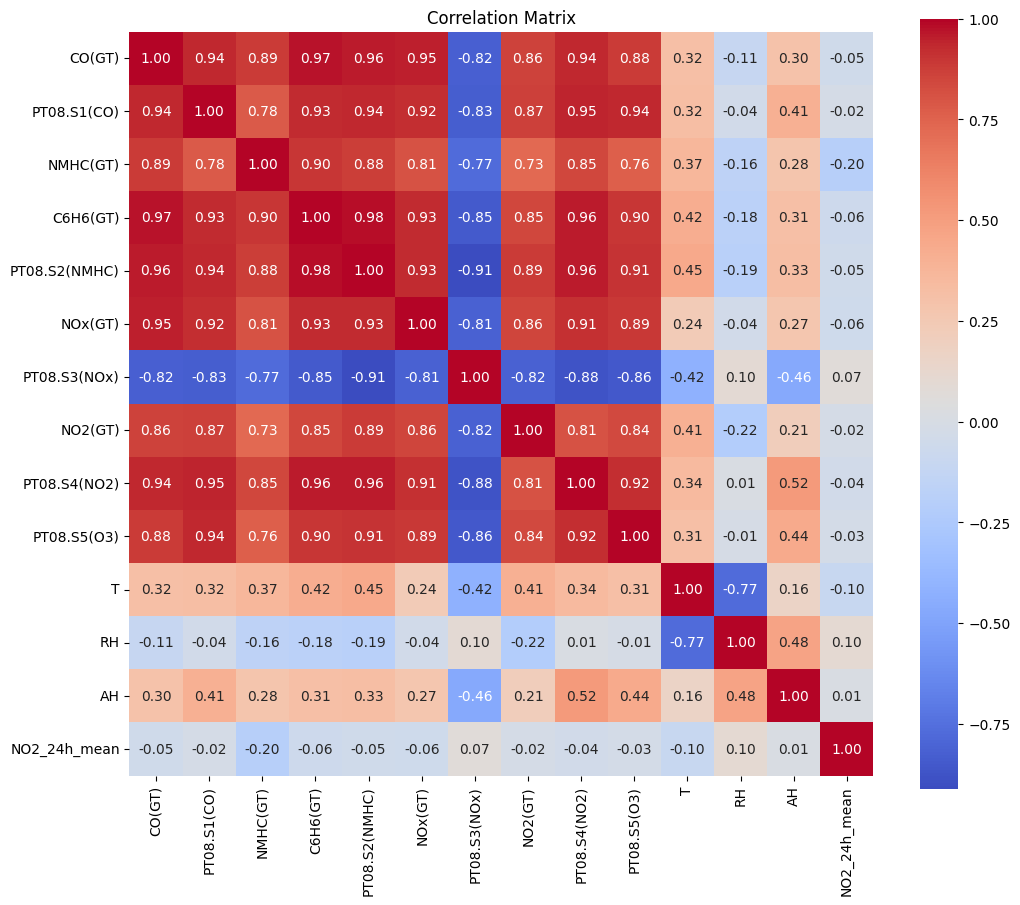
Firstly, the distribution of the frequency of carbon monoxide (CO) levels in the data was represented in a histogram as shown in Figure 1. The concentration ranges (maximum and minimum) of CO are between 1 and 3mg/m 3 with skewed distribution to the right meaning that there are some higher concentration extremes. This implies that low to moderate levels of CO are normal, but there exist times of high pollution, which require close scrutiny.

Figure 1 Distribution of CO(GT)



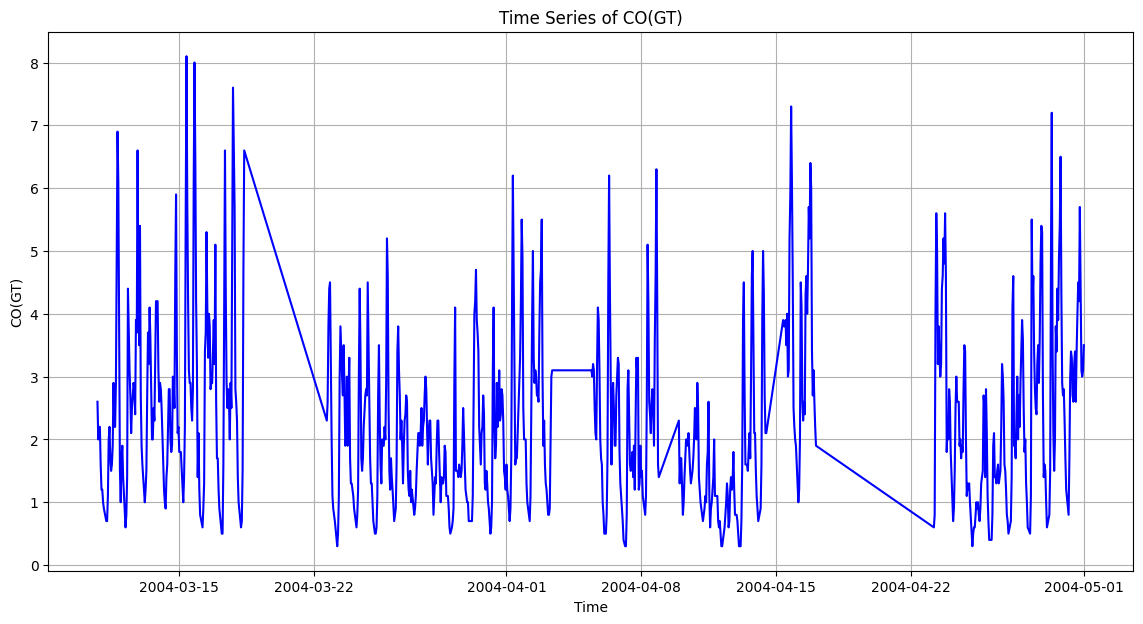
Secondly, the analysis of correlation (Figure 2) demonstrates that many pollutants, i.e., CO, NMHC, C6H6, and NOx, have strong positive correlations, which can be explained by the fact that most of these emissions are produced by similar sources, such as automobile traffic, and industry. On the other hand, certain negative relationships are found between the readings of the gas sensor like PT08.S3 and the concentration of the pollutants, which indicate that some inverse response of the sensor. Such environmental factors as temperature (T), relative humidity (RH), and absolute humidity (AH) indicate lower correlations with pollutants, which indicate their indirect role in determining the quality of the air.

Figure 2 Correlation Matrix



The time series plot (Figure 3) provides time variation in the CO concentrations during the time of observation. Severe peaks reflect pollution peaks that can be related to traffic jams, industries or unfavorable weather patterns. The uneven variability brings out the importance of having predictive modeling to capture the short term and long term variations of air quality trends.

Figure 3 Time series plot



# Methodology

## Data Preparation

The dataset was initially normalized through dealing with the missing values and feature scaling with MinMaxScaler. In the case of time-series analysis, the variable of interest CO(GT) was scaled between 0 and 1, and sequence of 24 hourly observations were prepared as the input to estimate the next hour concentration. The data was then divided into training and testing data sets in the ratio of 80% and 20% respectively. To conduct a causal analysis, a group of the most significant variables were chosen which included CO (GT), NOx(GT), C6H6(GT), temperature (T), and relative humidity (RH). These variables were further categorized in treatment (T), outcome (CO(GT)) and covariates (X) to organize the causal estimation framework. The names of the features were stored to the preprocessing pipeline in case of reproducibility during deployment.

## Causal Discovery

Three complementary frameworks were used to perform the causal discovery. Initially, a PCMCI ( Peter- Clark Momentary Conditional Independence ) approach was conducted with a maximum lag of six that identified the time dependency among air quality variables and generated a lagged causal graph. Second, LiNGAM (Linear Non-Gaussian Acyclic Model) algorithm was used to analyze the current data, which resulted in a directed acyclic graph (DAG) and it demonstrated causal edges with weights. Third, a Structural Causal Model (SCM) was constructed using DoWhy, with temperature as the treatment, CO(GT) the outcome and RH and NOx(GT) confounders. This gave an interpretable DAG and quantitative causal effect estimates. The insights provided by these methods were used to select variables used in the final causal forest mode.

## Causal Forest Estimation

Implementation of the Causal Forest model was done by EconML through CausalForestDML. The treatment and outcome models were configured to use random Forest regressors, with the following hyperparameters; 100 estimators, minimum 10 samples per leaf, three-fold cross-validation and constant random seed to ensure reproducibility. The model estimated the Conditional Average Treatment Effect (CATE) of the individual observations and the Average Treatment Effect (ATE) with confidence intervals, which provide a strong causal understanding of the effect of temperature on concentrations of pollutants.

## Deep Learning Models

Furthermore, deep learning models were trained to learn temporal behavior in the data. The network was trained on sequential pollutant data and had an architecture of a single LSTM layer of 50 units and a dense layer. The model was optimized using Adam and trained on mean squared error loss over 50 epochs. In PyTorch, a Temporal Convolutional Network (TCN) was proposed, with dilated causal convolutions and residual connections to model long-range temporal dependencies. The final layer was fully interconnected and produced the forecast of the next time step.

## Hybrid Model

The hybrid model came up as a result of combining the causal forests with deep learning. This allowed the neural network to enable a trade-off between the causal interpretability and the prediction accuracy by incorporating CATE estimates as independent features. Performance measures of the model on the test set were used to balance predictive accuracy and explanatory information by using the RMSE and R2.

# Results

Three deep learning models namely LSTM, Temporal Convolutional Network (TCN) and Transformer have been tried. The LSTM model also worked well with a MSE of 0.621, MAE of 0.587 and an R of 0.681 and therefore, it was able to fit the temporal dependencies. The Transformer model was also working; with the RMSE of 0.884, MAE of 0.645 and R2 of 0.613, which proved to be competitive in the sequential dynamics modeling. Conversely, TCN model was poor in its performance and RMSE of 13.81, MAE of 10.55 and a negative R 2 value (-93.29) means that it failed to generalize to the inner structure of the data.

The PCMCI model showed that there existed notable causes between the most important pollutants and meteorological variables. As an example, CO(GT) was greatly affected by C6H6(GT) (val = 0.958, p < 0.001) and NOx(GT) (val = 0.901, p < 0.001), same time and with short time lags. Temperature (T) showed negative correlation with CO(GT), both at the present time step and with lagged values, indicating that an increase in temperatures is likely to lower carbon monoxide. Likewise, relative humidity (RH) was found to correlate positively with CO(GT) and NOx(GT), which indicates the influence of the weather on the concentration of air pollutants.

The Causal Forest model was used to measure the causal impact of temperature on concentration of pollutants. The Mean Conditional Average Treatment Effect (CATE) was estimated to be negative ( = -0.0204), and the confidence intervals of the first five test samples were all negative, thus showing that higher temperatures cause a decrease in the CO(GT) levels. The Average Treatment Effect (ATE) was also determined as: -0.0204 (95% CI = -0.0396 -0.0012), giving additional evidence to the negative causal impact of temperature effect.

Finally, the hybrid model that employed the estimates of CATE obtained by the causal forest, as well as the characteristics of deep learning displayed the strongest predictive accuracy. It achieved RMSE = 0.312, R2 = 0.953, the best among single LSTM and Transformer models. It demonstrates that neural networks augmented with causal knowledge significantly increase predictive accuracy and, besides, ensure interpretability. It is not only that the hybrid model is more accurate in the predictions of the level of pollution, but also that it provides causal explanations of the contribution of the environmental factors by bridging gaps between the black-box prediction and the provide explanations.

# Discussion

This paper demonstrates that a combination of deep learning and causal inference is useful in modeling how air pollution dynamics. Despite the LSTM and Transformer models being able to model any temporal relationships and make valid forecasts, the TCN model did not work well, likely due to being not supported adequately with this dataset. Causal discovery methods such as PCMCI and LiNGAM have shown that pollutants (e.g., CO(GT), NOx(GT), C6H6(GT)) were highly interdependent and that meteorological parameters (temperature and humidity) had important roles in the interaction of the underlying environmental and chemical processes. These results were also determined in the Causal Forest analysis that revealed temperature to have negative causal impact on the uniform CO(GT) concentrations that are consistent with environmental theory on the dispersion of pollutants. It is noteworthy that the hybrid deep-learning model, achieved by combining the causal estimates, had a significantly higher predictive accuracy (R 2 = 0.953 ) indicating that the integration of interpretability and predictive power is a more cohesive and effective view of predicting the air quality.

# Conclusion

This article shows that a hybrid method that integrates causal inference and deep learning can provide both high predictive and meaningful interpretability to air pollution analysis. The research determined that the LSTM and Transformer models can be useful in prediction, whereas causal discovery and estimation models can predetermine further insight into the origin of the pollution, especially the major weather conditions influence. The hybrid method was more effective than the individual predictive methods and provided viable information to support management of the environment according to the elements of causality in predictive models. This means that causal deep-learning models can be a useful resource to predict, respond to and monitor air quality in real-time, and the future research needs to focus on growing data sets, developing more hybrid approaches, and treating the framework to a broader group of locations to have a greater impact.

# References

Gautam, A. K., & Gautam, P. (2020). Air Pollution Forecasting Using Machine Learning Techniques: A Survey. Journal of Physics: Conference Series, 1478(1), 012028. doi:10.1088/1742-6596/1478/1/012028

Li, S., Li, Y., & Gao, F. (2017). Deep recurrent neural network for air pollution forecasting: A case study in Beijing. International Journal of Environmental Research and Public Health, 14(11), 1385. doi:10.3390/ijerph14111385

Li, A. Z., Li, J., & Wang, J. (2019). An air pollution prediction model based on LSTM with attention mechanism. Journal of Cleaner Production, 239, 118042. doi:10.1016/j.jclepro.2019.118042

World Health Organization. (n.d.). Ambient (outdoor) air pollution. Retrieved from https://www.who.int/news-room/fact-sheets/detail/ambient-(outdoor)-air-pollution