University of Milan-Bicocca
Department of Informatics, Systems and Communication
Master's Degree in Data Science

Study of key Deep Learning techniques applied to Open Data for air quality monitoring in Smart Cities

Candidate:
Matteo Lanzillotti
843283

Supervisor:
Prof. Simone Bianco

Co-Supervisors:
Prof. Paolo Napoletano
Dr. Luigi Celona

Air pollution in Urban Spaces

PROBLEM

Outdoor air pollution is the estimated cause of **4.2 million premature deaths** worldwide in 2019 (Source: World Health Organization [1]).

Some of the most dangerous pollutants in urban areas are:

- Particulate matter ($PM_{2.5}$ & PM_{10})
- Ground-level Ozone (O_3)
- Sulfure Dioxide (SO_2)
- Carbon Monoxide (CO)
- Nitrogen Dioxide (NO_2)


L'EMERGENZA

Allarme smog: Milano è la terza città peggiore al mondo per inquinamento

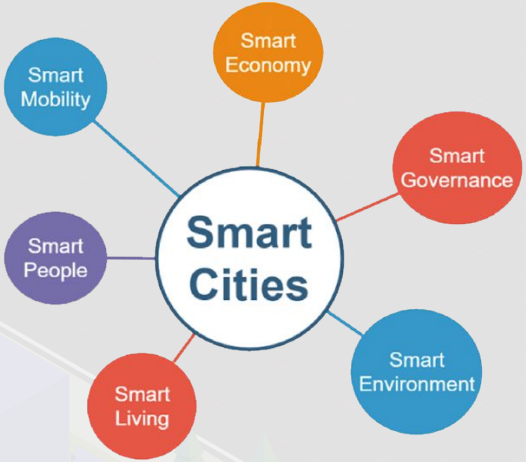
Non accennano a calare i livelli di inquinanti nell'aria di Milano. A spaventare sono i dati emersi dalla rilevazione dell'indice Aqi

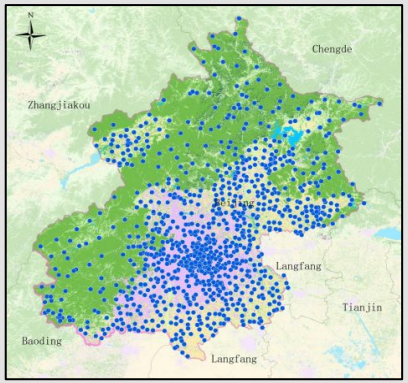
These pollutants can cause damage to the respiratory and cardiovascular systems, especially in sensitive people.

Smart City

INTRODUCTION

- *Citizen-centric town*
- *Sustainable and resilient*
- *Balances innovation and future needs*





Capillary monitoring of air quality is possible thanks to the large number of sensors employed in smart cities.

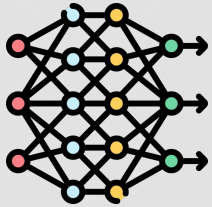
Task definition

CONTRIBUTION

Input data

(24 h x 7 days, $n_{features}$)

Deep learning model



Output data

(24 h, $n_{pollutants}$)

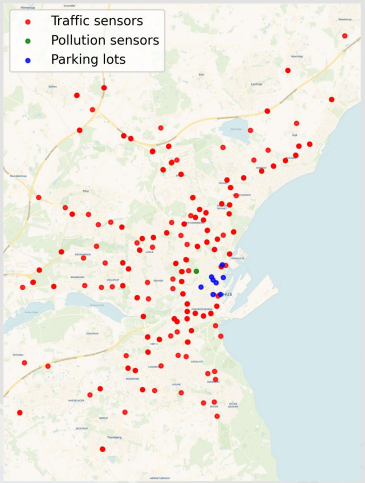
Main challenges:

1. Missing or incorrect data
2. Collinearity between dependent variables
3. Complex and multiple seasonal patterns

Datasets (1)

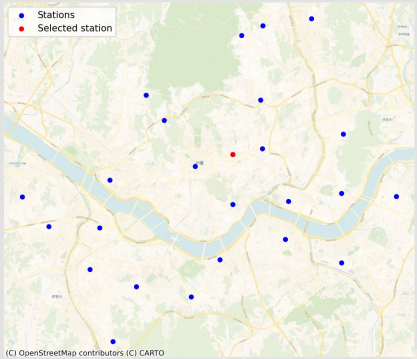
Public datasets have been used for this projects:

Citypulse



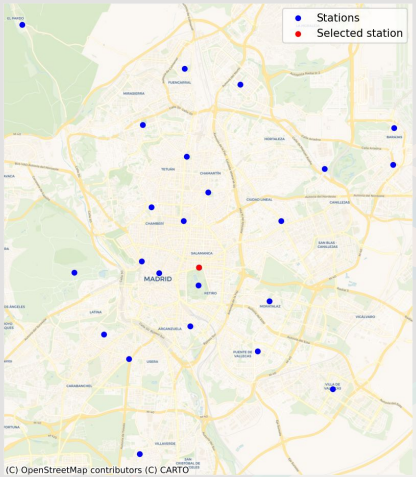
Source: Citypulse dataset collection [2]

Seoul



Source: Kaggle[3]

Madrid



Source: Kaggle [4]

Datasets (2)

Public datasets have been used for this project:

Dataset	Duration	Available Pollutants						External features		
		O3	PM10	PM2.5	NO2	SO2	CO	Traffic	Parkings	Weather
Citypulse	3 Months Aug – Nov 2014	✓	✓	✓	✓			1	1	7
Seoul	3 Years 2017 - 2019	✓	✓	✓	✓	✓	✓	0	0	7
Madrid	3 Years 2015 - 2019	✓	✓	✓	✓	✓	✓	0	0	10

Data preparation

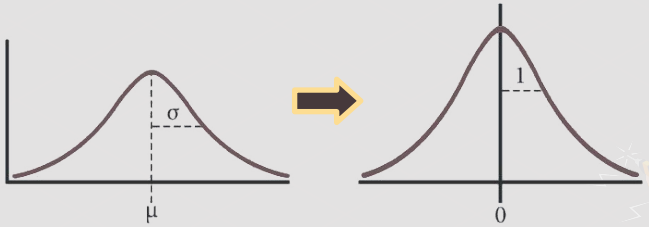
TECHNIQUES

Data Splitting

- 80% Train
- 10% Validation
- 10% Test

Data Scaling

- Z-score normalization



Feature engineering

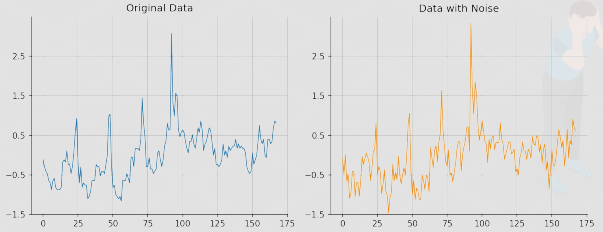
- Creation of time based features

Date
23/02/2024 11:00

Day of week	Month	Hour
Friday	February	11

Data augmentation

- Time series jittering



Proposed & implemented models

Proposed models:

- **Linear model**
- LSTM model
- Bi-LSTM Model
- Dense Encoder-Decoder Model
- CONV-LSTM
- Wavelet Model

Implemented model:

- TSMixer

- Very simple, lot of parameters
- Used for benchmark

Proposed & implemented models

Proposed models:

- Linear model
- **LSTM model**
- Bi-LSTM Model
- Dense Encoder-Decoder Model
- CONV-LSTM
- Wavelet Model

Implemented model:

- TSMixer

- Concatenation of two LSTM layers

Proposed & implemented models

Proposed models:

- Linear model
- LSTM model
- **Bi-LSTM Model**
- Dense Encoder-Decoder Model
- CONV-LSTM
- Wavelet Model

Implemented model:

- TSMixer

- Similar to LSTM Model, implements Bidirectional processing in LSTM layers

Proposed & implemented models

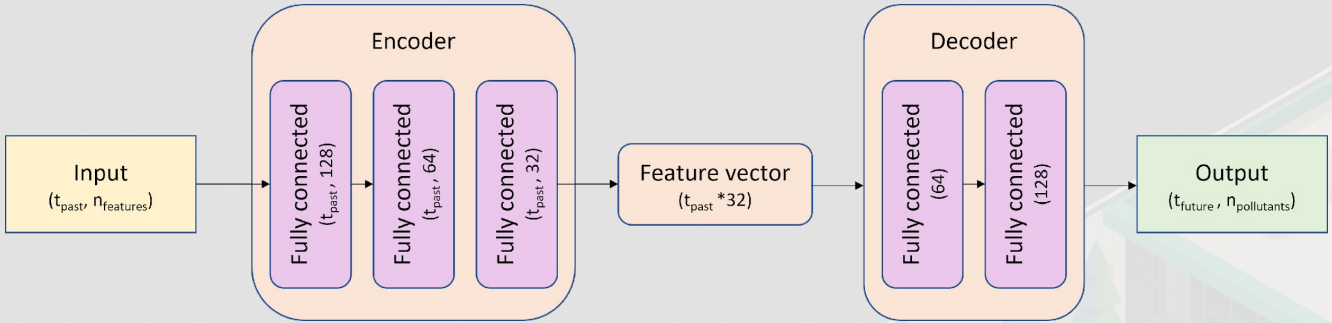
Proposed models:

- Linear model
- LSTM model
- Bi-LSTM Model
- **Dense Encoder-Decoder Model**
- CONV-LSTM
- Wavelet Model

Implemented model:

- TSMixer

- Use of two MLPs to process input sequence
- Encoder: Embed the input data on a feature vector
- Decoder: decodes feature vector to forecast future values



Proposed & implemented models

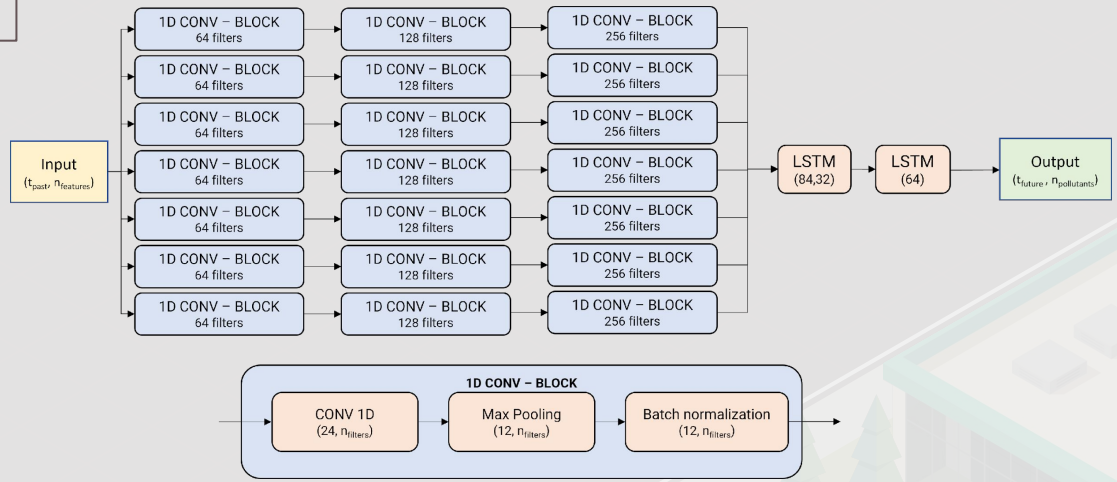
Proposed models:

- Linear model
- LSTM model
- Bi-LSTM Model
- Dense Encoder-Decoder Model
- **CONV-LSTM**
- Wavelet Model

Implemented model:

- TSMixer

- Process separatedly the 7 days of data through convolutional 1D layers, then feeds features to LSTM layers.



Proposed & implemented models

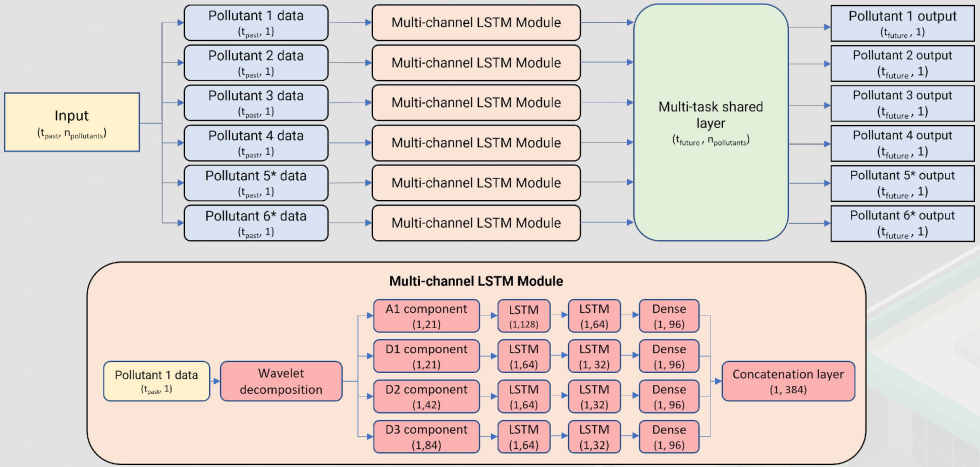
Proposed models:

- Linear model
- LSTM model
- Bi-LSTM Model
- Dense Encoder-Decoder Model
- CONV-LSTM
- **Wavelet Model**

Implemented model:

- TSMixer

- Split each pollutant sequence in 4 components via Wavelet transform
- Process each component separately through two LSTM layers, then concatenate the features in a shared layer to make predictions



Proposed & implemented models

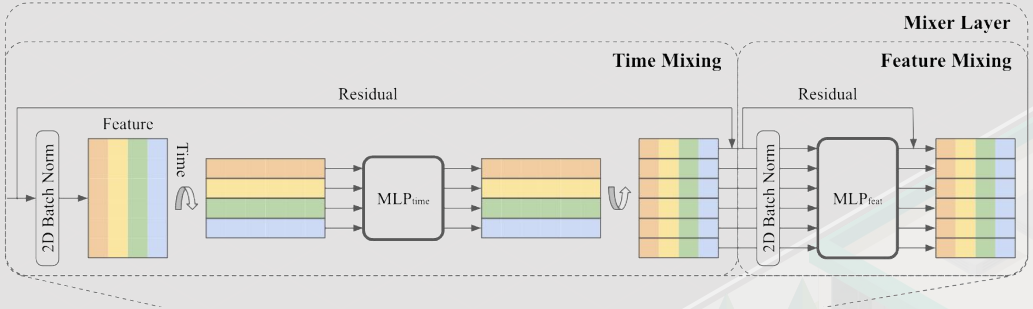
Proposed models:

- Linear model
- LSTM model
- Bi-LSTM Model
- Dense Encoder-Decoder Model
- CONV-LSTM
- Wavelet Model

Implemented model:

- **TSMixer**

- Proposed by Google in 2023 [5]
- MLP based architecture
- Uses matrix transposition to process data on both the time and feature axes



Experiments

Metrics

- Mean Absolute Error (MAE)
- Symmetric Mean Absolute Percentage Error (sMAPE)

Training configuration

- Loss: Mean Squared Error
- Optimizer ADAM (learning rate = 0.001)
- Early Stopping
- Reduce learning rate on plateau

Performed tests

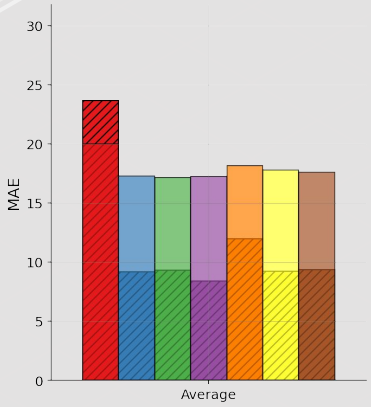
- Examination of the performance of the models on different subsets of the datasets

Dataset	Time horizon	Original Data	Data with Noise
Citypulse	3 Months	✓	✓
Seoul	1 year	✓	
	3 years	✓	✓
Madrid	1 year	✓	
	3 years	✓	✓

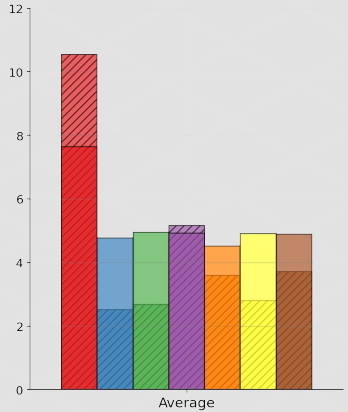
Performances on Test set

RESULTS

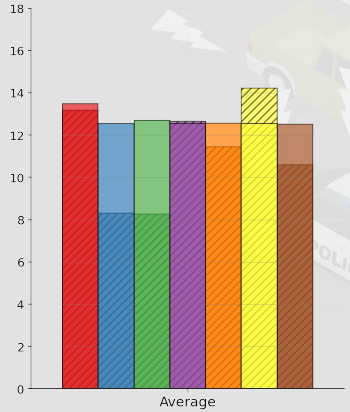
Citypulse

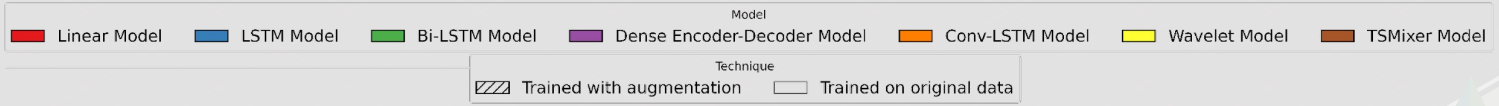


Seoul



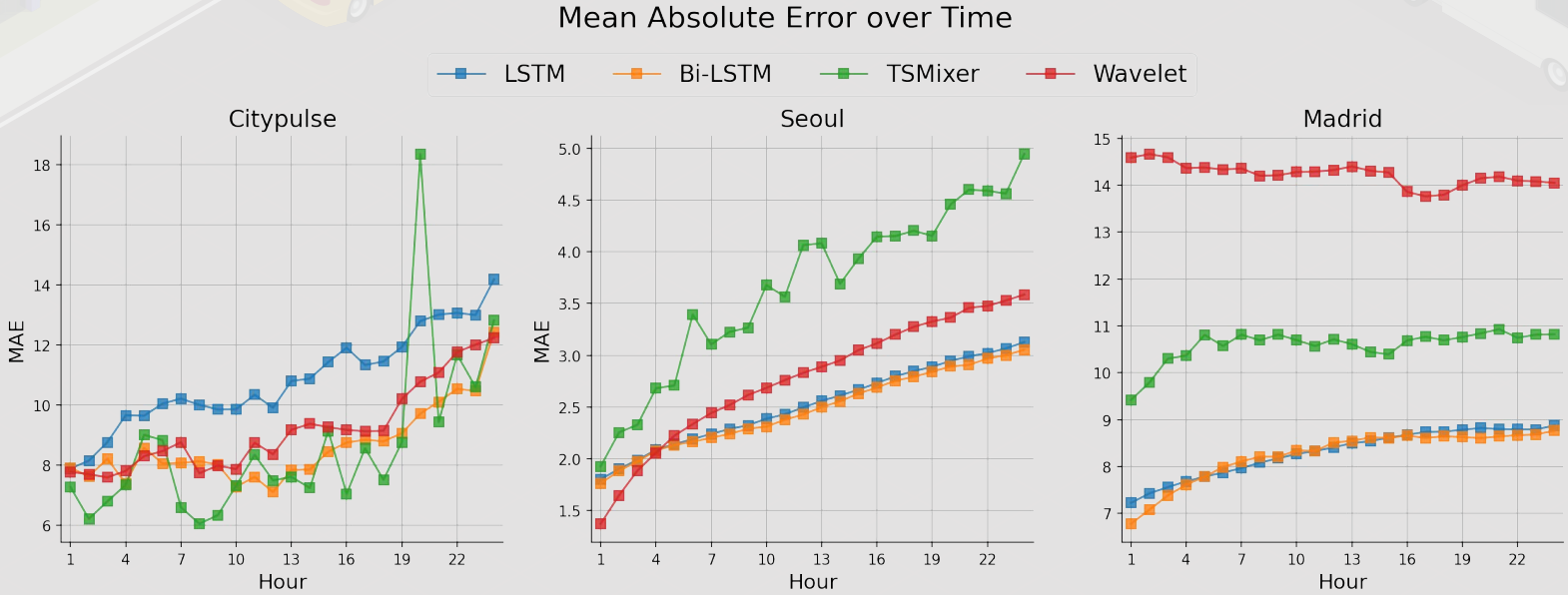
Madrid





- The most consistent models in the three datasets are LSTM, Bi-LSTM, Wavelet, and TSMixer models.
- Augmentation considerably improves the results in most cases.

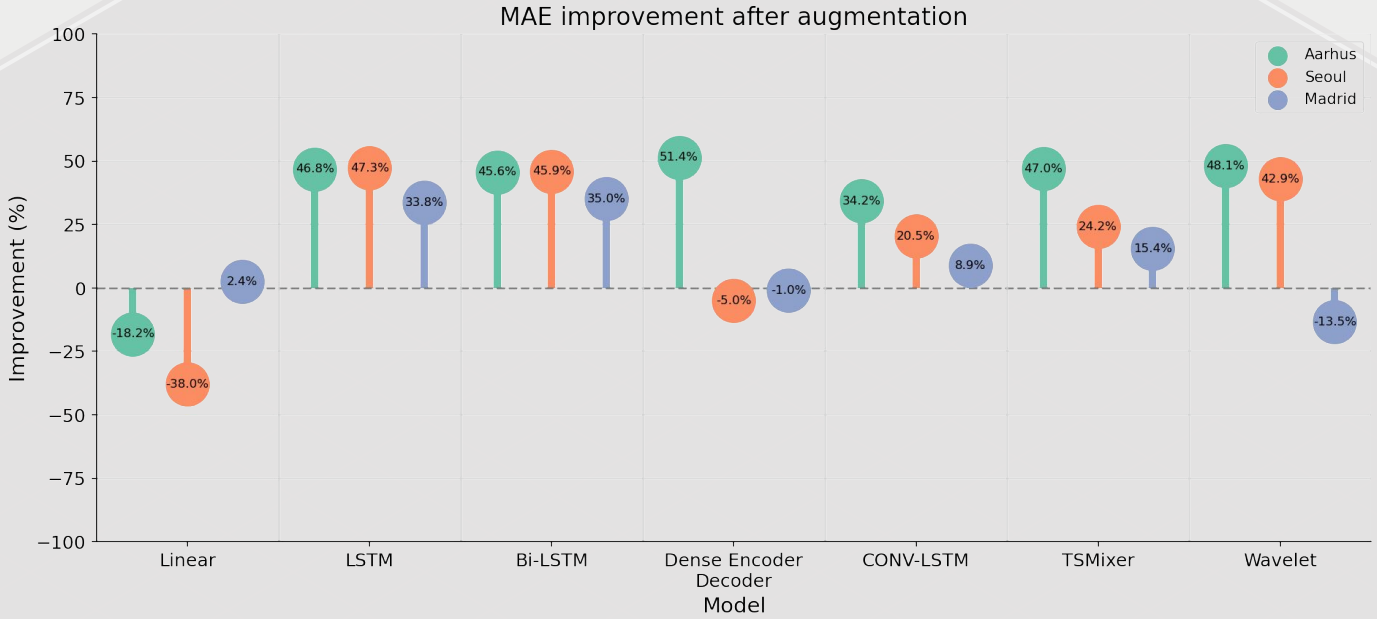
Error over time



- LSTM and Bi-LSTM models are the most reliable in the three datasets.

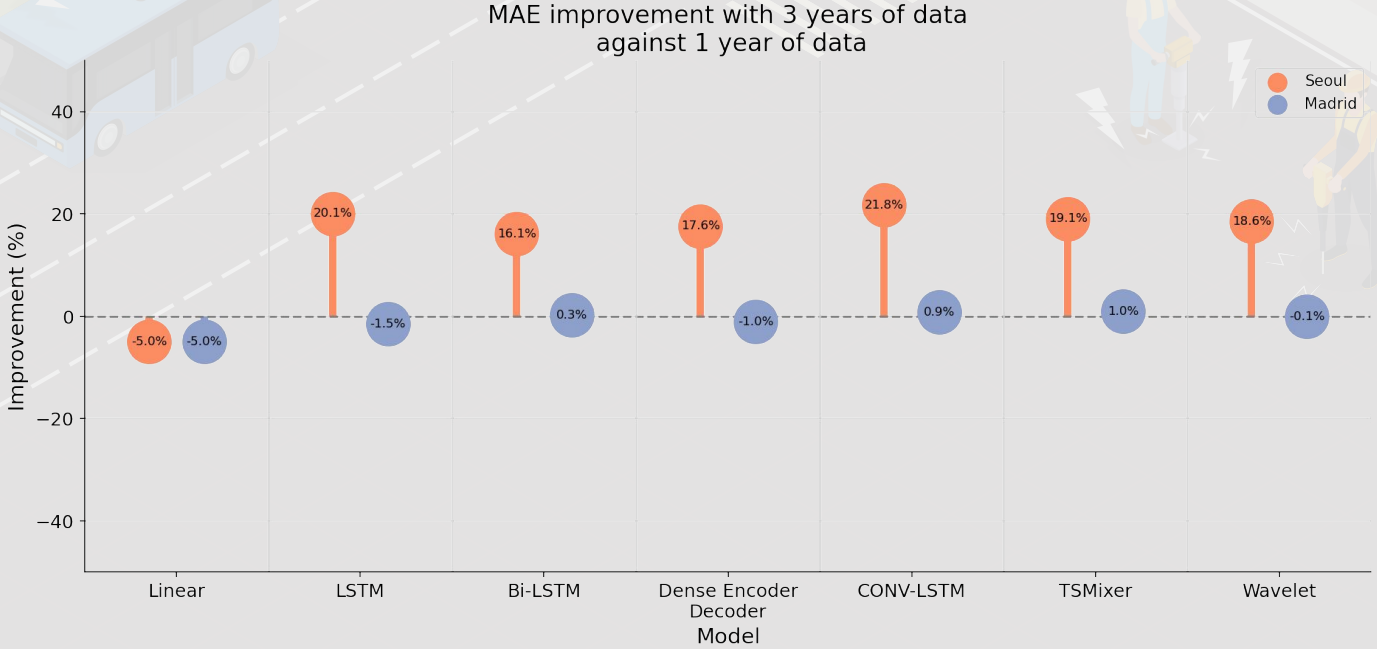
Data augmentation benefits

RESULTS



Here the usefulness of **data augmentation** in the training process is confirmed, as it **dramatically improves results** in some cases.

Historical data volume influence



- Complex models such as **CONV-LSTM** are affected by the larger volume of training data.
- **Data quality** in the Madrid dataset is **questionable**, as no improvement is seen despite the huge increase in data volume (**3x**).

Final considerations

CONCLUSIONS

- Models that take into account the temporal dimension have superior predictive performance:
  - Implementing **memory mechanisms**: LSTM, Bi-LSTM and Wavelet
  - Deriving **features from the temporal dimension**: TSMixer
- **Data augmentation** is a powerful technique, both during the training phase to **prevent overfitting**, both in terms of **error**.
- **Data collection should be conducted rigorously**, as demonstrated by the results obtained from the Seoul dataset, which has the least amount of missing data.

Future developments

- Experimentation with various configurations in the model's architecture is planned, **testing deeper or simpler networks**, specifically in LSTM, Bi-LSTM, and TSMixer networks.
- Utilization of **advanced augmentation strategies** for addressing data variability.
- Utilizing spatial data presents a significant opportunity to enhance air pollution forecast accuracy, for example using **advanced spatial modeling techniques**, such as Spatio-Temporal Graph Neural Networks, improve predictions by considering influences from surrounding areas.
- Increase **interpretability** for real-world scenarios to gain confidence in predictions.

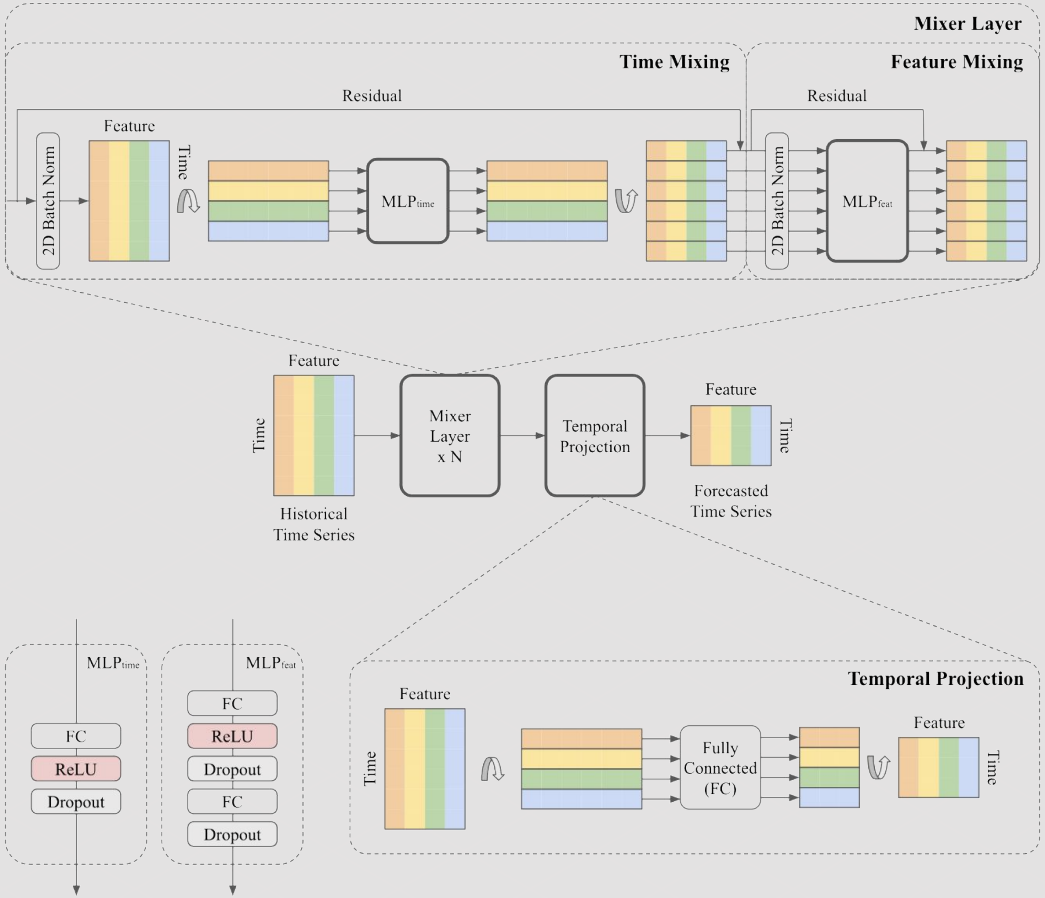

THANKS FOR THE ATTENTION

Short Bibliography

1. World Health Organization (WHO). Ambient (Outdoor) Air Pollution. Dec. 2022. URL
2. CityPulse EU FP7 Project. CityPulse Dataset Collection. URL
3. Kaggle - Air Pollution in Seoul. URL
4. Kaggle - Air Quality in Madrid (2001-2018). URL
5. CCAC Secretariat. Beijing's air quality improvements are a model for other cities. Mar. 2019. URL
6. Si-An Chen et al. "TSMixer: An All-MLP Architecture for Time Series Forecasting". In: (2023). arXiv: 2303.06053 [cs.LG].
7. G.E.P. Box and G.M. Jenkins. "Time Series Analysis: Forecasting and Control". In: Holden-Day series in time series analysis and digital processing (1970).

TSMixer architecture

EXTRA – MODEL ARCHITECTURES

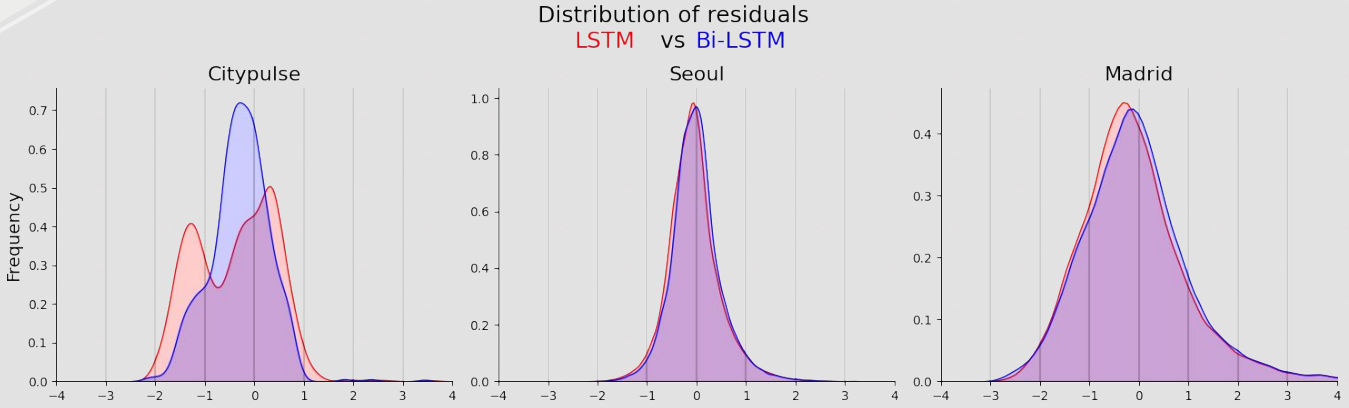


Baseline comparison

Model / Dataset	Citypulse	Seoul	Madrid
ARIMA [7]	17.41	6.92	10.32
LSTM	9.18	2.51	8.31
Bi-LSTM	9.33	2.68	8.25
TSMixer [5]	9.34	3.71	10.58
Wavelet	9.23	2.80	14.23

- Models demonstrate superior performance relative to the baseline in Citypulse and Seoul datasets.
- In Madrid dataset, TSMixer and Wavelet models fails to surpass the baseline

Residuals distribution

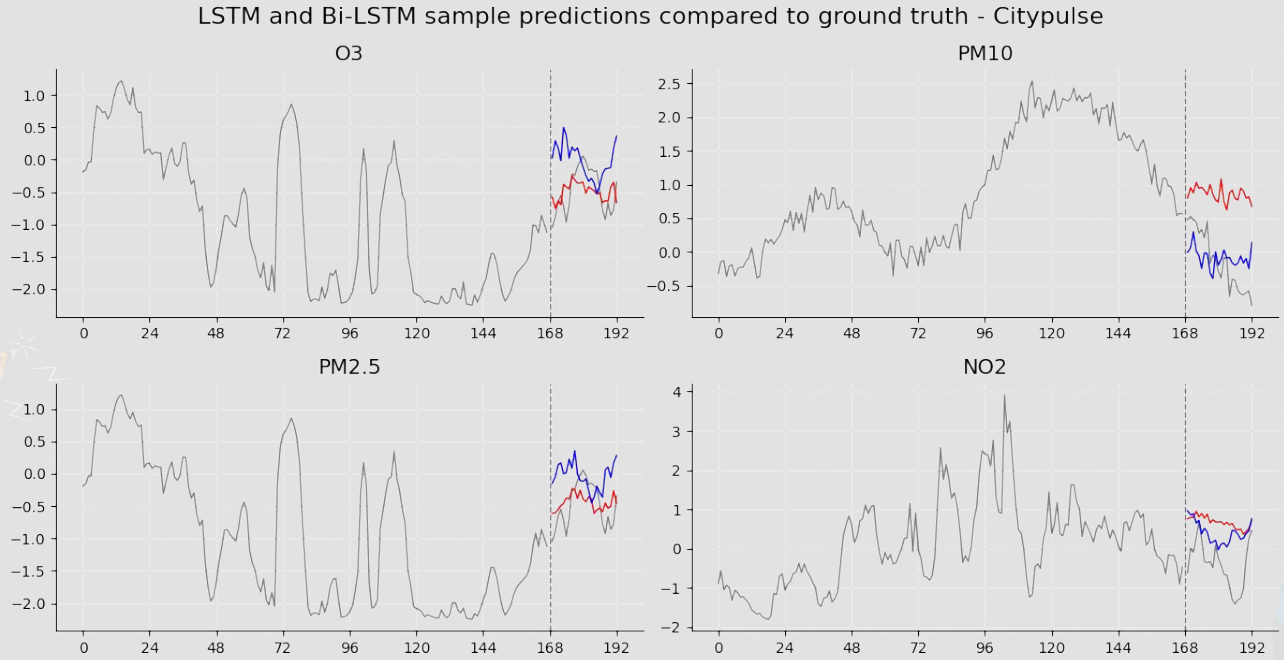


- Residuals of the LSTM model predictions in Citypulse have a strange distribution and makes us think about the confidence in MAE.
- Madrid has a wider distribution in residuals for both models.

Qualitative results

EXTRA - RESULTS

Citypulse | Seoul | Madrid

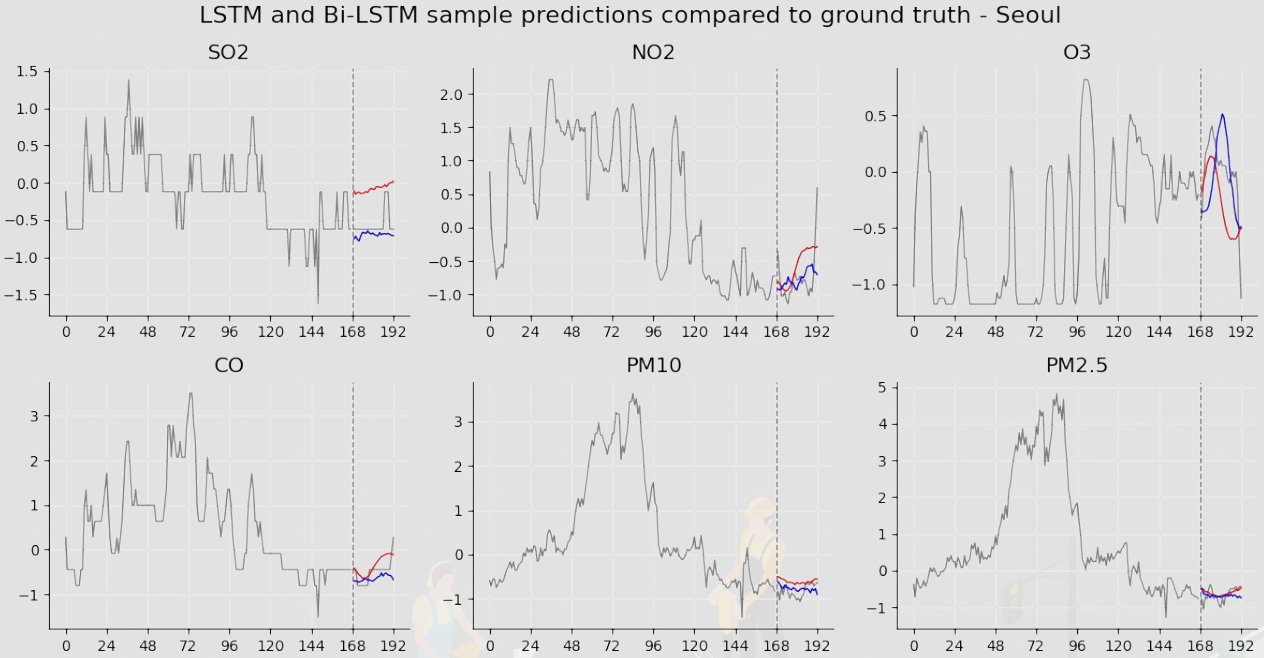


Both models forecasts are displayed, LSTM and Bi-LSTM

Qualitative results

EXTRA - RESULTS

Citypulse | **Seoul** | Madrid

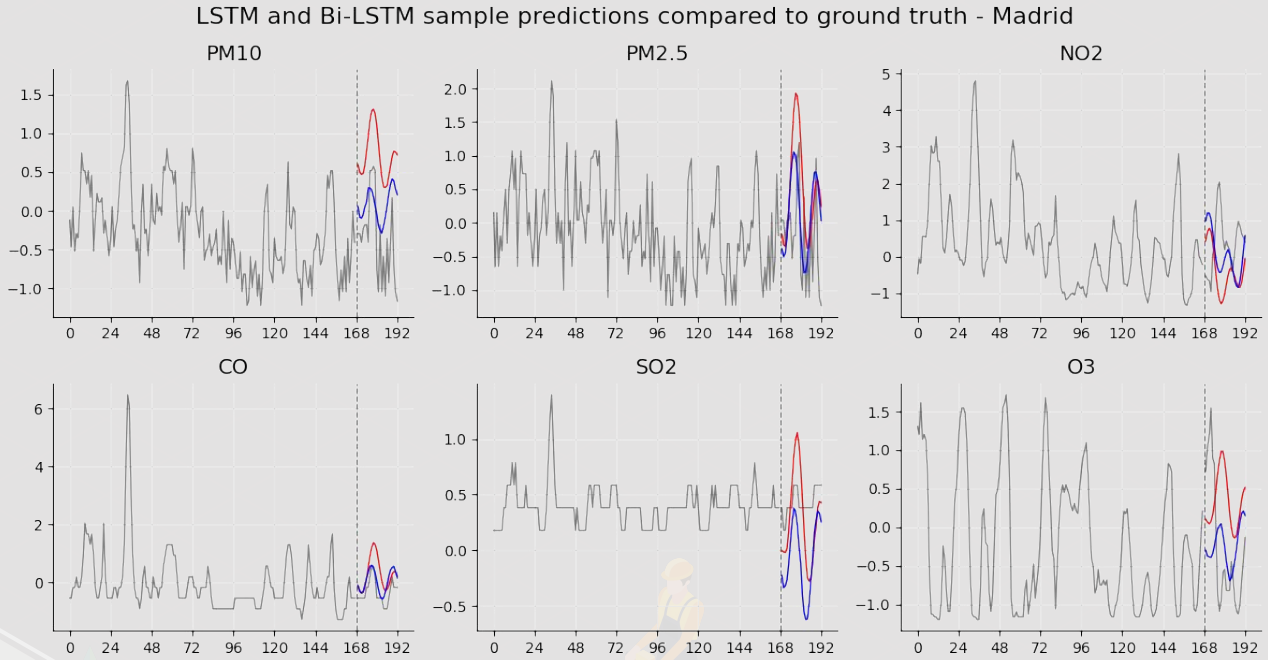


Both models forecasts are displayed, LSTM and Bi-LSTM

Qualitative results

EXTRA - RESULTS

Citypulse | Seoul | **Madrid**



Both models forecasts are displayed, LSTM and Bi-LSTM