Prädiktion von Feinstaubdaten in Polen

Projektpräsentation

Inhalt

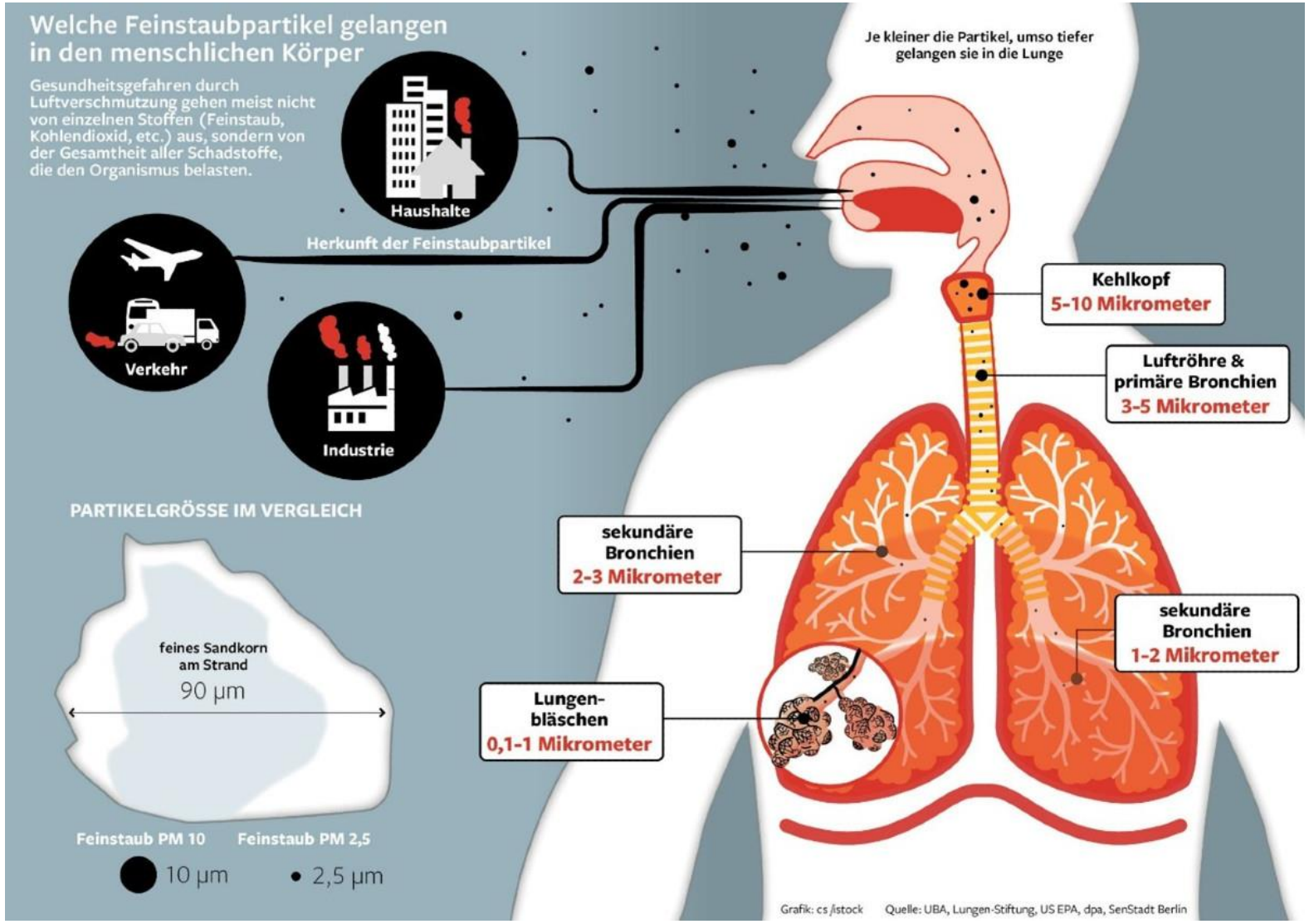
- Vorstellung der Gruppenmitglieder
- Motivation & interdisziplinäre Grundlagen
- Forschungsfragen
- Datenbeschaffung
- Vorverarbeitung
- Feature Engineering
- Data Windowing
- Neuronales Netz
- Ergebnisse
- Softwaredemo
- Zusammenfassung & Ausblick

Gruppenmitglieder

- Teamaufteilung:
  - Team Datenbeschaffung
    - Niklas Theis, Sarah Flohr, Niklas Lange
  - Team Machine Learning Grundlagen
    - Tobias Kaps, Svea Worms
  - Team Machine Learning Experimente
    - Niklas Theis, Sarah Flohr, Niklas Lange, Tobias Kaps, Svea Worms

Feinstaub

- Gesundheitsrisiko für den Menschen
- Mindestens 238.000 vorzeitige Todesfälle in der EU im Jahr 2020



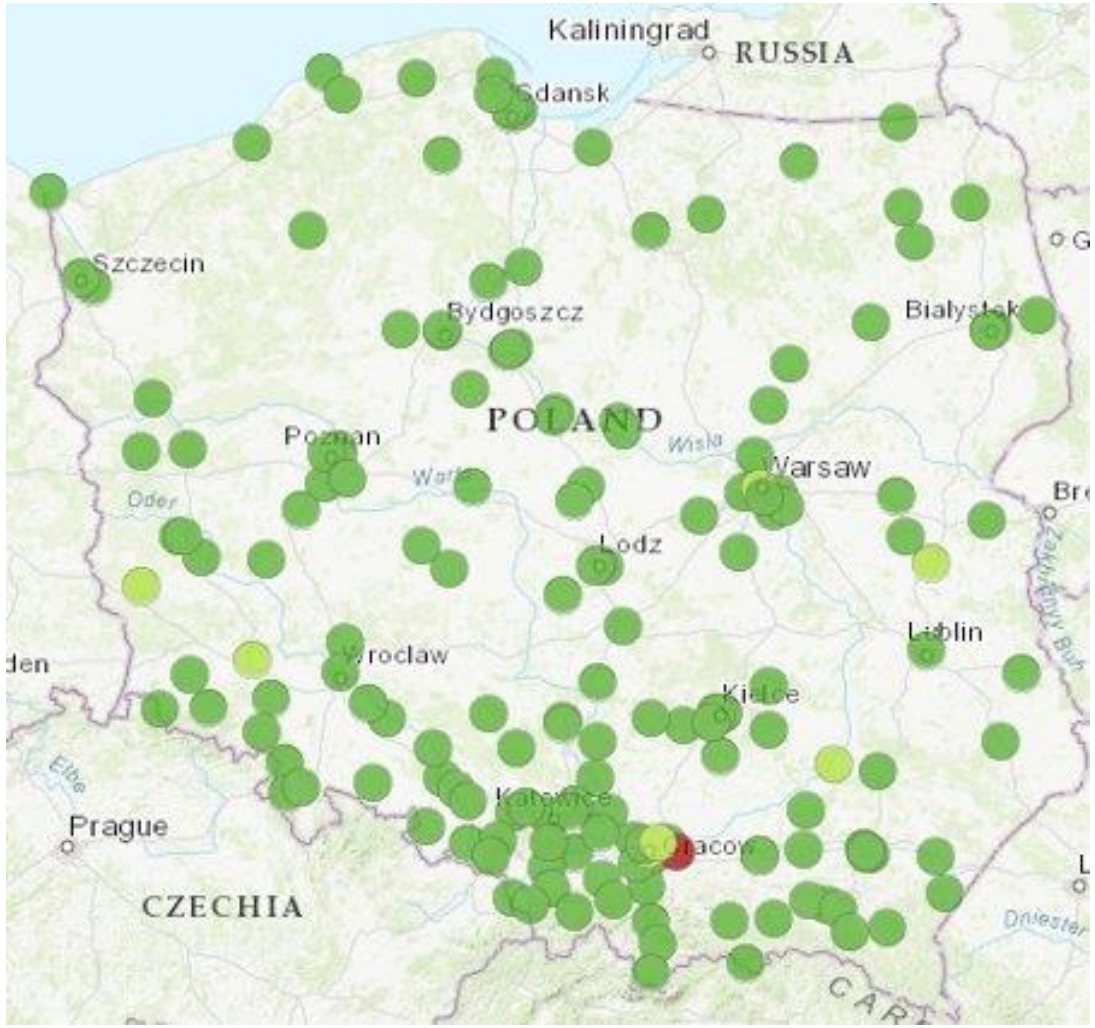
Feinstaub

- Feinstaub wird nach Größe unterteilt
  - PM_{10} bezeichnet Partikel mit einem Durchmesser < 10 µm
  - $PM_{2.5}$ ist eine Teilmenge von PM_{10} mit Partikel von einem Durchmesser < 2.5 µm

Grenzwerte

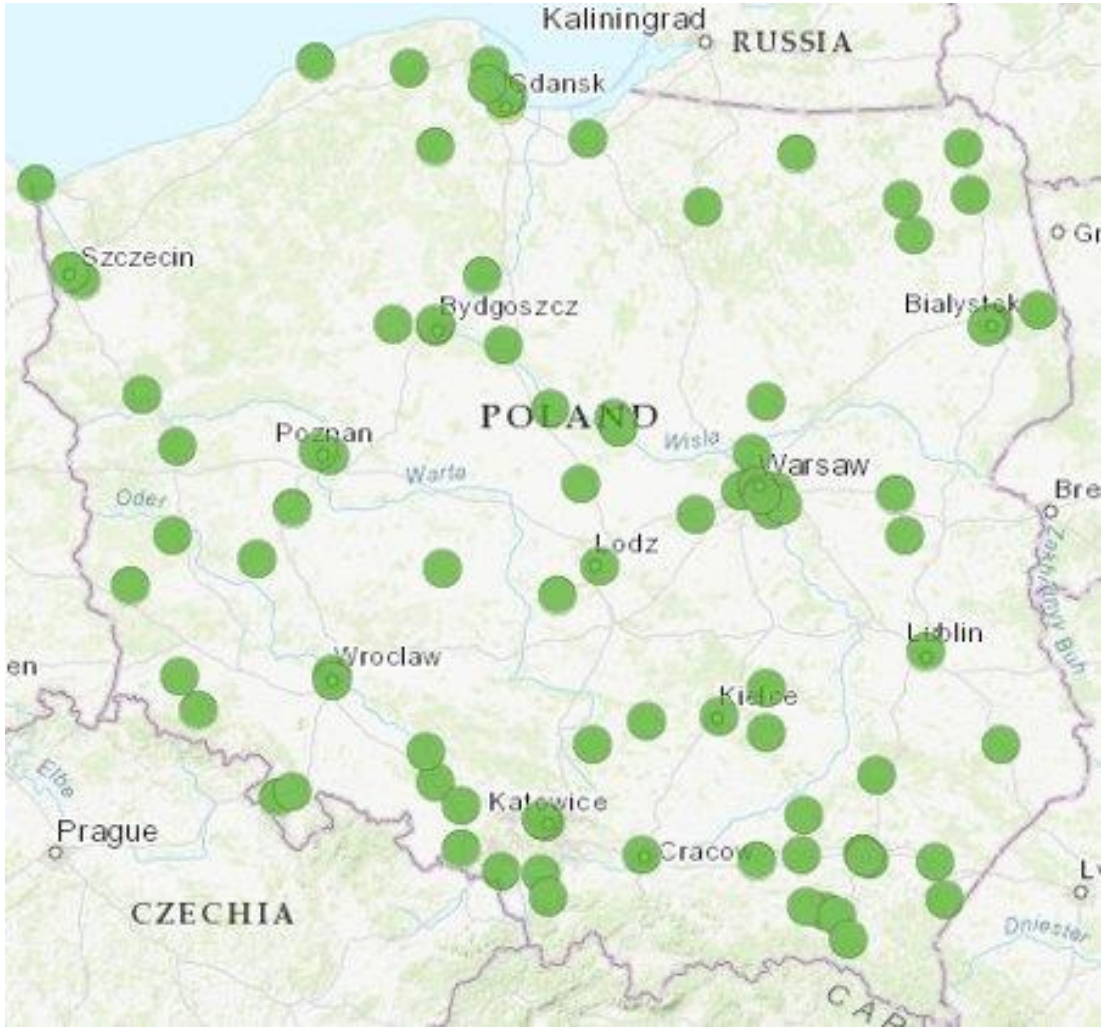
Feinstaub	EU	WHO	Mitteilungszeitraum
PM_{10}	40 µm³ 50 µm³ 35 Tage/Jahr	15 µm³ 45 µm³ 3-4 Tage/Jahr	1 Jahr 24 Stunden Erlaubte Überschreitung
$PM_{2.5}$	25 µm³ - -	5 µm³ 15 µm³ 3-4 Tage/Jahr	1 Jahr 24 Stunden Erlaubte Überschreitung

Untersuchungsgebiet Polen PM_{10}



PM_{10} Grenzwerte in µg/m³	$PM_{2.5}$ Grenzwerte in µg/m³	Kategorie
0 – 20	0 – 13	Very good
20.1 – 50	13.1 – 35	Good
50.1 – 80	35.1 – 55	Moderate
80.1 – 110	55.1 – 75	Sufficient
110.1 – 150	75.1 – 110	Bad
> 150	> 110	Very bad

Untersuchungsgebiet Polen $PM_{2.5}$



PM_{10} Grenzwerte in µg/m³	$PM_{2.5}$ Grenzwerte in µg/m³	Kategorie
0 – 20	0 – 13	Very good
20.1 – 50	13.1 – 35	Good
50.1 – 80	35.1 – 55	Moderate
80.1 – 110	55.1 – 75	Sufficient
110.1 – 150	75.1 – 110	Bad
> 150	> 110	Very bad

Forschungsfragen

- Lässt sich mit Hilfe eines neuronalen Netzes unter Verwendung einer CNN-LSTM Kombination eine stündliche Prognose von Feinstaubdaten für die nächsten 14 Tage realisieren?
  - Ist es möglich den PM_{10} Wert mit einem MAE unter 10 vorherzusagen?
  - Ist es möglich den $PM_{2.5}$ Wert mit einem MAE unter 10 vorherzusagen?
  - Gibt es einen Zusammenhang zwischen PM_{10} und $PM_{2.5}$, sodass $PM_{2.5}$ mit dem Modell für PM_{10} vorhergesagt werden kann?
  - Wie sehen unsere Prognosen im Vergleich mit denen des polnischen Umweltamts aus? (für einen Tag)
  - Ist es sinnvoll, Stationen zu Gebieten zusammenzufassen, sodass die Aussagekräftigkeit der Prädiktion im Vergleich zu den einzelnen Stationen gleich bleibt oder verbessert wird?

Feinstaub

- Bereitstellung durch REST-API
- Crawling
  - Beschaffung der:
    - Station-IDs
      - Mit dazugehörigen Sensoren
  - Stationen durchlaufen
    - Jeden Sensor abfragen
    - Ergebnisse sind Paginiert
- Restriktionen durch Timeout
  - Zwei Anfragen pro Minute


```
[{"id":114,"stationName":"Wrocław, ul. Bartnicza","gegrLat":"51.115933","gegrLon":"17.141125","city":
{"id":1064,"name":"Wrocław","commune":
{"communeName":"Wrocław","districtName":"Wrocław","provinceName":"DOLNOŚLĄSKIE"}},"addressStreet":"ul.
Bartnicza"},{"id":117,"stationName":"Wrocław, wyb. Conrada-
Korzeniowskiego","gegrLat":"51.129378","gegrLon":"17.029250","city":{"id":1064,"name":"Wrocław","commune":
{"communeName":"Wrocław","districtName":"Wrocław","provinceName":"DOLNOŚLĄSKIE"}},"addressStreet":"ul. Wyb.
J.Conrada-Korzeniowskiego 18"},{"id":129,"stationName":"Wrocław, al.
Wiśniowa","gegrLat":"51.086225","gegrLon":"17.012689","city":{"id":1064,"name":"Wrocław","commune":
{"communeName":"Wrocław","districtName":"Wrocław","provinceName":"DOLNOŚLĄSKIE"}},"addressStreet":"al.
Wiśniowa/ul. Powst. Śląskich"},{"id":52,"stationName":"Legnica, al.
Rzeczypospolitej","gegrLat":"51.204503","gegrLon":"16.180513","city":{"id":453,"name":"Legnica","commune":
{"communeName":"Legnica","districtName":"Legnica","provinceName":"DOLNOŚLĄSKIE"}},"addressStreet":"al.
Rzeczypospolitej 10/12"},{"id":109,"stationName":"Wałbrzych, ul.
Wysockiego","gegrLat":"50.768729","gegrLon":"16.269677","city":{"id":998,"name":"Wałbrzych","commune":
{"communeName":"Wałbrzych","districtName":"Wałbrzych","provinceName":"DOLNOŚLĄSKIE"}},"addressStreet":"ul.
Wysockiego 11"},{"id":11,"stationName":"Czerniawa","gegrLat":"50.912475","gegrLon":"15.312190","city":
{"id":142,"name":"Czerniawa","commune":{"communeName":"Świeradów-
Zdrój","districtName":"lubański","provinceName":"DOLNOŚLĄSKIE"}},"addressStreet":"ul. Strażacka 7"},
{"id":38,"stationName":"Kłodzko, ul. Szkolna","gegrLat":"50.433493","gegrLon":"16.653660","city":
{"id":368,"name":"Kłodzko","commune":
{"communeName":"Kłodzko","districtName":"kłodzki","provinceName":"DOLNOŚLĄSKIE"}},"addressStreet":"ul. Szkolna
8"},{"id":70,"stationName":"Oława, ul. Żołnierzy Armii
Krajowej","gegrLat":"50.942073","gegrLon":"17.291333","city":{"id":642,"name":"Oława","commune":
{"communeName":"Oława","districtName":"oławski","provinceName":"DOLNOŚLĄSKIE"}},"addressStreet":"ul. Żołnierzy AK
9"},{"id":74,"stationName":"Osieczów","gegrLat":"51.317630","gegrLon":"15.431719","city":
{"id":648,"name":"Osieczów","commune":
{"communeName":"Osiecznica","districtName":"bolesławiecki","provinceName":"DOLNOŚLĄSKIE"}},"addressStreet":"bez
ulicy"},{"id":84,"stationName":"Śnieżka","gegrLat":"50.736389","gegrLon":"15.739722","city":
{"id":346,"name":"Karpacz","commune":
{"communeName":"Karpacz","districtName":"karkonoski","provinceName":"DOLNOŚLĄSKIE"}},"addressStreet":"Śnieżka"},
```

Feinstaub

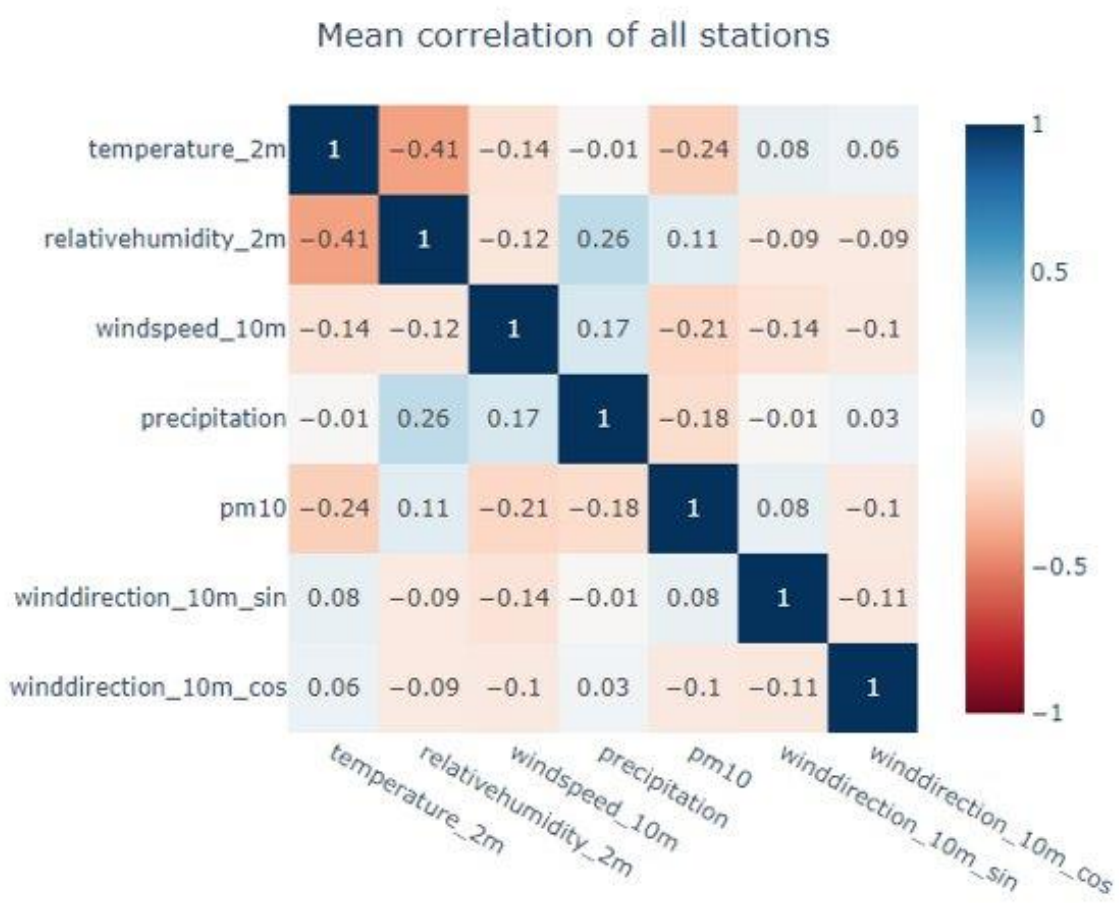
- Stündliche Feinstaubdaten
- Für Zeitraum 2018 bis 2021 aus Excel Daten extrahiert
  - 2018: 98 Stationen
  - 2019: 108 Stationen
  - 2020: 125 Stationen
  - 2021: 153 Stationen
  - 2022: 160 Stationen
- 99 Stationen mit validen Werten für mindestens Hälfte der Jahre
- Einige Stationen im Zeitraum neu aufgebaut
- Einige Stationen im Zeitraum abgeschaltet

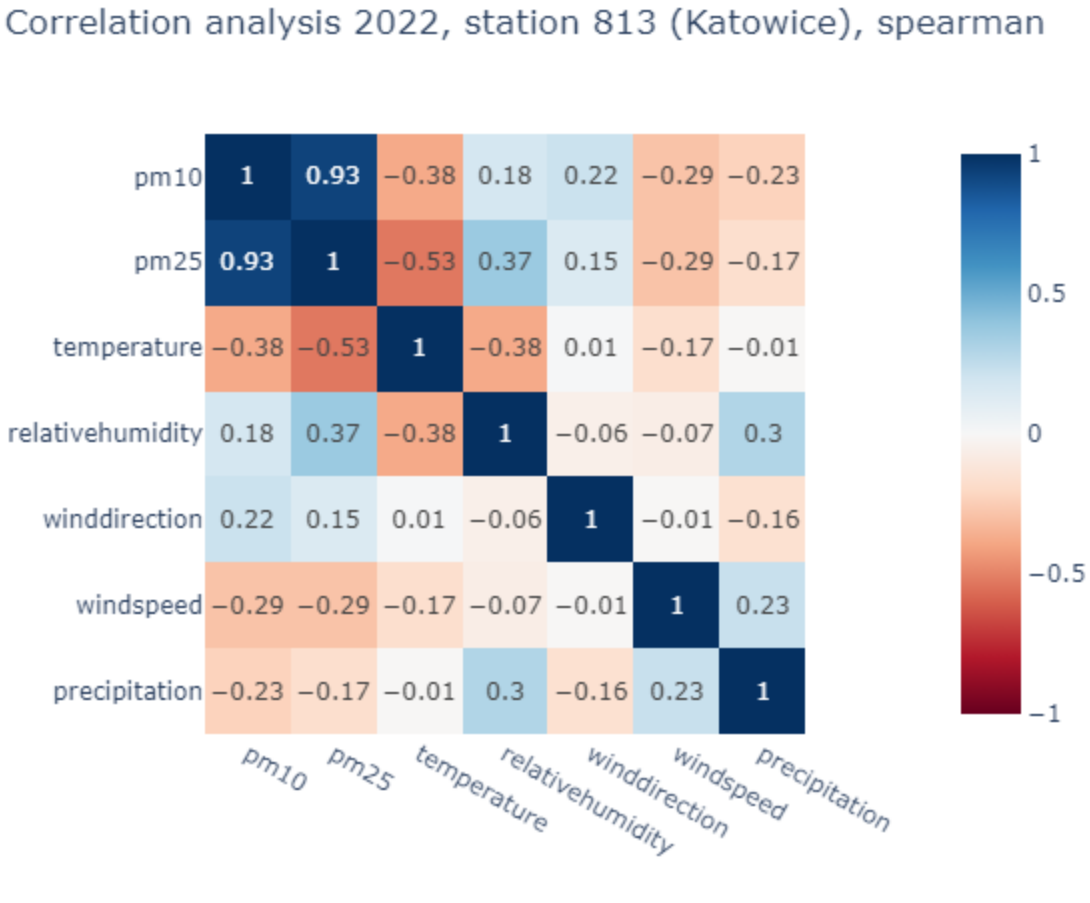
Wetterdaten

- Daten von Open-Meteo - Modellbasierte Interpolation für alle Koordinaten
- Modell verwendet Daten von Wetterstation, Flugzeugen und Satelliten
- Dadurch genauere Werte, wenn Wetterstation weit von Feinstaubstation entfernt
- Model liefert vergleichbare Werte zu Messstationen

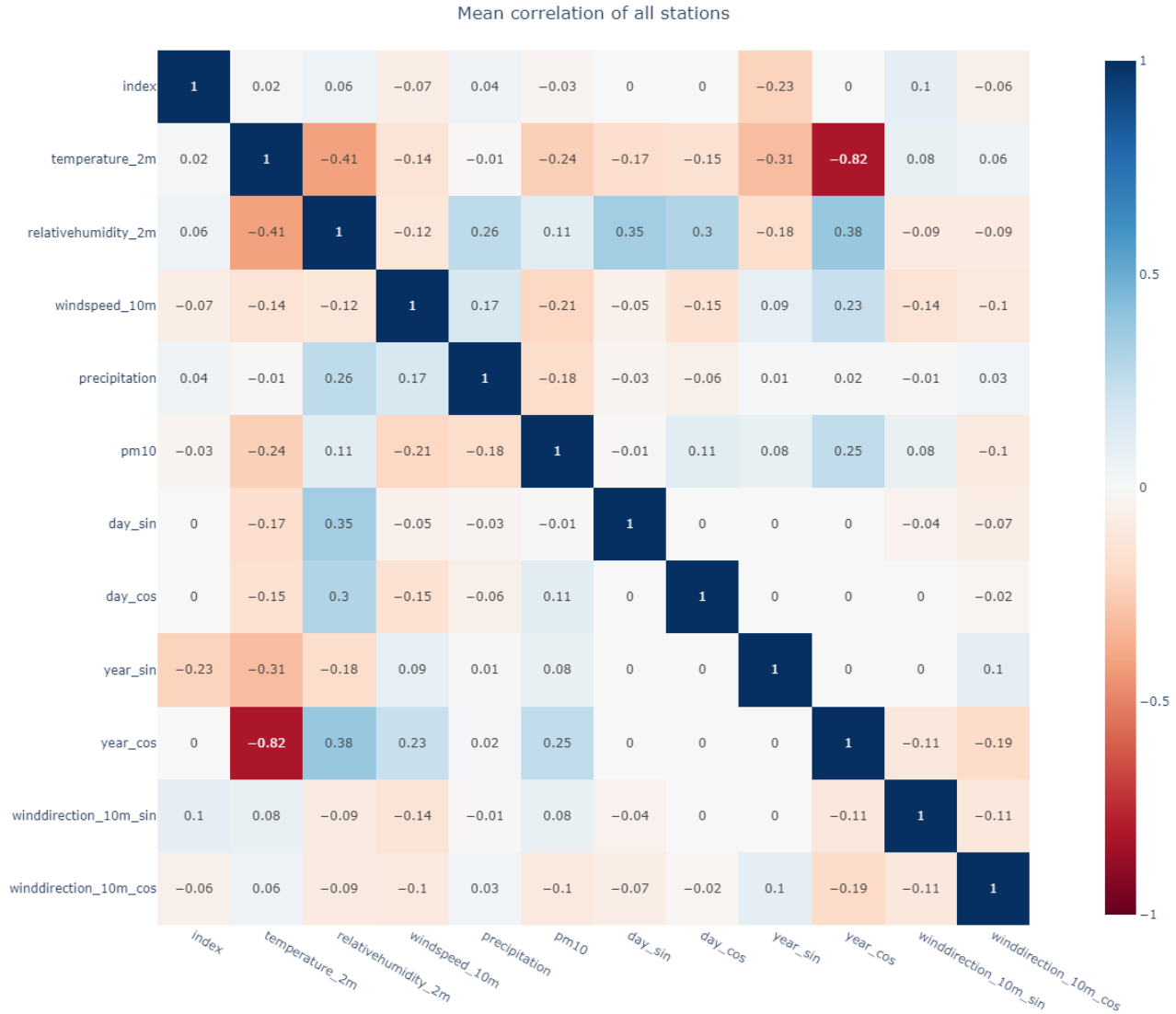
```
latitude = 50.732817; longitude = 16.648050
start_date = datetime(2022, 1, 1); end_date = datetime(2022, 1, 31)
options = ModelBasedOptions(
    hourly=[
        HourlyData.Temperature_2m, HourlyData.RelativeHumidity_2m, HourlyData.WindDirection_10m,
        HourlyData.WindSpeed_10m, HourlyData.Precipitation_rain_showers_snow,
    ]
)
meta_data_model, daily_model, hourly_model = WeatherData.getModelBasedData(
    latitude, longitude, start_date, end_date, options
)
```

Korrelationplots





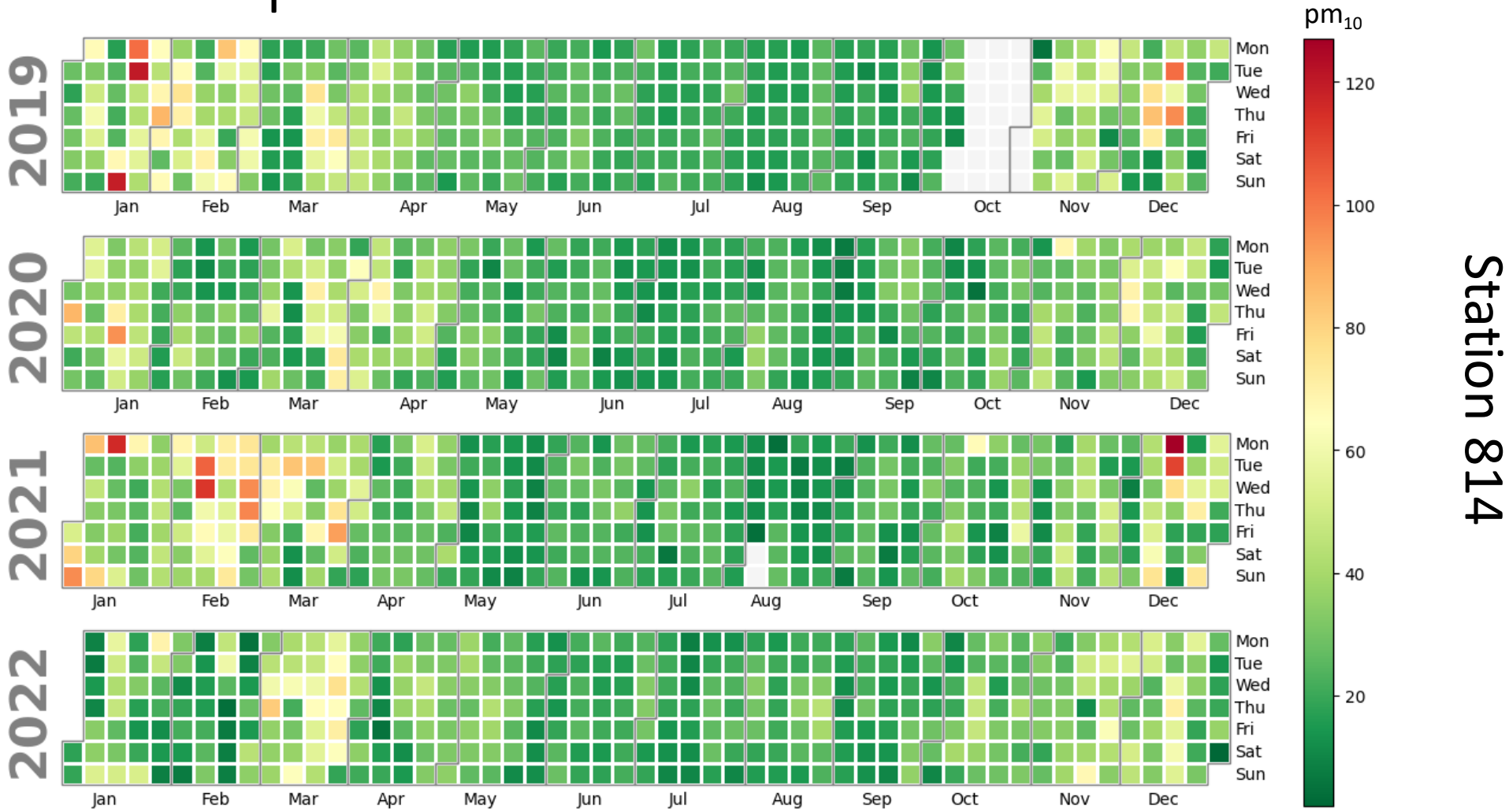
Korrelationplot alle Stationen



- Korrelation für alle Stationen
- Sinus and Cosinus Features
  - Tag
  - Jahr
  - Windrichtung

➢ Keine sinnvollen Korrelationen in den Daten vorhanden

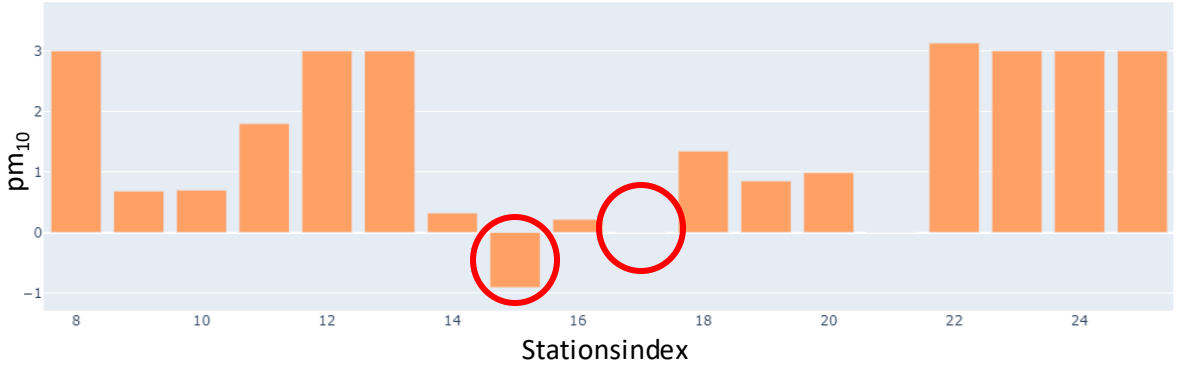
Heatmaps


2019
2020
2021
2022
Mon
Tue
Wed
Thu
Fri
Sat
Sun
Jan
Feb
Mar
Apr
May
Jun
Jul
Aug
Sep
Oct
Nov
Dec
pm10
120
100
80
60
40
20
Station 814

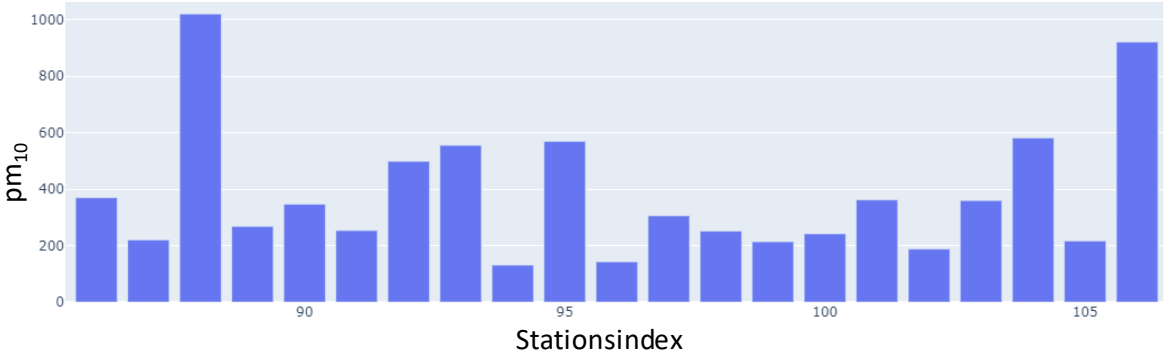
Interpolation PM_{10}

- Maximalwerte aller Stationen valide
  - Outlier Detection nicht notwendig
- Inkorrekte Werte
  - Fehlende Werte → NaN
  - Negative PM_{10} Werte
- Interpolation inkorrekter Werte für bis zu 5 aufeinander folgende Stunden
- Entfernen der restlichen Zeiträume

Minimalwerte der Datasets einiger Stationen:

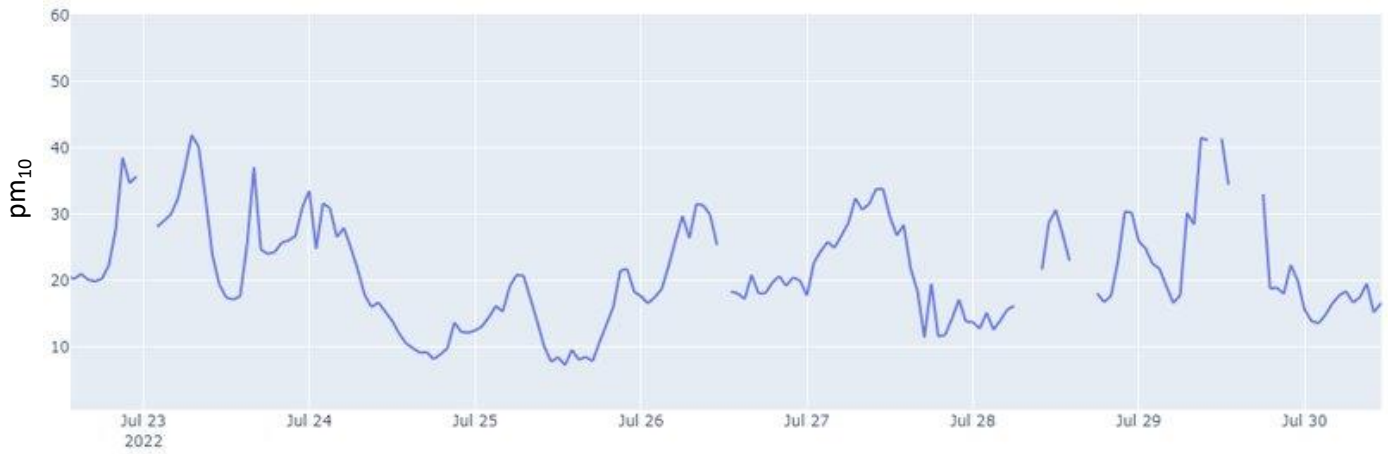


Maximalwerte der Datasets einiger Stationen:

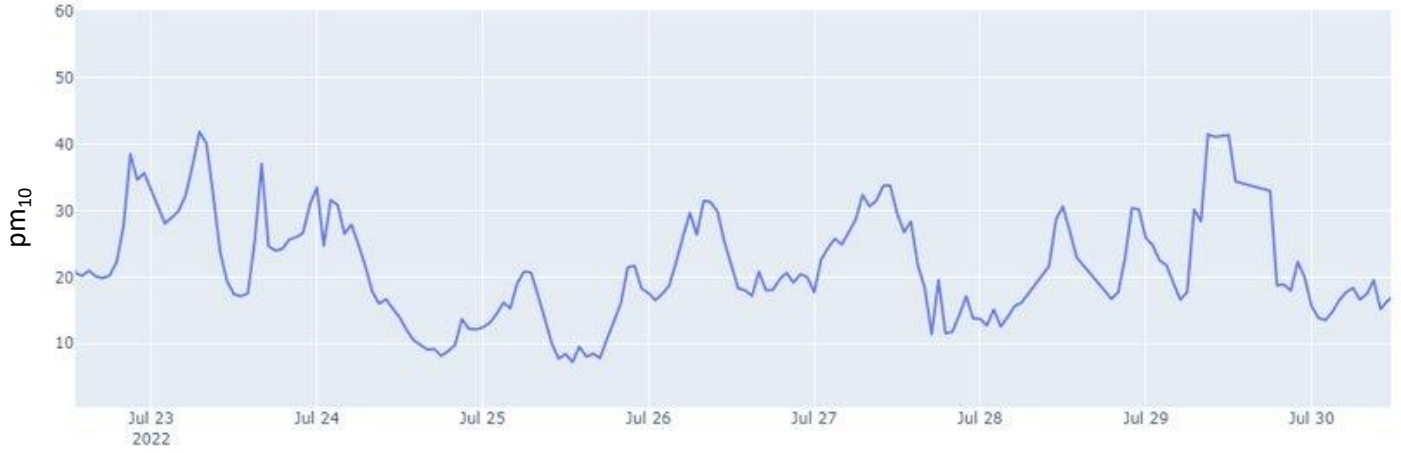


Interpolation PM_{10}

Original Daten für PM_{10}

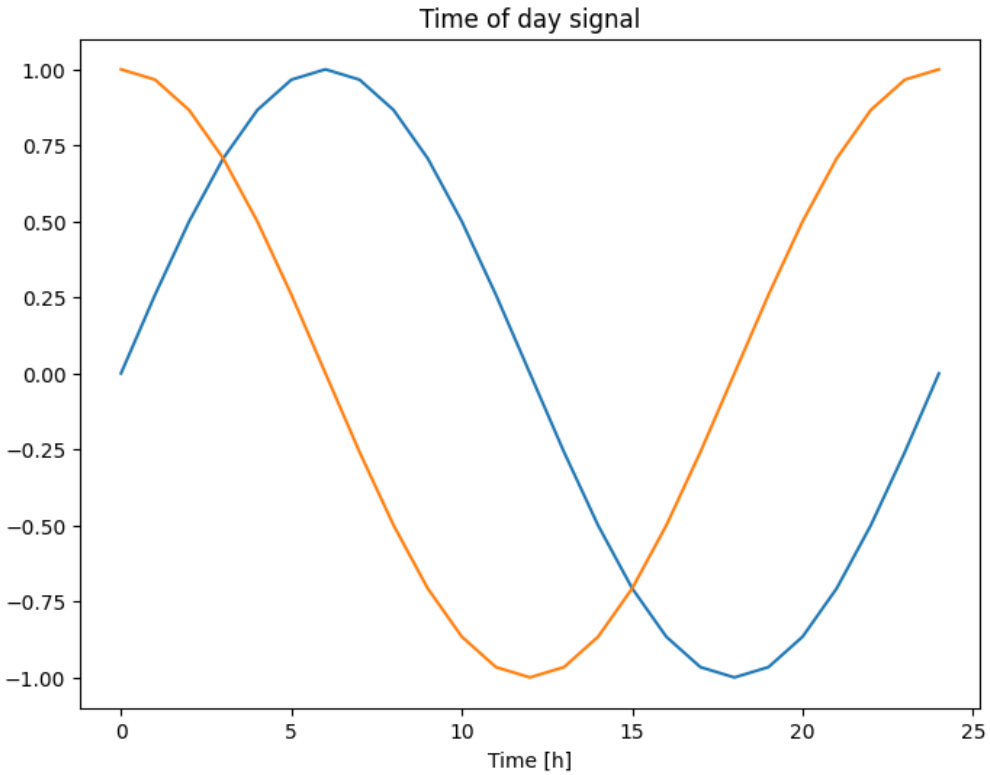


Interpolierte Daten für PM_{10}



Zeit

- Zeit periodisch angeben statt absolut
- Perioden für Tag und Jahr
- Für Modell besser verwendbar
- Periode durch Sin und Cos abbilden

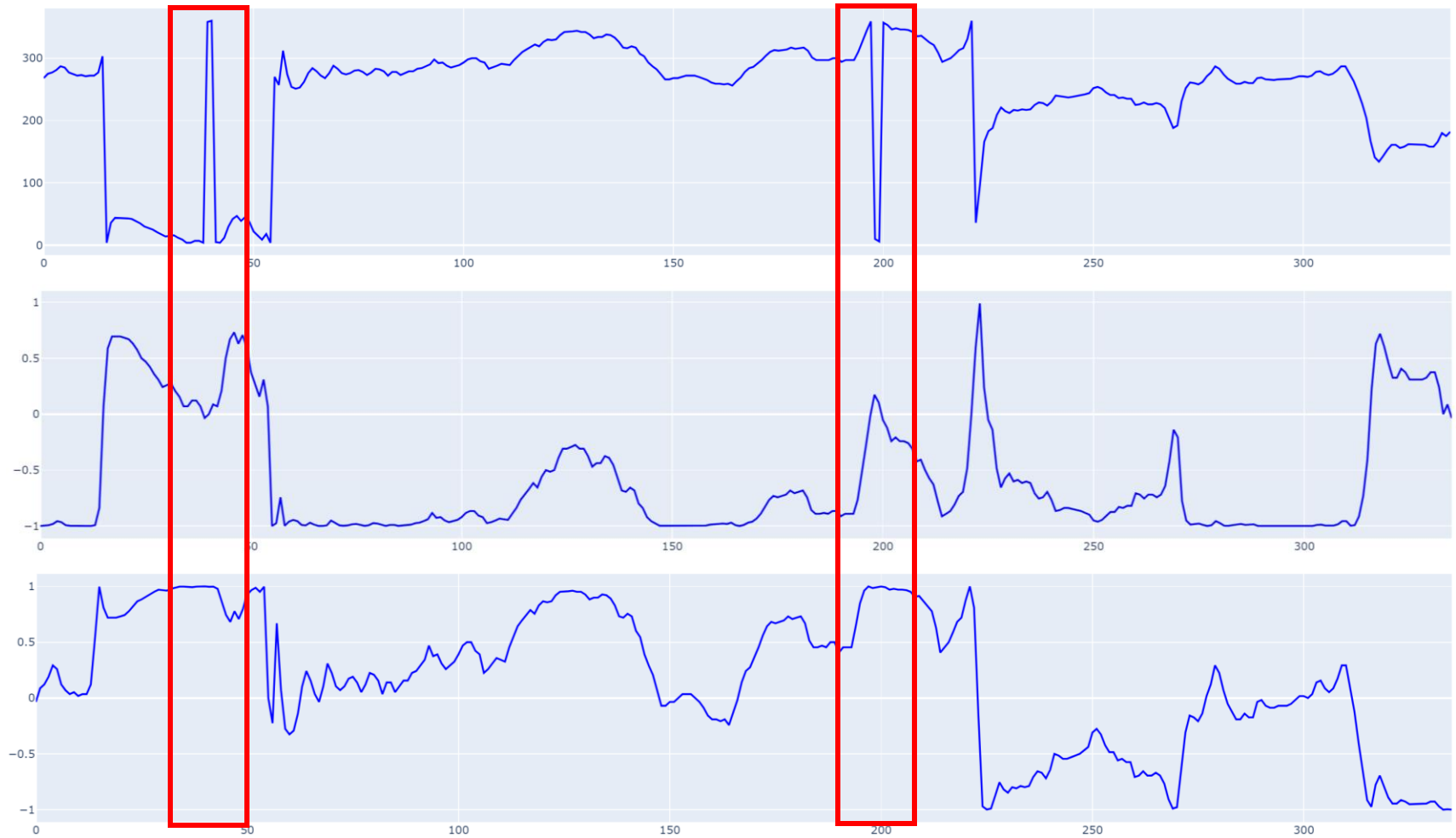


Windrichtung

Windrichtung in Grad

Windrichtung in Radiant (Sinus)

Windrichtung in Radiant (Cos)



14 Tage Fenster

Feature Vector

- Temperatur (°C)
- Luftfeuchtigkeit (%)
- Windgeschwindigkeit (m/s)
- Niederschlag (l/m²)
- Windrichtung (sin & cos)
- Tag (sin & cos)
- Jahr (sin & cos)
- PM_{10}

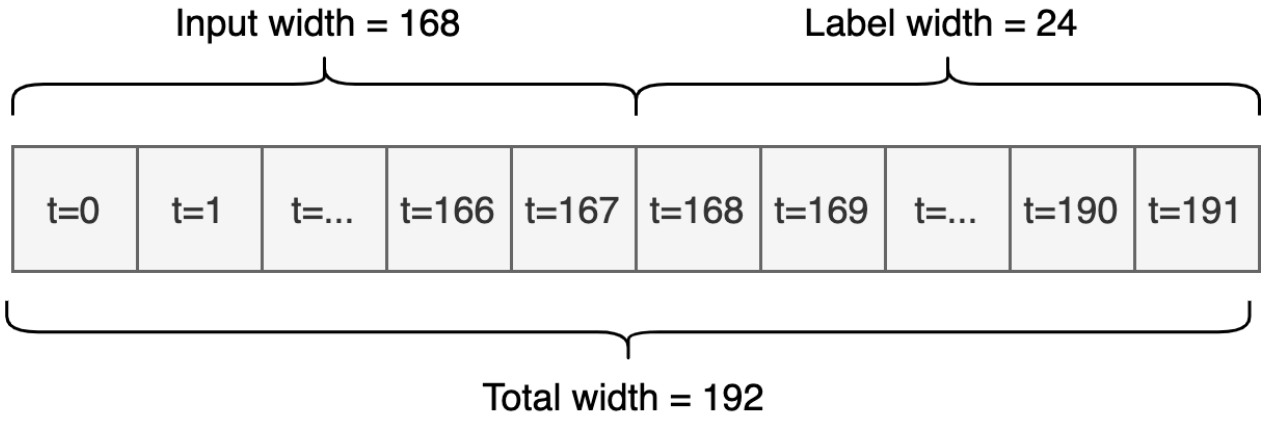
➢ 11 Features

Normalisierung

- Normalisierung der numerischen Werte mittels Standard Skalierung
  - (value - mean) / sqrt(var)
- Skalierung durch Normalization-Layer im neuronalen Netz
- Mittelwert und Varianz der Daten werden während des Trainings gelernt

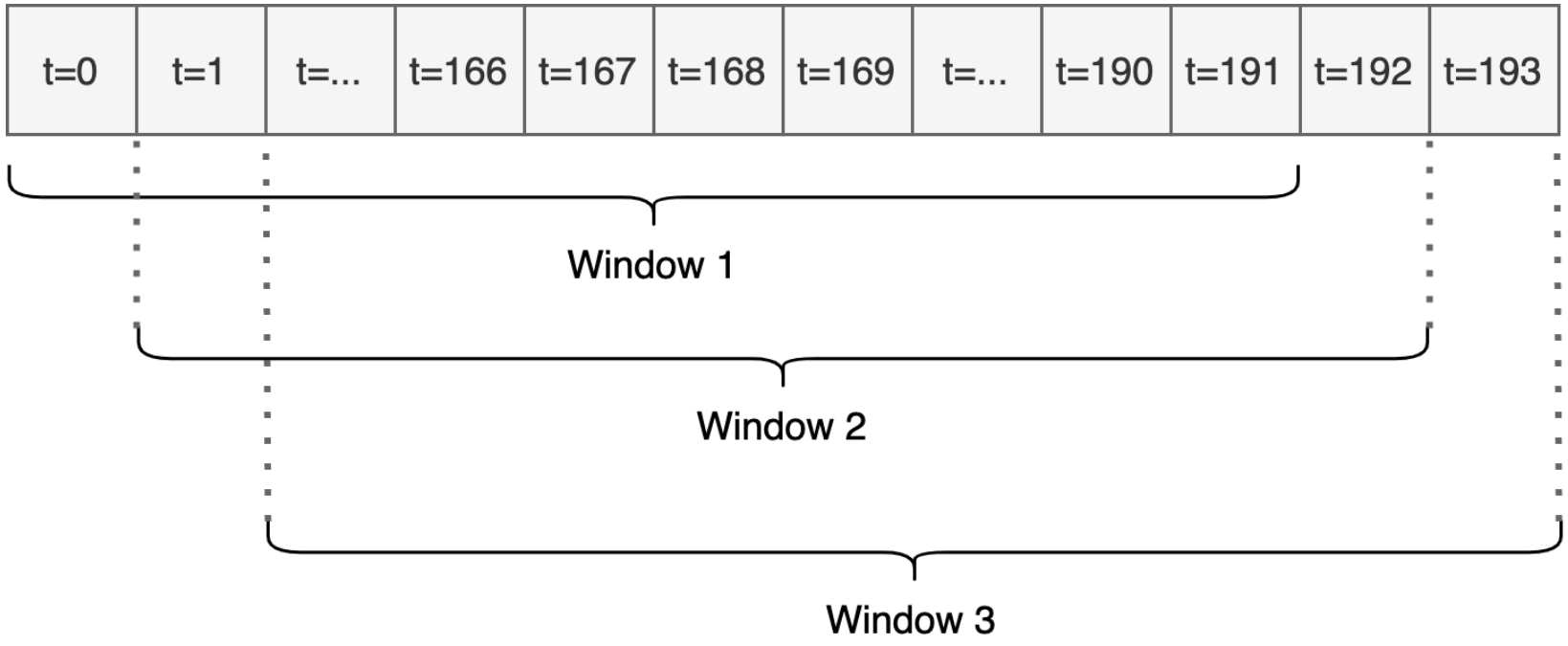
Window Generator

- Erstellen von 8-tägigen Trainingswindows aus den Daten
  - Generierung von N Windows mit vollständigen Feature Vektoren
    - ➢ Keine Zeitsprünge vorhanden
- Länge des Windows variabel einstellbar
  - Definition von Input- und Labelbreite der Daten



Window Generator

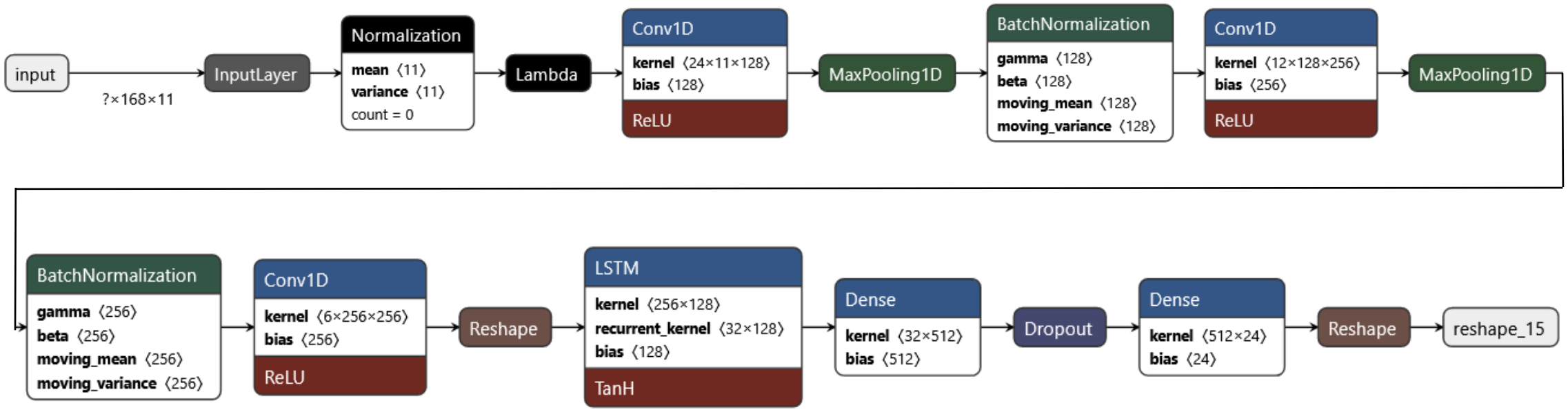
- Erstellung des Windows in stündlichen Schritten
  - Überprüfung auf zeitliche Lücken



Data Windowing

Window Generator

Architektur des Neuronalen Netzes

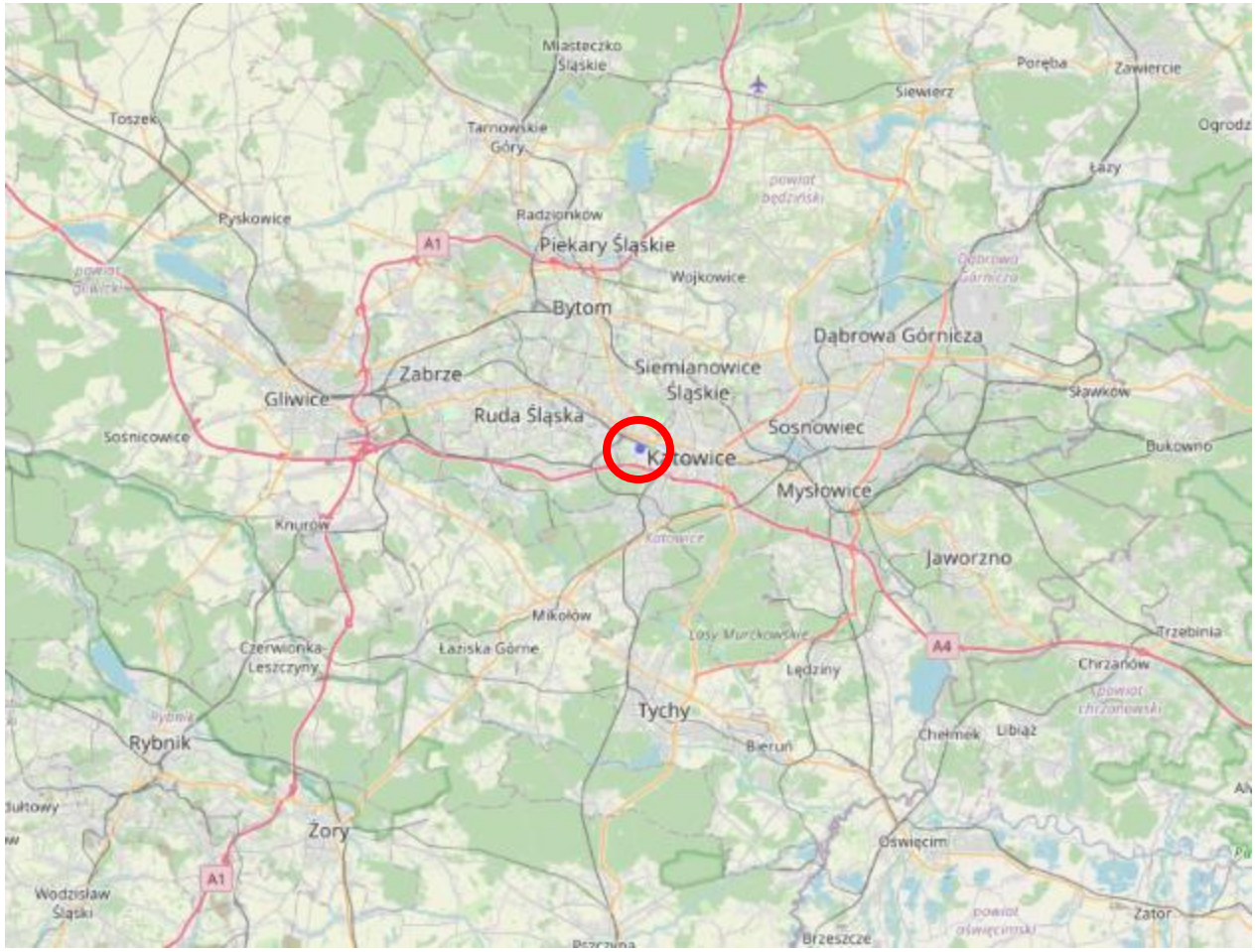


Architektur des Neuronalen Netzes

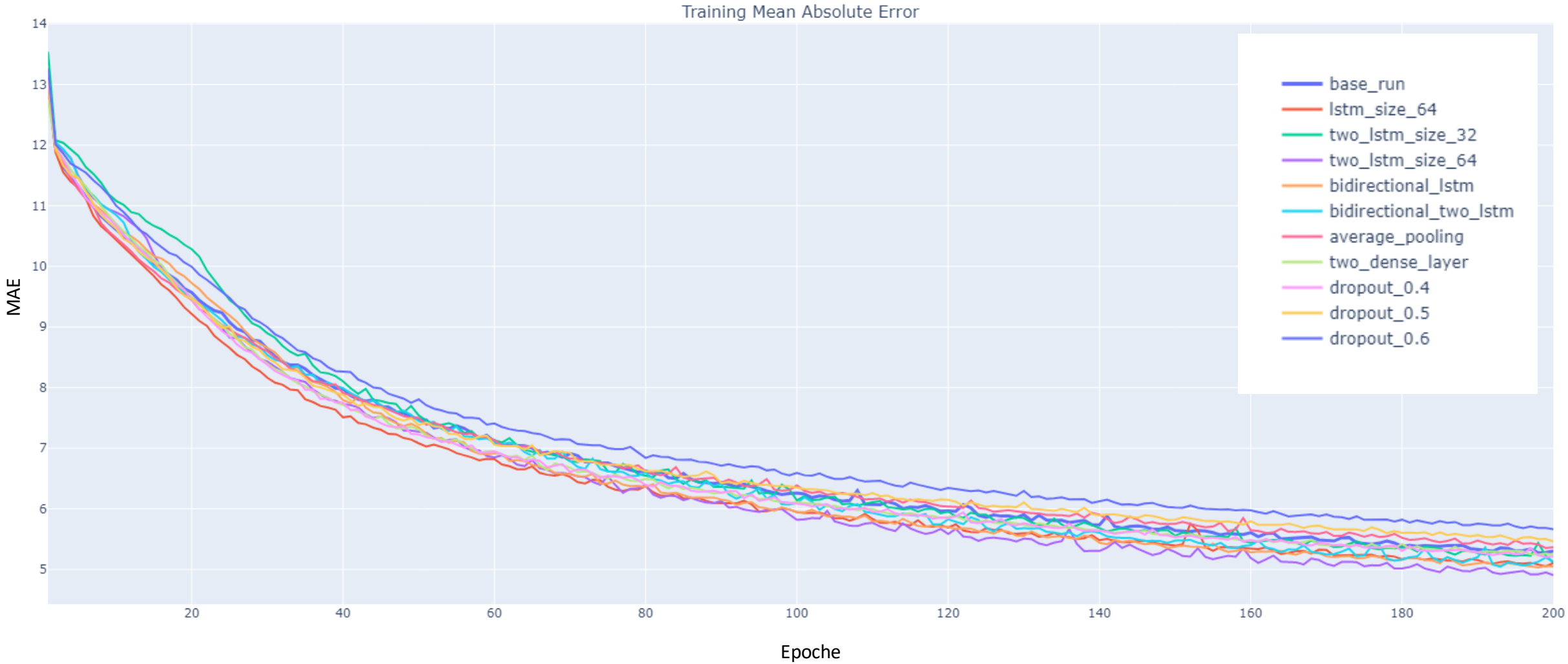
```
# Convolution
cnn_lstm_model = tf.keras.models.Sequential()
cnn_lstm_model.add(tf.keras.layers.Normalization())
cnn_lstm_model.add(tf.keras.layers.Lambda(lambda x: x[:, -6:, :]))
cnn_lstm_model.add(tf.keras.layers.Conv1D(128, activation="relu", kernel_size=(24), padding="same"))
cnn_lstm_model.add(tf.keras.layers.MaxPooling1D())
cnn_lstm_model.add(tf.keras.layers.BatchNormalization())
cnn_lstm_model.add(tf.keras.layers.Conv1D(256, activation="relu", kernel_size=(12), padding="same"))
cnn_lstm_model.add(tf.keras.layers.MaxPooling1D())
cnn_lstm_model.add(tf.keras.layers.BatchNormalization())
cnn_lstm_model.add(tf.keras.layers.Conv1D(256, activation="relu", kernel_size=(6), padding="same"))
cnn_lstm_model.add(tf.keras.layers.Reshape((-1, 256)))
# LSTM
cnn_lstm_model.add(tf.keras.layers.LSTM(32, return_sequences=True))
# Dense
cnn_lstm_model.add(tf.keras.layers.Dense(512))
cnn_lstm_model.add(tf.keras.layers.Dropout(0.2))
cnn_lstm_model.add(tf.keras.layers.Dense(24, kernel_initializer=tf.initializers.zeros()))
# Output
cnn_lstm_model.add(tf.keras.layers.Reshape([24, 1]))
```

Training

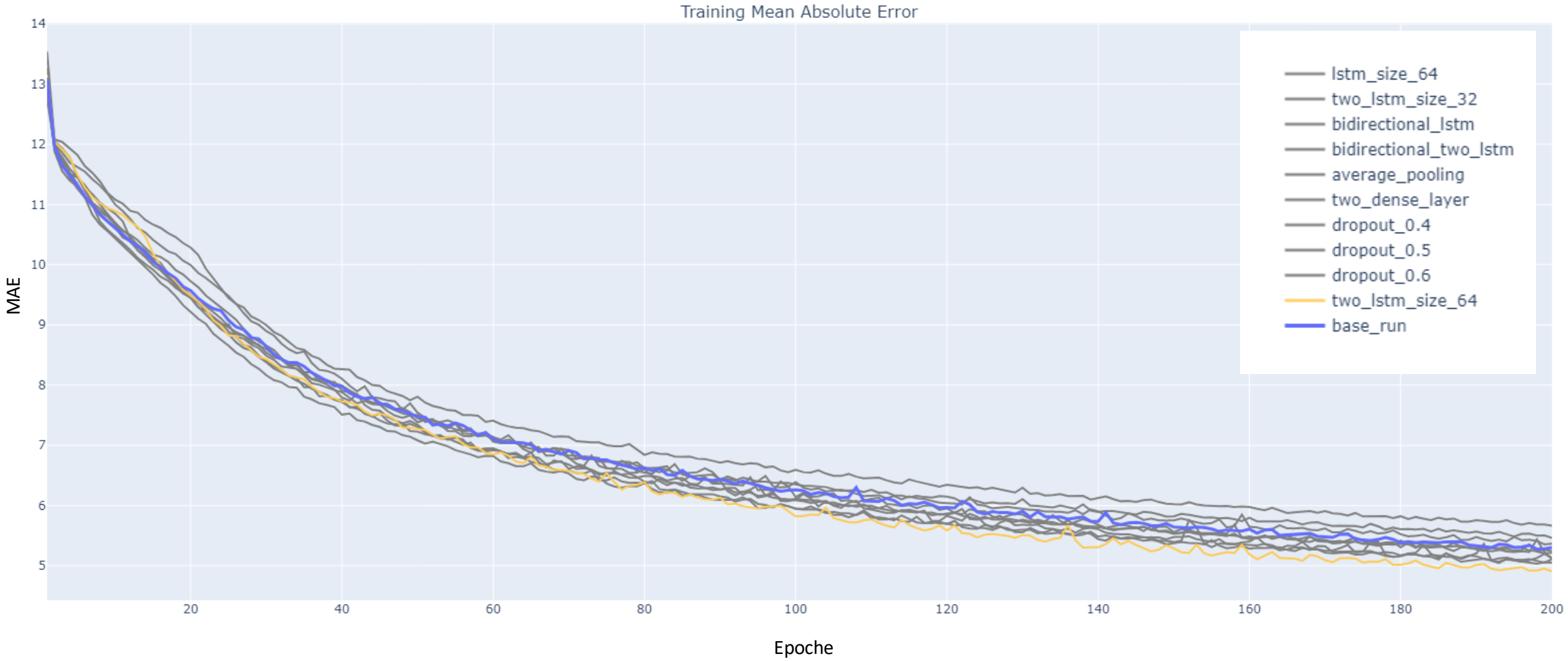
- Training auf Daten von Station 814
- Trainingsdaten von 2019 – 2022
- Windows:
  - Trainingswindows: 70 %
  - Validierungswindows: 10 %
  - Testwindows: 20 %
- Testwindows am Ende der Trainingsdaten Zeitspanne
- Train und Validation Windows werden geshuffelt



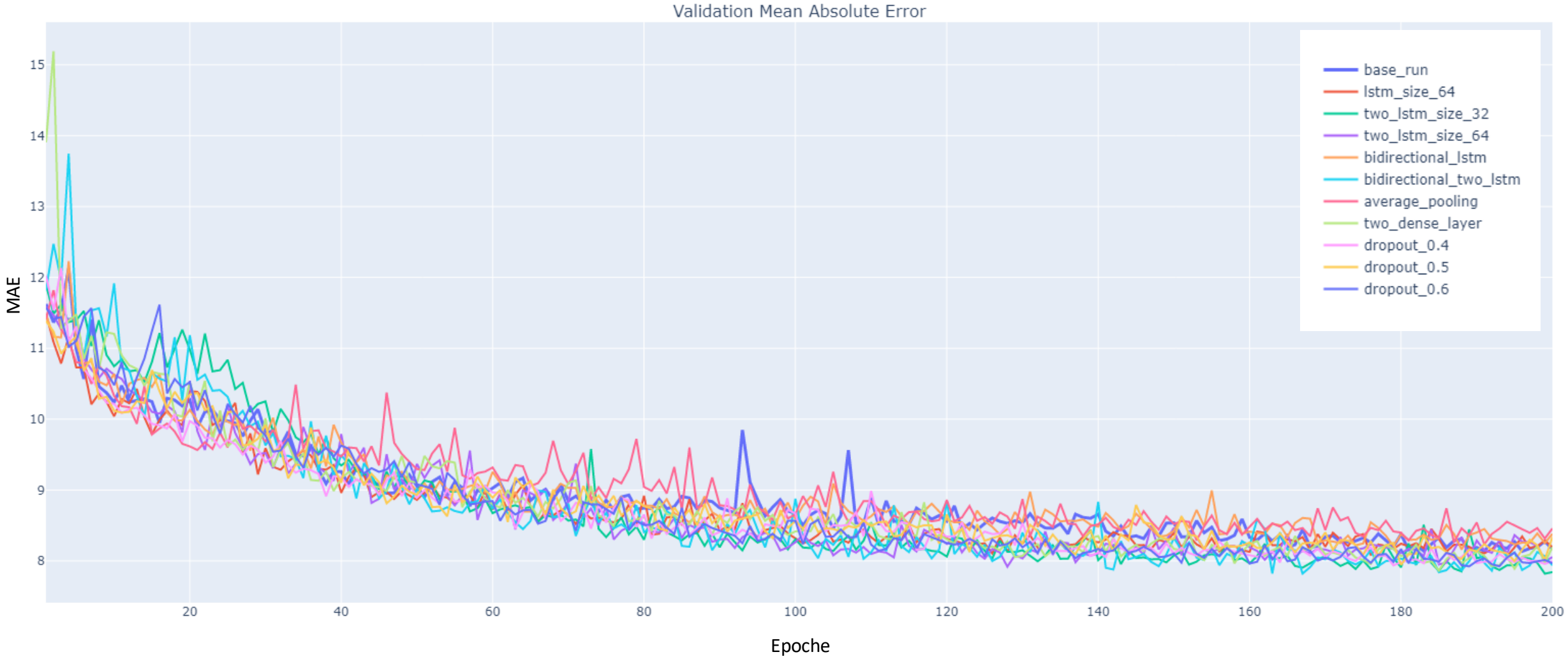
Versuchsergebnisse – Training MAE aller Modelle



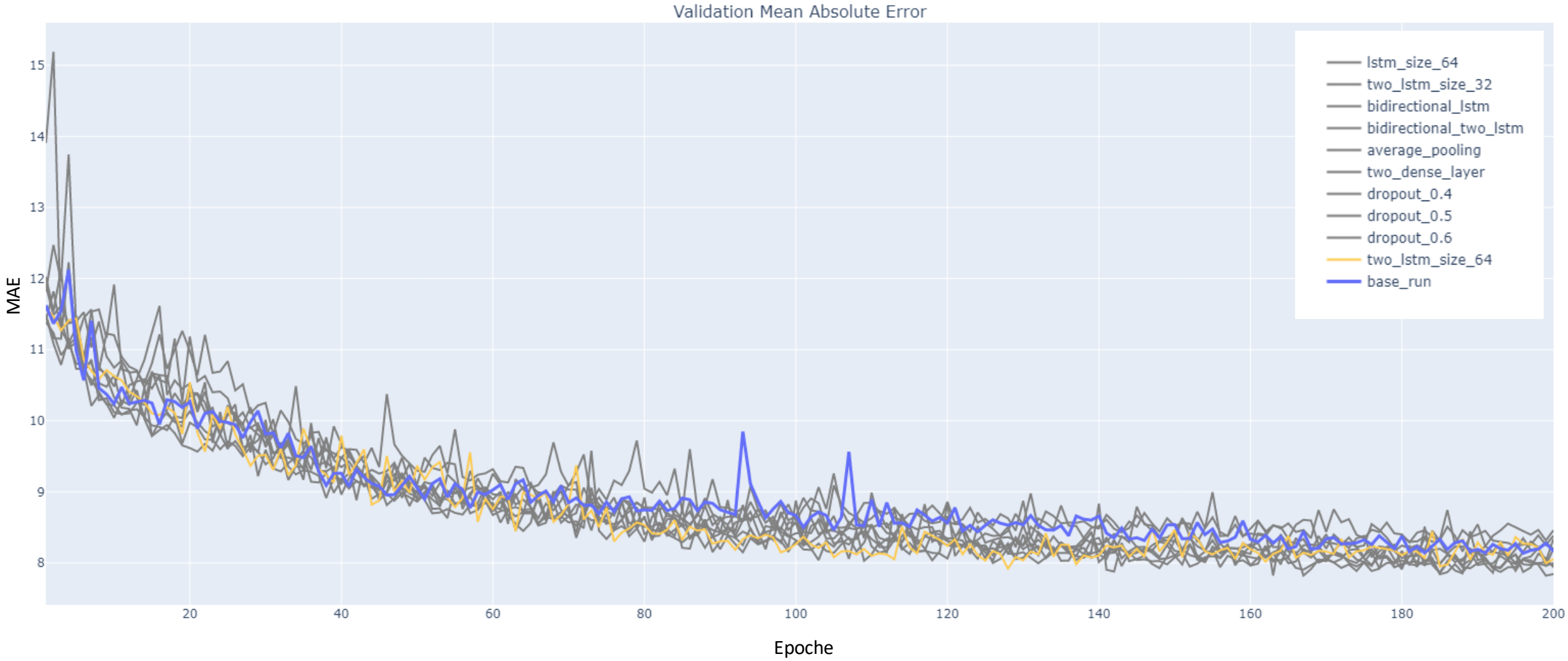
Versuchsergebnisse – Training MAE aller Modelle



Versuchsergebnisse – Validation MAE aller Modelle

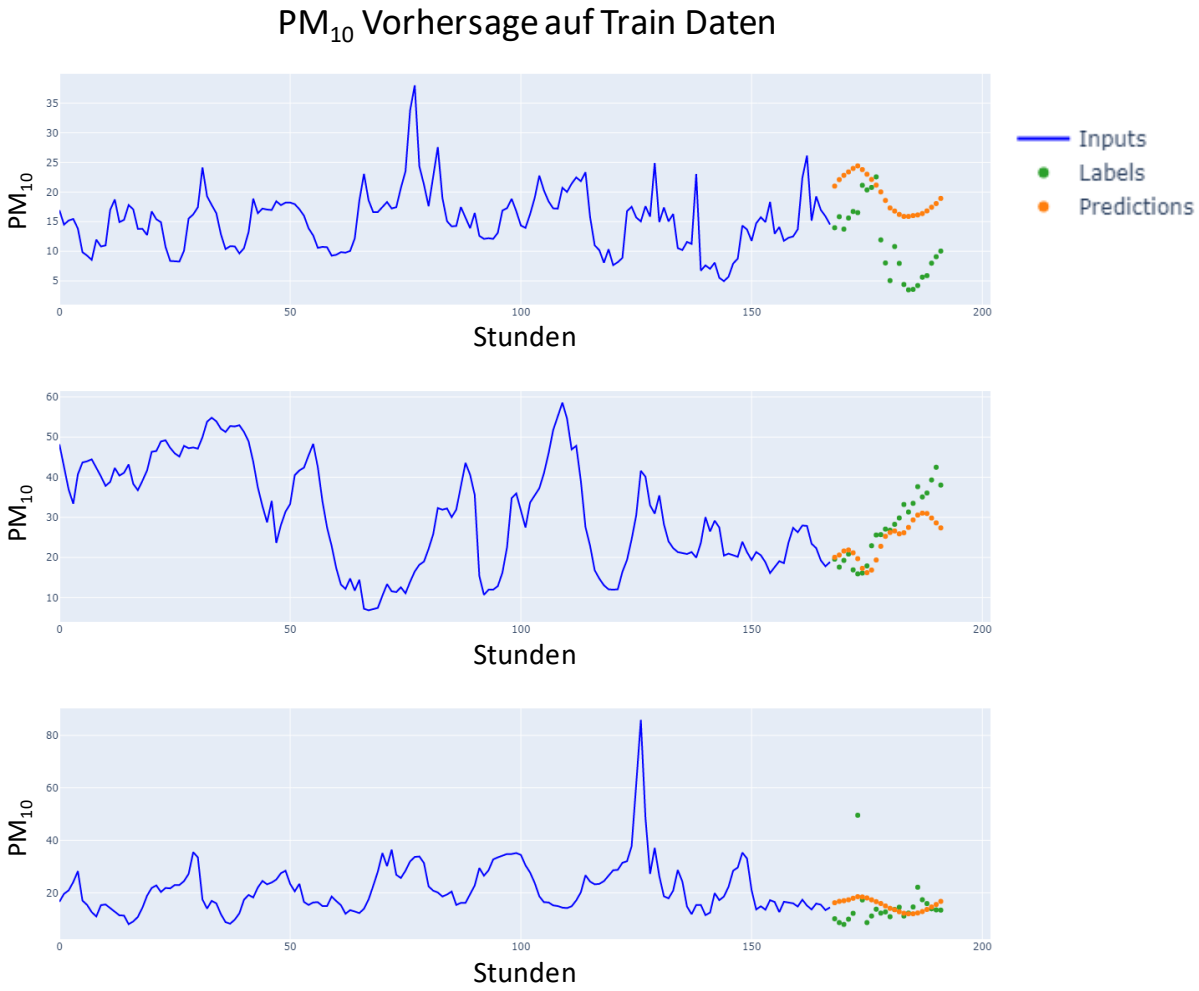


Versuchsergebnisse – Validation MAE aller Modelle



Bestes Modell

- Parametrisierung von IMT-14:
  - zwei LSTM-Layer mit Größe 64
  - 200 Epochen
  - Dropout 0.2
  - 7 Tage Eingabe
  - 24 Stunden Vorhersage
- Auswertung
  - Train MAE = 4,91
  - Val MAE = 8,05
  - Test MAE = 10,21



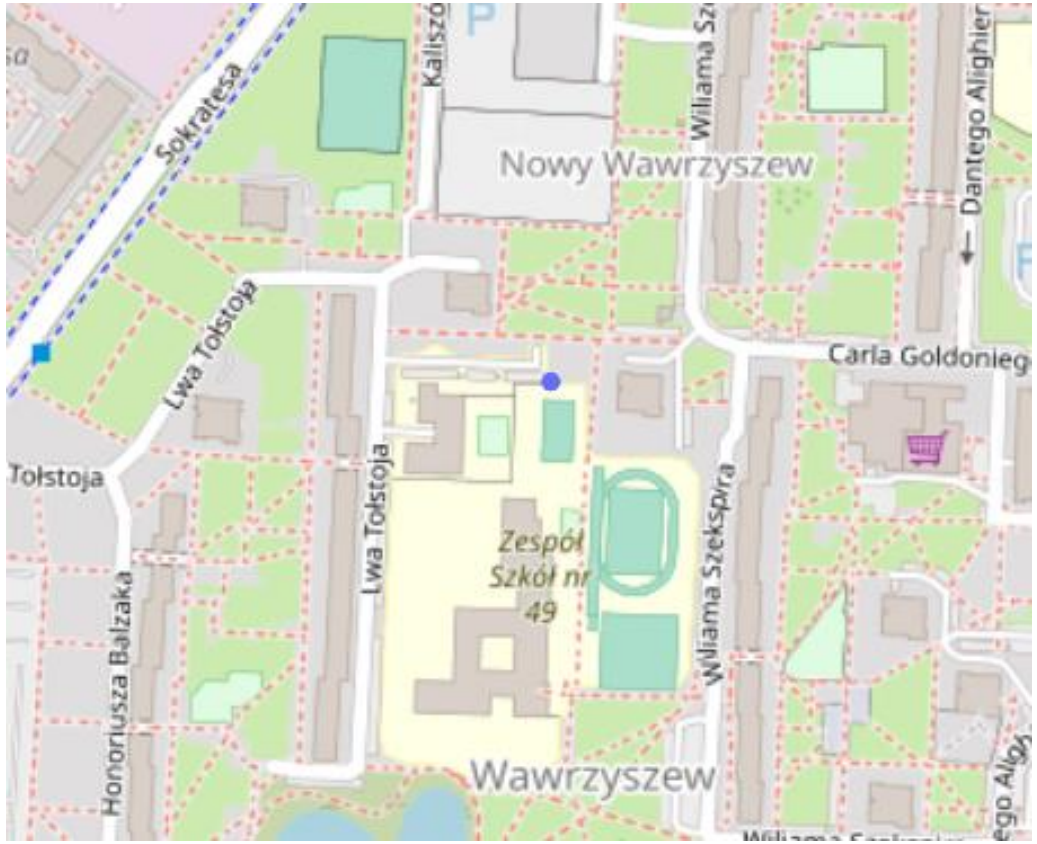
Ergebnisse – Station 530 & 538

Voraussage durch Modell einer naheliegenden Station

- Drei Modelle trainiert:
  - Für Station 538, 530 und Kombination beider
  - Testen auf Daten beider Stationen
- Ziel:
  - Vorhersage der PM10 Daten
  - Auf naheliegender Station trainierten Modell
- Ergebnis:
  - Performance schlecht bis medium
- Erkenntnisse:
  - Verschiedene Standort Begebenheiten

Lage der Station 530 und 538

Station 538



Stationstyp: Background

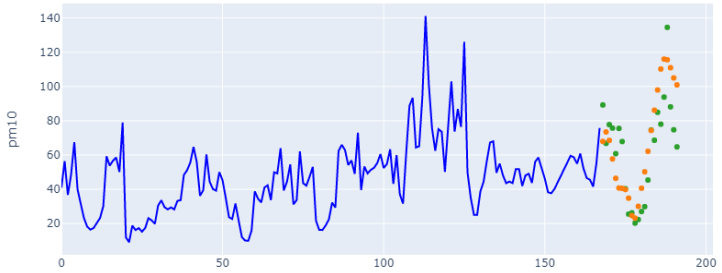
Station 530

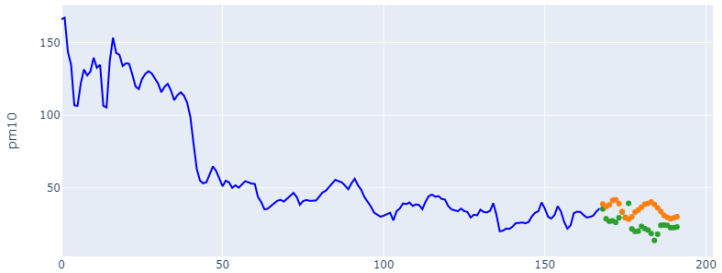
Stationstyp: Traffic

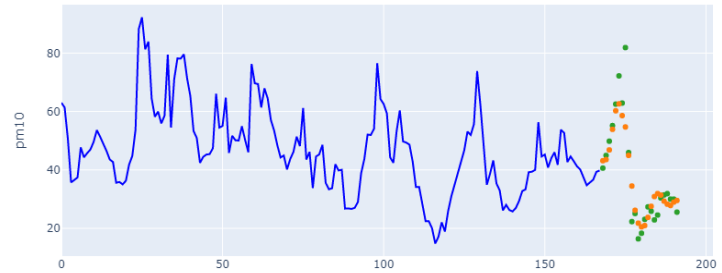
Ergebnisse – Station 530 & 538

Messdaten der Station 530

Modell trainiert auf 530

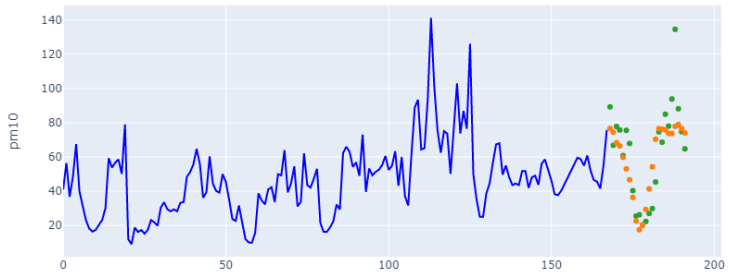


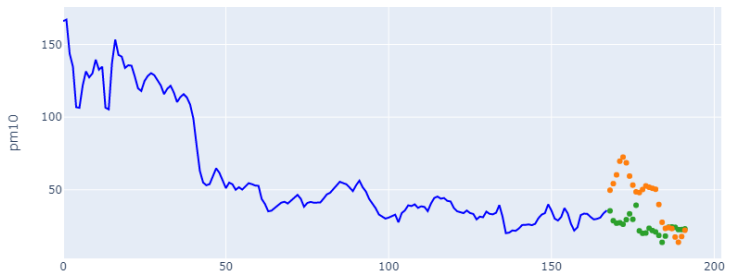


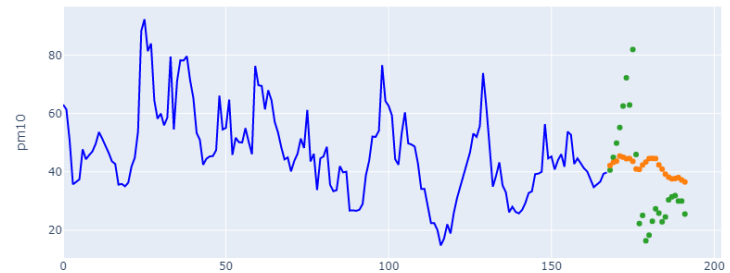


Vorhersage auf Train Daten

Modell trainiert auf 530 + 538

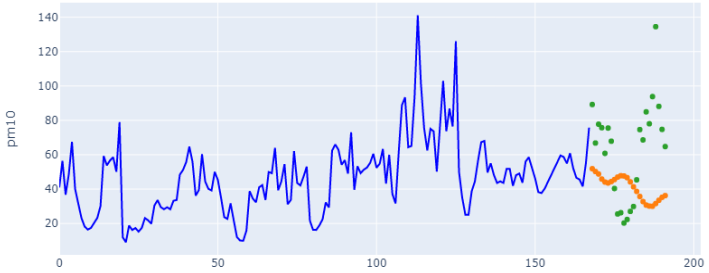


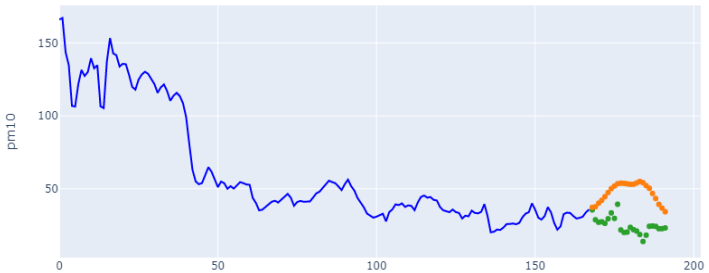


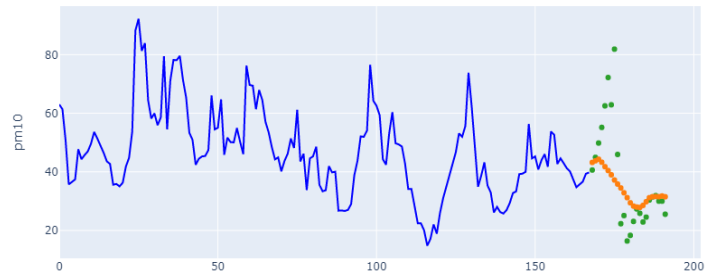


Vorhersage auf Train Daten

Modell trainiert auf 538





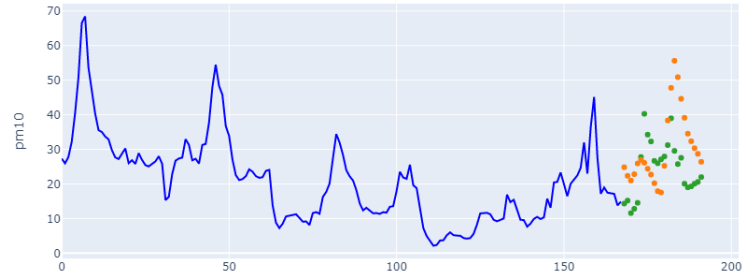


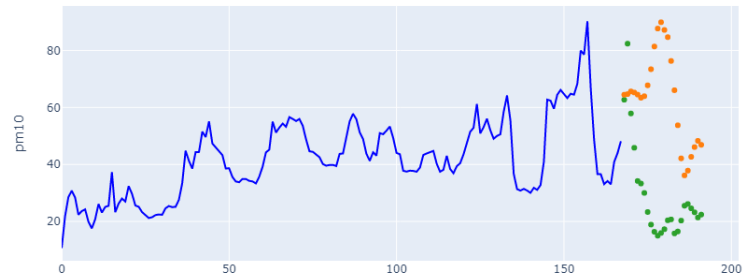
Vorhersage auf Train Daten

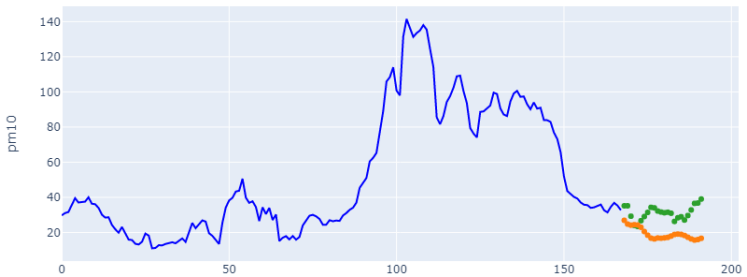
Ergebnisse – Station 530 & 538

Messdaten der Station 538

Modell trainiert auf 530

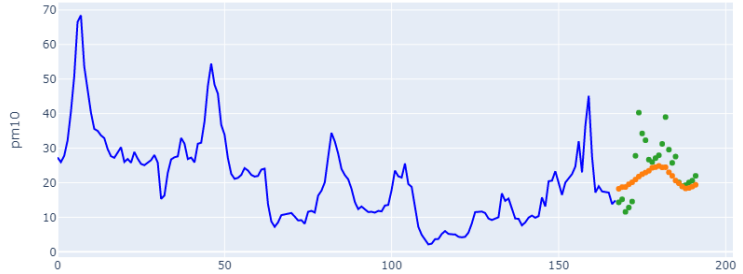


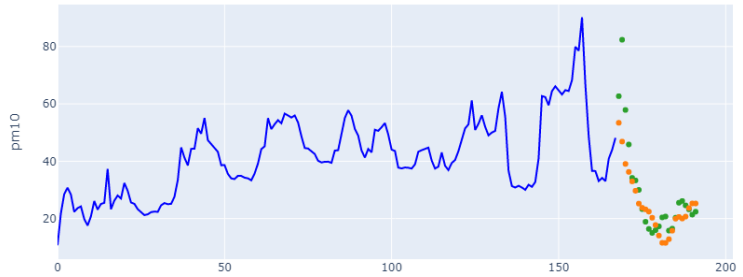


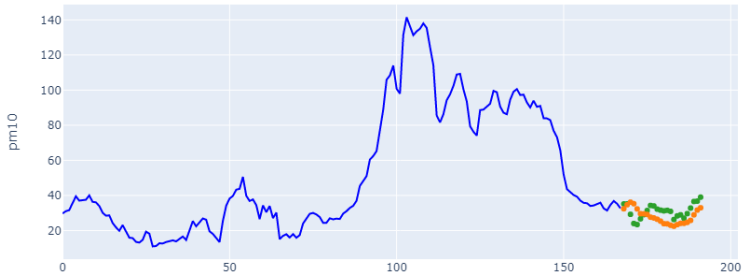


Vorhersage auf Train Daten

Modell trainiert auf 530 + 538

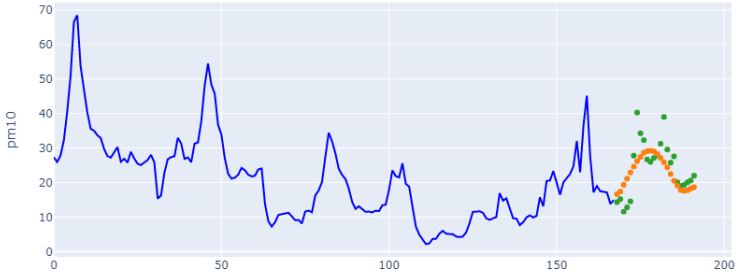


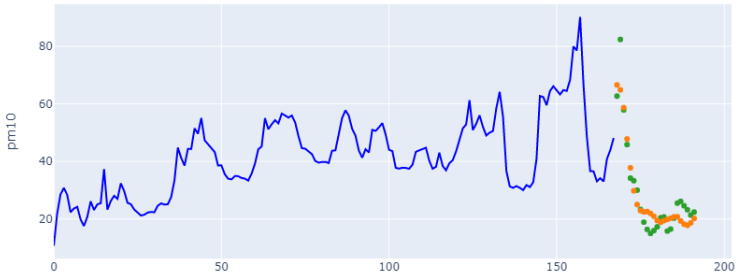


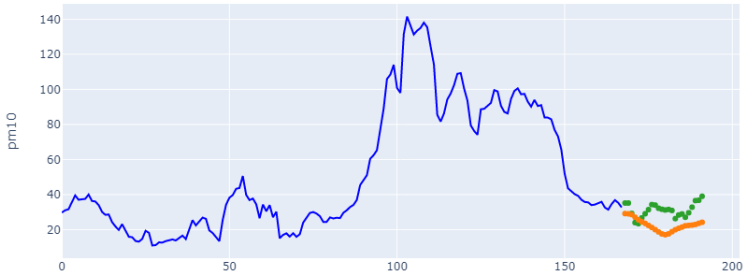


Vorhersage auf Train Daten

Modell trainiert auf 538





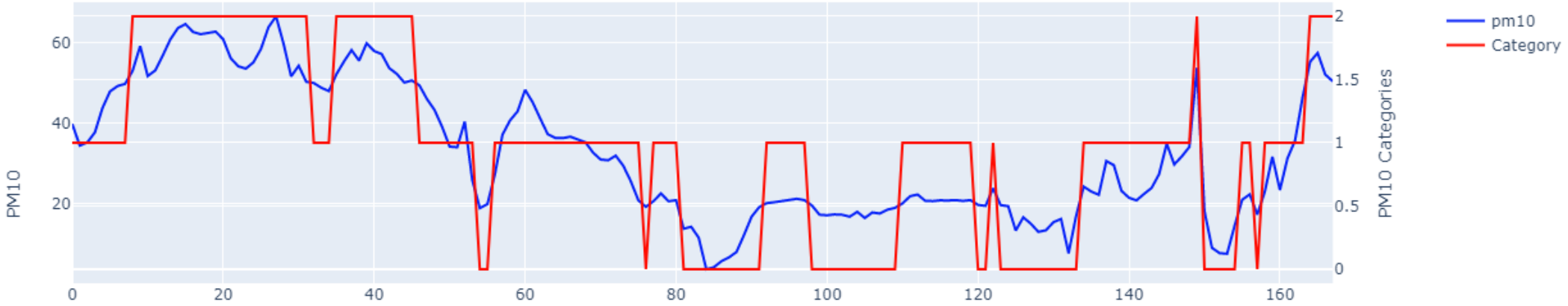


Vorhersage auf Train Daten

Klassifikationsmodell

- Feinstaubwerte können in 6 verschiedene Gruppen aufgeteilt werden
  - 6 Luftqualitätsgruppen für PM10 Werte
- Gleiche Gruppierung wie polnisches Bundesamt
- Feature Vector um Kategorie ergänzt

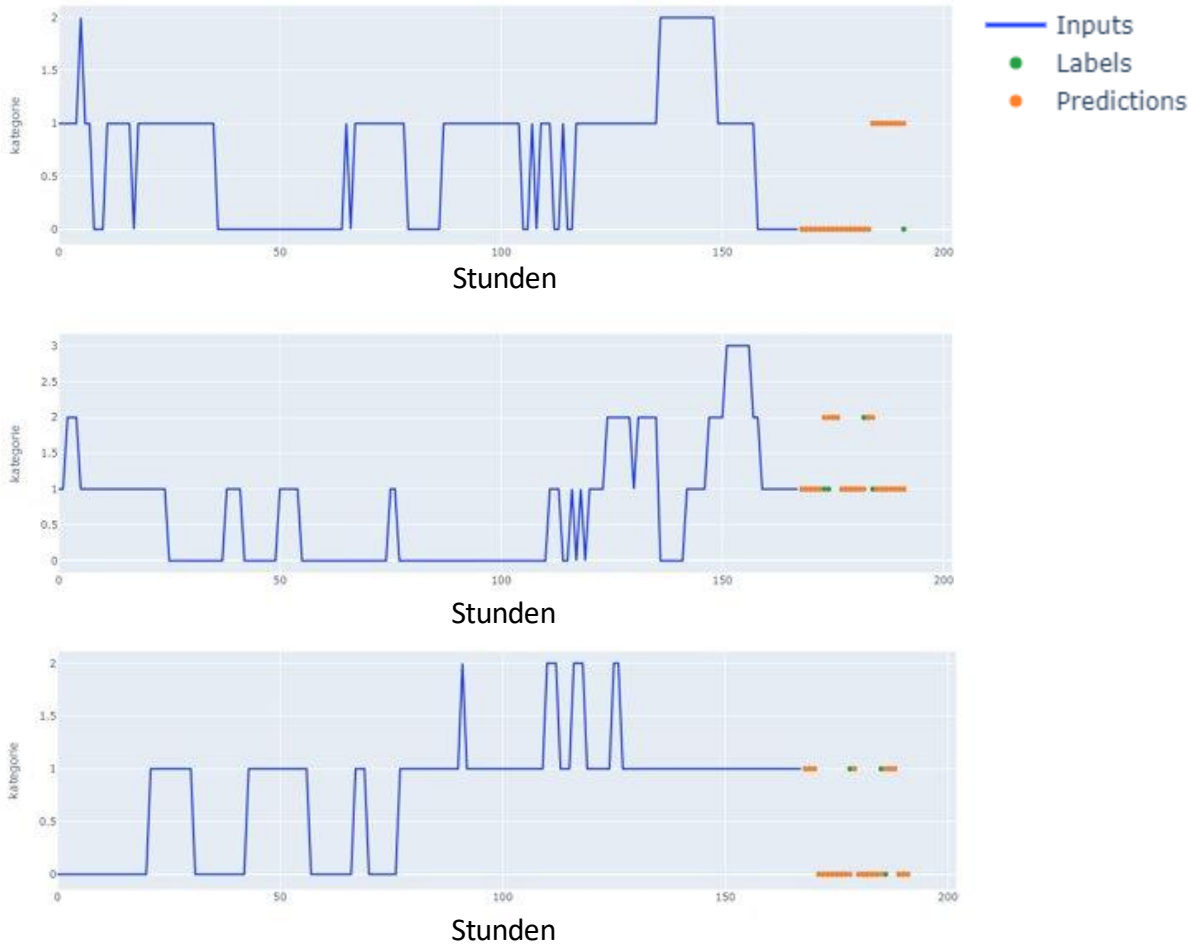
PM_{10} Grenzwerte in µg/m³	Kategorie
0 – 20	Very good
20.1 – 50	Good
50.1 – 80	Moderate
80.1 – 110	Sufficient
110.1– 150	Bad
> 150	Very bad



Klassifikationsmodell

- Änderungen im Vergleich zum besten Modell
  - Dropout 0.4
  - Letzter Dense Layer angepasst auf Klassifikation
    - Softmax Aktivierungsfunktion
- Metriken:
  Train Accuracy: 0.97
  Validation Accuracy: 0.96
  Test Accuracy: 0.59

PM_{10} Vorhersage auf Train Daten (Station 538)



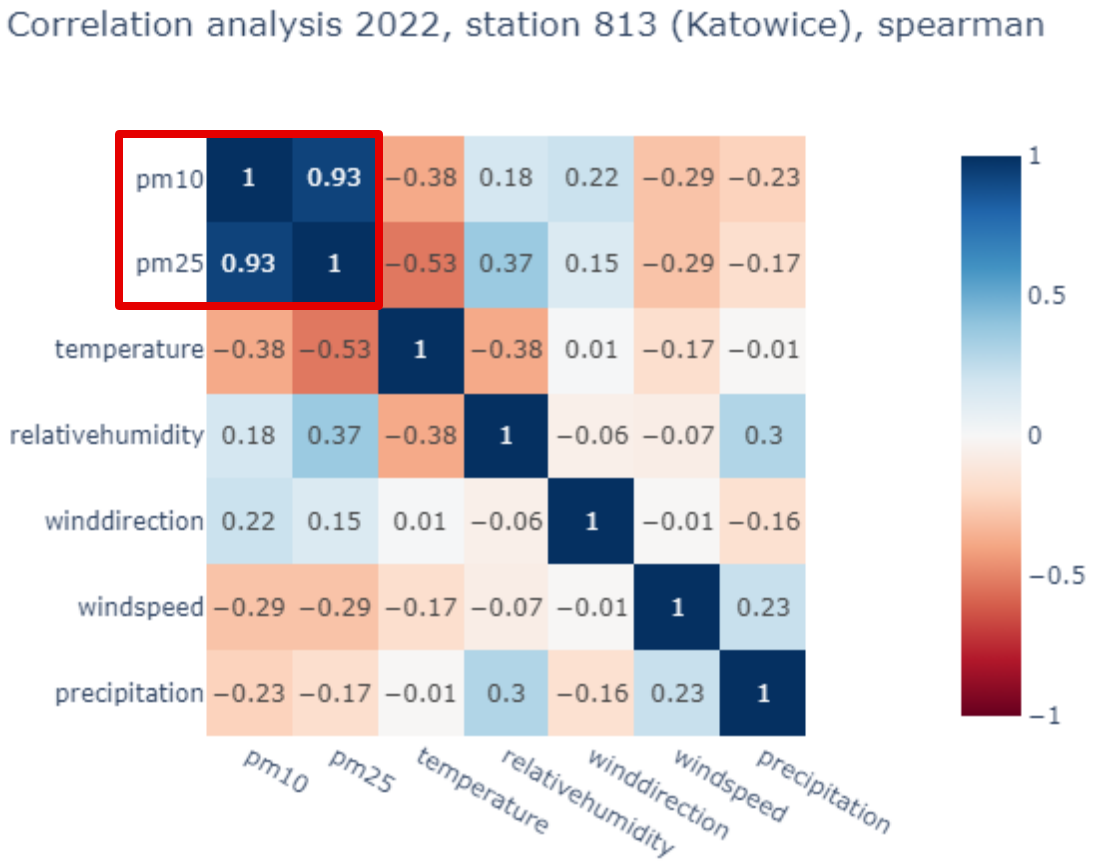
Softwaredemo

Forschungsfragen - Ergebnisse

- Lässt sich mit Hilfe eines neuronalen Netzes unter Verwendung einer CNN-LSTM Kombination eine stündliche Prognose von Feinstaubdaten für die nächsten 14 Tage realisieren?
  - Nein, eine Prognose über 14 Tage ist nicht umsetzbar, der Prädiktionszeitraum ist zu lang
  - Daher die weiteren Forschungsfragen mit Prädiktion von 1 Tag evaluiert
- Ist es möglich den PM_{10} Wert mit einem MAE unter 10 vorherzusagen?
  - Ja, eine Vorhersage von einem Tag ist möglich

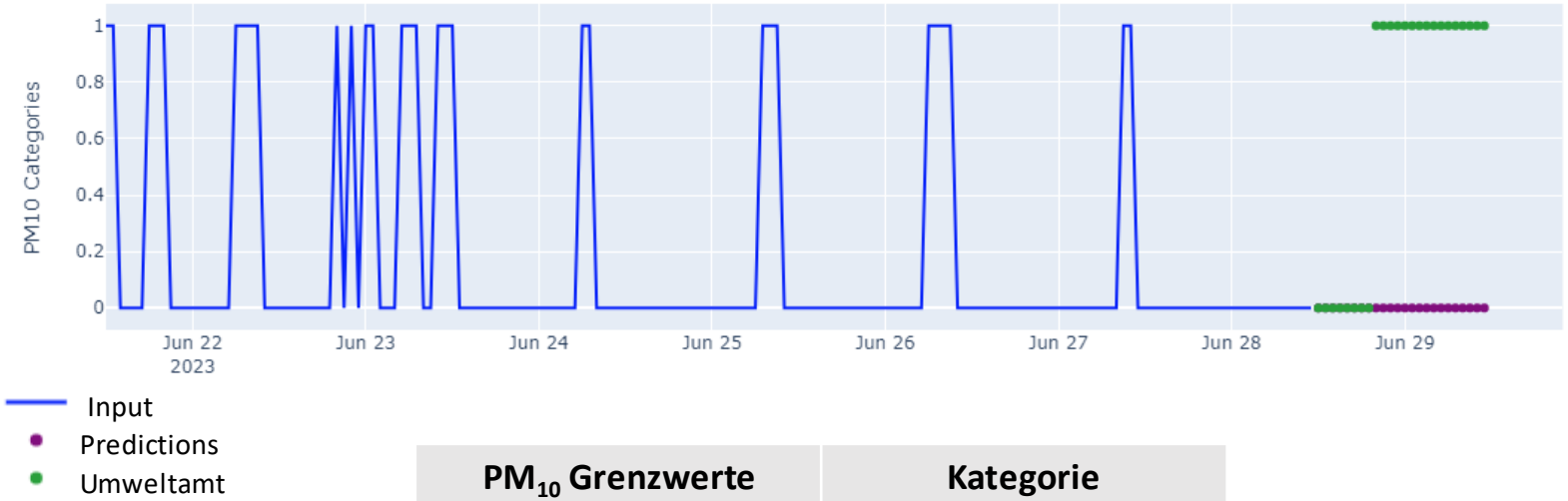
Forschungsfragen - Ergebnisse

- Gibt es einen Zusammenhang zwischen PM_{10} und $PM_{2.5}$, sodass $PM_{2.5}$ mit dem Modell für PM_{10} vorhergesagt werden kann?
  - Aufgrund der starken Korrelation (>0,9) von PM_{10} und $PM_{2.5}$ kann davon ausgegangen werden, dass $PM_{2.5}$ wie PM_{10} funktioniert
- Ist es möglich den $PM_{2.5}$ Wert mit einem MAE unter 10 vorherzusagen?



Forschungsfragen - Ergebnisse

- Wie sehen unsere Prognosen im Vergleich mit denen des polnischen Umweltamts aus? (für einen Tag)



PM_{10} Grenzwerte in µg/m³	Kategorie
0 – 20	Very good
20.1 – 50	Good

Uhrzeit	Modell	Umweltamt	Echte Daten
28.06. 12:00	0	0	0
28.06. 13:00	0	0	0
28.06. 14:00	0	0	0
28.06. 15:00	0	0	0
28.06. 16:00	0	0	1
28.06. 17:00	0	0	0
28.06. 18:00	0	0	0
28.06. 19:00	0	0	0
28.06. 20:00	0	1	0
28.06. 21:00	0	1	0
28.06. 22:00	0	1	0
28.06. 23:00	0	1	0
29.06. 00:00	0	1	0
29.06. 01:00	0	1	0
29.06. 02:00	0	1	0
29.06. 03:00	0	1	0
29.06. 04:00	0	1	1
29.06. 05:00	0	1	1
29.06. 06:00	0	1	1
29.06. 07:00	0	1	1
29.06. 08:00	0	1	1
29.06. 09:00	0	1	1
29.06. 10:00	0	1	1
29.06. 11:00	0	1	0

https://api.gios.gov.pl/pjp-api/v1/rest/data/getData/5376?size=500

Forschungsfragen - Ergebnisse

- Ist es sinnvoll, Stationen zu Gebieten zusammenzufassen, sodass die Aussagekräftigkeit der Prädiktion im Vergleich zu den einzelnen Stationen gleich bleibt oder verbessert wird?
  - Die Prädiktion von Daten einer benachbarten Station ist nicht möglich
  - Aufgrund der Menge an Stationen sind benachbarte Stationen dennoch meist mehrere Kilometer voneinander entfernt
  - Benachbarte Stationen stehen meist auch in unterschiedlichen Umgebungen
  - Trainieren eines Modells mit Daten von zwei Stationen liefert gute Ergebnisse für beide Stationen, die aber schlechter als die stationsspezifischen Modelle sind

Zusammenfassung

- Feinstaub schädlich für Gesundheit
  - Prognose von zukünftigen Werten relevant
- Korrelationsanalyse zwischen verschiedenen Wetterdaten und Feinstaubwerten
  - ➢ Keine ausschlaggebende Korrelation
- Feinstaubdaten gefiltert, vorbereitet und interpoliert
- Data-Windowing
- Neuronales Netz mit den Daten trainiert
- Model jeweils nur für eine Station
  - Zusammenfassen mehrerer Stationen liefert nicht erfolgreiche Ergebnisse
  - Klassifikation liefert bessere Ergebnisse
  - Deutliche Differenz zwischen Ergebnissen von Validation- und Testset
    - Overfitting
    - Verursacht durch Corona-Pandemie?

Ausblick

- Einteilung von Feinstaubdaten in Kategorien sinnvoll
  - Weitere Experimente mit Klassifikationsmodell
- Modelle für $PM_{2.5}$ trainieren
- Inwiefern spiegelt sich die Corona Pandemie in den Feinstaubdaten wider?
- Kann die Feinstaubbelastung in bestimmten Gebieten vorausgesagt werden?
  - Ähnlichkeit der Feinstaubbelastung in ländlichen und/oder urbanen Gebieten
- Wettervorhersagen in den Prädiktionszeitpunkten mit einbinden

➢ Für genauere Prädiktionen ist Aufstellung weiterer Sensoren nötig