Interdisziplinäre und Machine Learning -Grundlagen

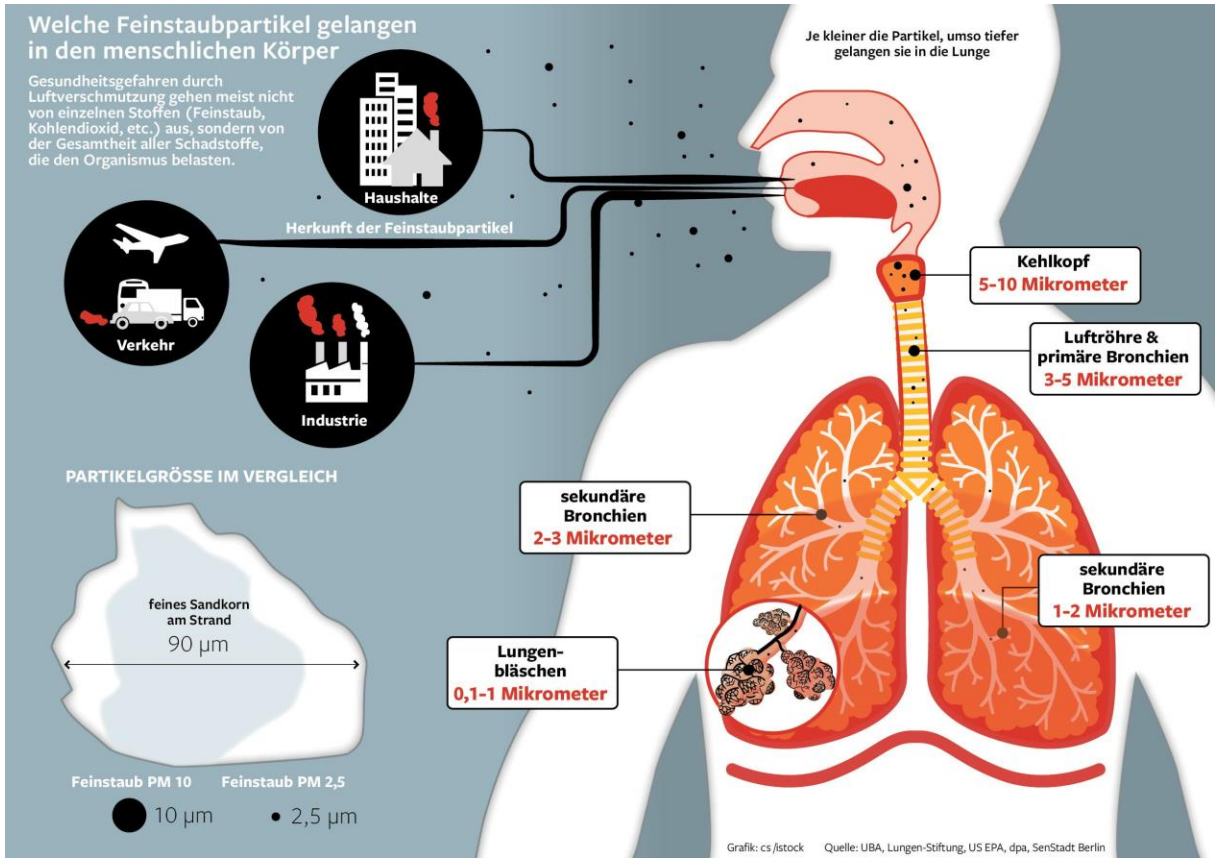
Meilenstein 1

Inhalt

- Motivation
  - Feinstaub
  - Untersuchungsgebiet Polen
  - Messdaten
- Related Work
- Forschungsfrage
- Methodik
  - CNN
  - RNN
  - Modell

Feinstaub

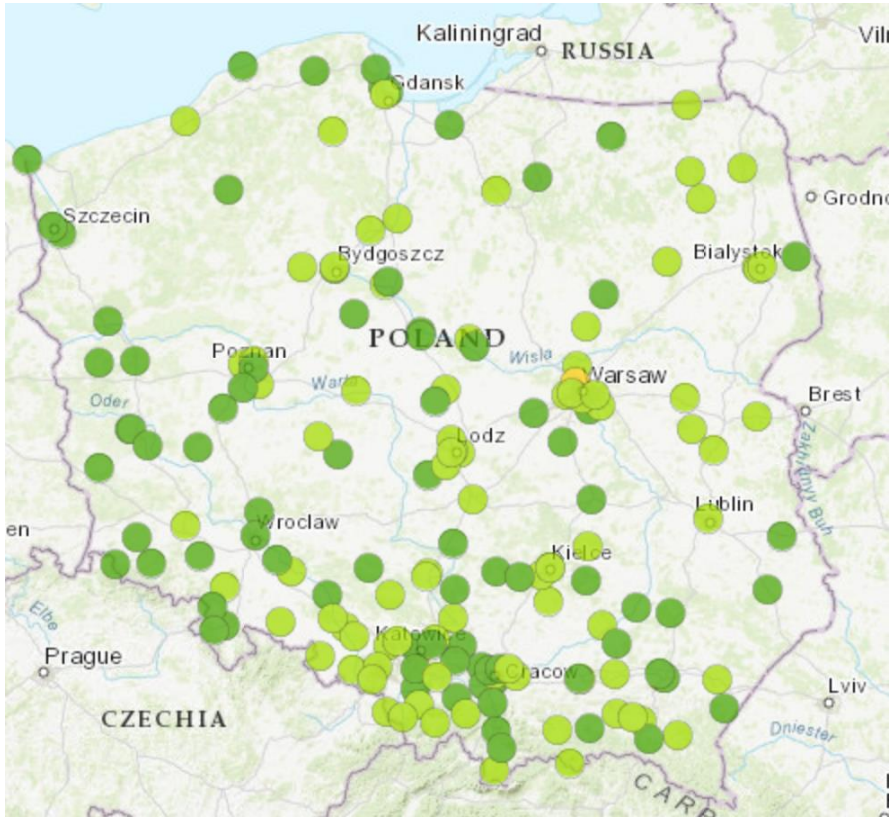
- Feinstaub wird nach Größe unterteilt
  - PM_{10} bezeichnet Partikel mit einem Durchmesser < 10 µm
  - $PM_{2.5}$ ist eine Teilmenge von PM_{10} mit Partikel von einem Durchmesser < 2.5 µm
- Grenzwerte



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

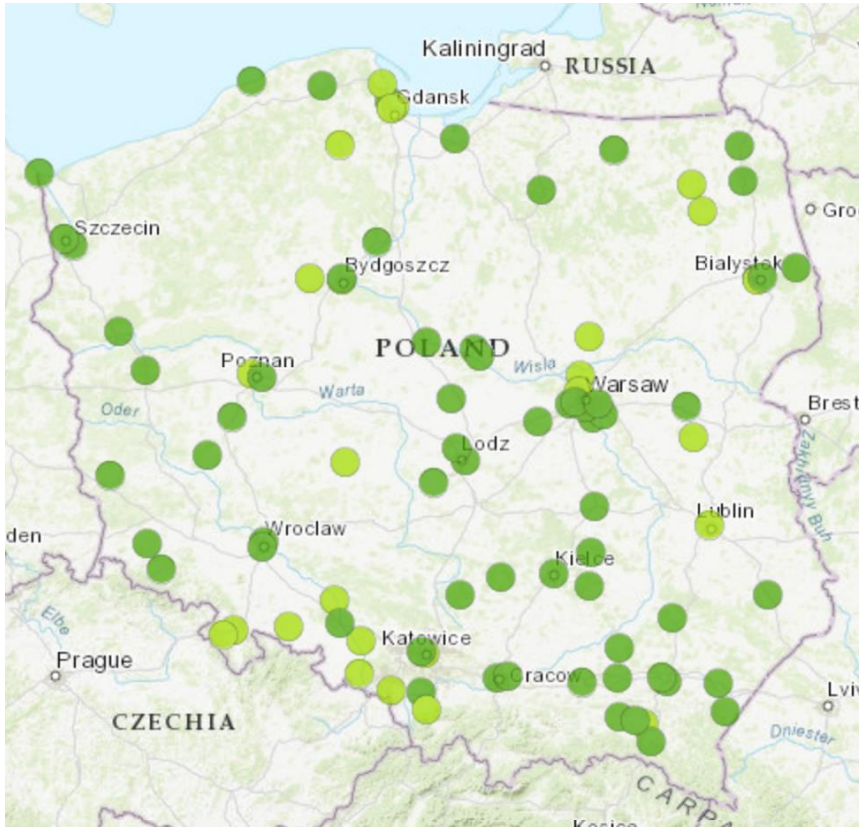
Untersuchungsgebiet Polen

PM_{10}



0 - 20 µg/m^3	Very good
20.1 - 50 µg/m^3	Good
50.1 - 80 µg/m^3	Moderate
80.1 - 110 µg/m^3	Sufficient
110.1 - 150 µg/m^3	Bad
> 150 µg/m^3	Very bad

$PM_{2.5}$



0 - 13 µg/m^3	Very good
13.1 - 35 µg/m^3	Good
35.1 - 55 µg/m^3	Moderate
55.1 - 75 µg/m^3	Sufficient
75.1 - 110 µg/m^3	Bad
> 110 µg/m^3	Very bad

Motivation

Wetterdaten

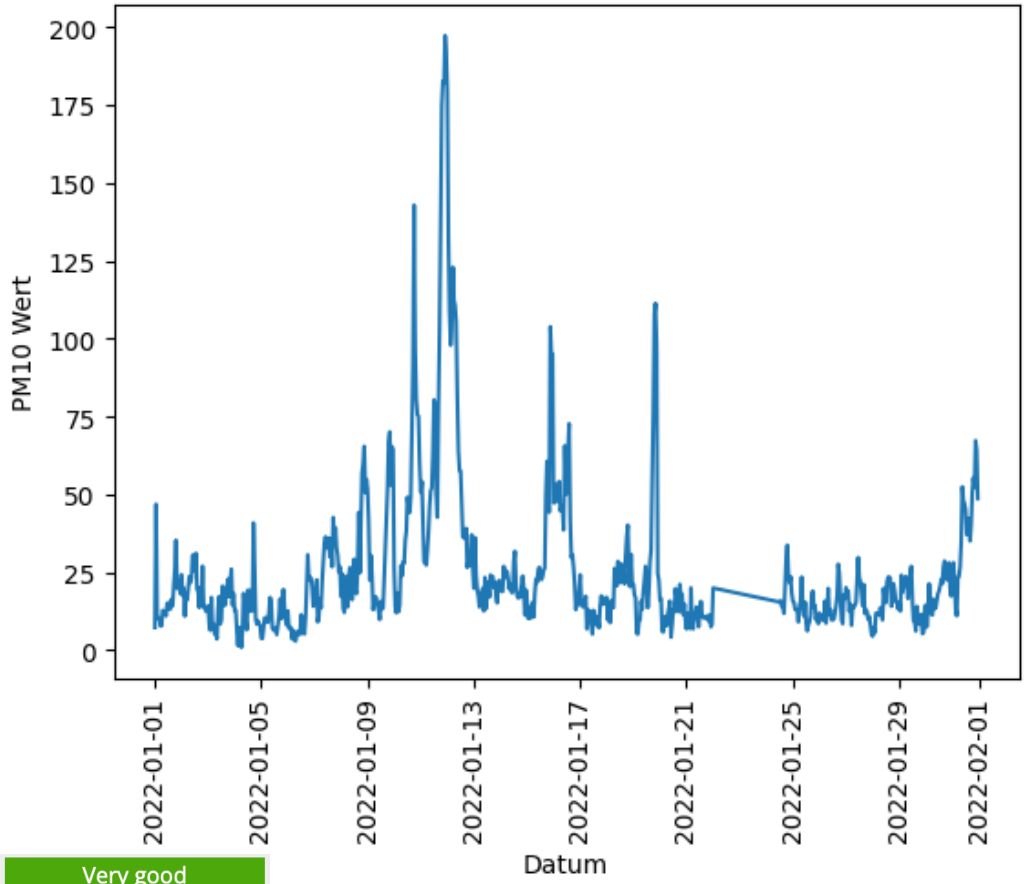
- Temperatur
- Wind/ Luftströmung
- Niederschlag
- Luftfeuchtigkeit

Inversionswetter

- Die oberen Luftschichten sind wärmer als die unteren
- Die kalte Luftschicht ist schwer und durchmischt sich nicht mit der oberen
- Die obere Luftschicht sperrt den Feinstaub wie eine Käseglocke ein

Beispiel Messdaten PM_{10}

Station 52 SensorID 14397

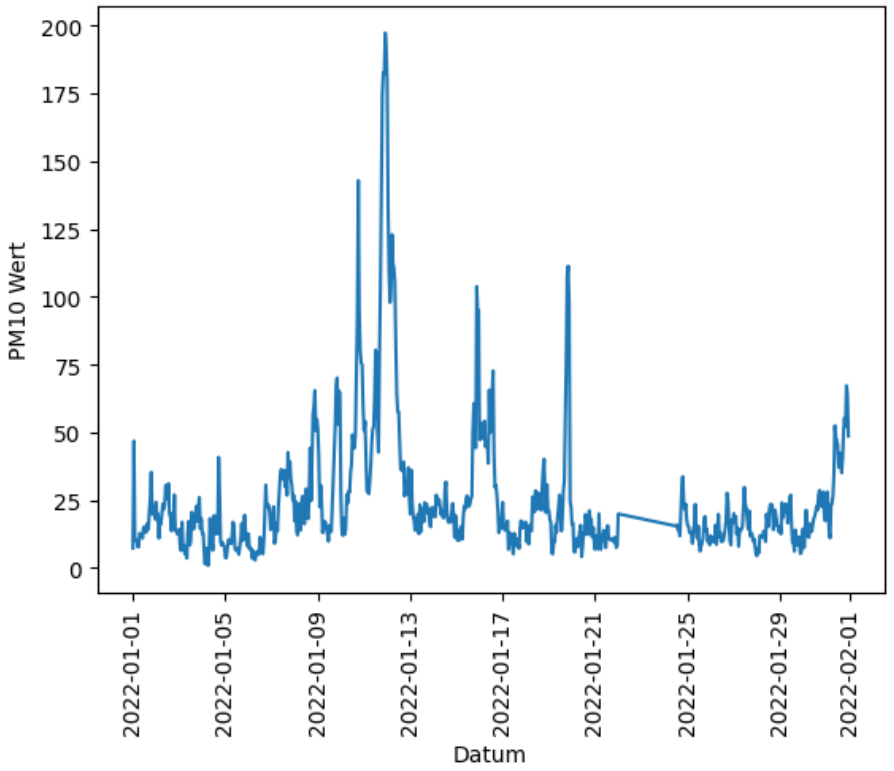


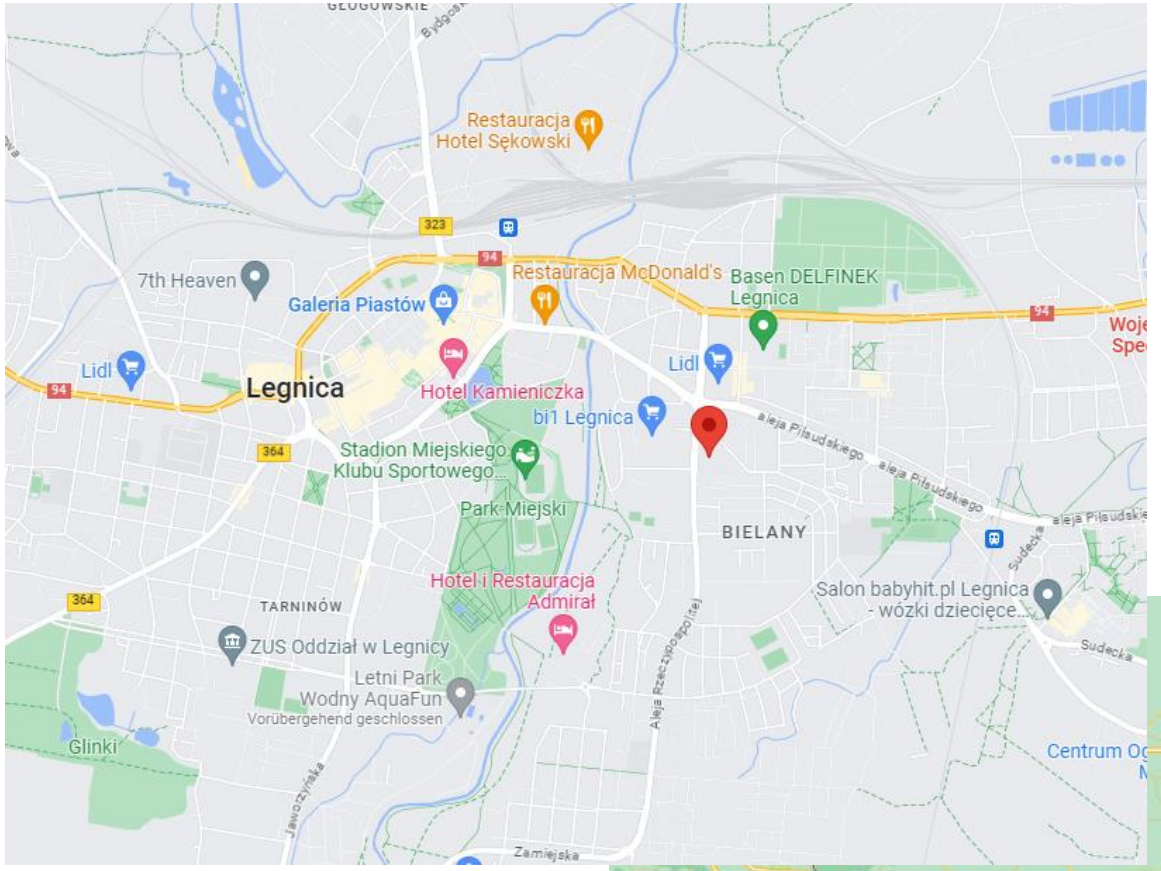
0 - 20 µg/m³	Very good
20.1 - 50 µg/m³	Good
50.1 - 80 µg/m³	Moderate
80.1 - 110 µg/m³	Sufficient
110.1 - 150 µg/m³	Bad
> 150 µg/m³	Very bad

Motivation

Beispiel Messdaten PM_{10}

Station 52 SensorID 14397





Daten

Revisiting air quality forecasting: a regression approach (2018)

- Vorhersage PM_{10} für nächsten Tag pro Messstation
- Daten:
  - PM_{10}
  - Temperatur
  - Luftfeuchtigkeit
  - Tag
  - Monat

Air-pollution prediction in smart city, deep learning approach (2021)

- Stündliche Vorhersage $PM_{2.5}$
- Daten:
  - PM_{10}
  - $PM_{2.5}$
  - Temperatur
  - Luftdruck
  - Regen
  - Windrichtung

Modelle

Revisiting air quality forecasting: a regression approach (2018)

- Regression
- Ein Modell pro Messstation
- Vergleich von Modellen
  - Lineare Regression
  - ANN (1 Hidden Layer (10 Neuronen))
  - Random Forest
  - ➢ 1.LR, 2. RF, 3. ANN
  - ➢ ANN recht klein

Air-pollution prediction in smart city, deep learning approach (2021)

- Regression
- Hybridmodell:
  - CNN (räumliche Merkmale) + LSTM (zeitliche Merkmale)
    - 3 Convolution Layer mit Batch Normalization
    - Maxpooling Layer
    - 2 LSTM Layer (100 und 50 Units)
    - 1 Dense Layer

Ziele

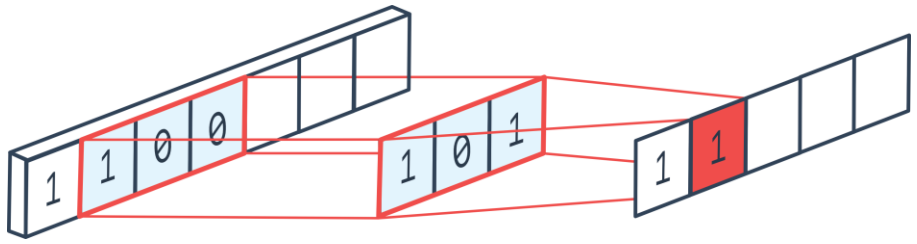
- Prognose von stündlichen Feinstaubdaten innerhalb der nächsten 14 Tage
  - Prädiktion von PM_{10}
    - optional $PM_{2.5}$
  - Aufteilen des Gebietes in Bereiche
    - Prädiktion pro Station
    - Prädiktion für ein Gebiet
  - Wie sehen unsere Prognosen im Vergleich mit denen des polnischen Umweltamts aus? (für einen Tag)
  - Besteht ein Zusammenhang zwischen PM_{10} und $PM_{2.5}$?

Forschungsfragen

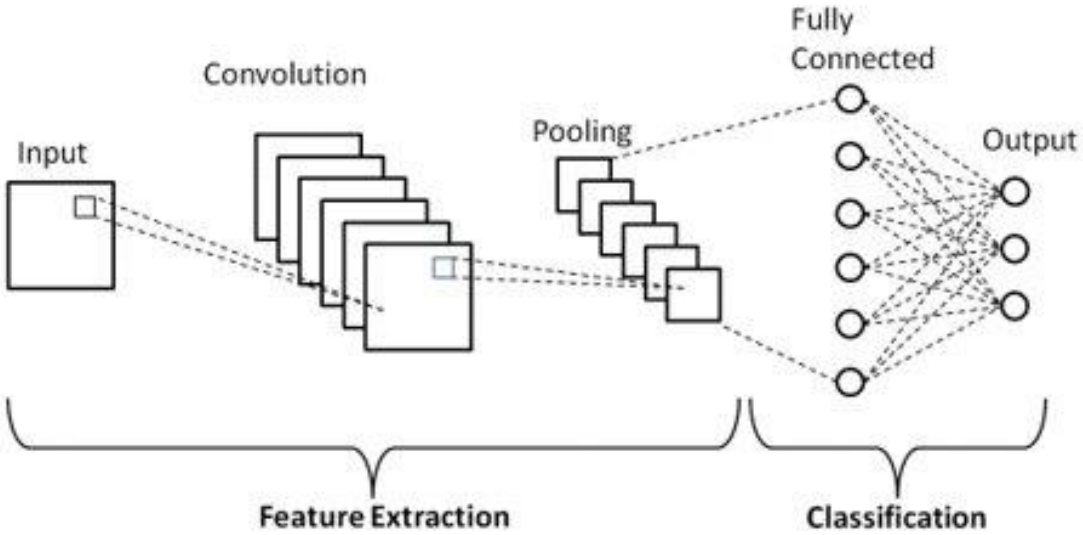
- Lässt sich mit Hilfe eines neuronalen Netzes unter Verwendung einer CNN-LSTM Kombination eine stündliche Prognose von Feinstaubdaten für die nächsten 14 Tage realisieren?
  - ⑩ Ist es möglich den PM_{10} Wert mit einem MAE unter 10 vorherzusagen?
  - ⑩ Ist es möglich den $PM_{2.5}$ Wert mit einem MAE unter 10 vorherzusagen?
  - ⑩ 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?

Convolutional neuronale Netze (CNN)

- maschinelles Verarbeiten von Daten
- Beruht auf Faltungsoperation (Convolution)
- Convolution
  - Berechnung des neuen Feldes durch Maske
  - Filterung der Daten



Quelle: https://ai.stackexchange.com/questions/28767/what-does-channel-mean-in-the-case-of-an-1d-convolution



Quelle: https://www.upgrad.com/blog/basic-cnn-architecture/

Rekurrente neuronale Netze (RNN)

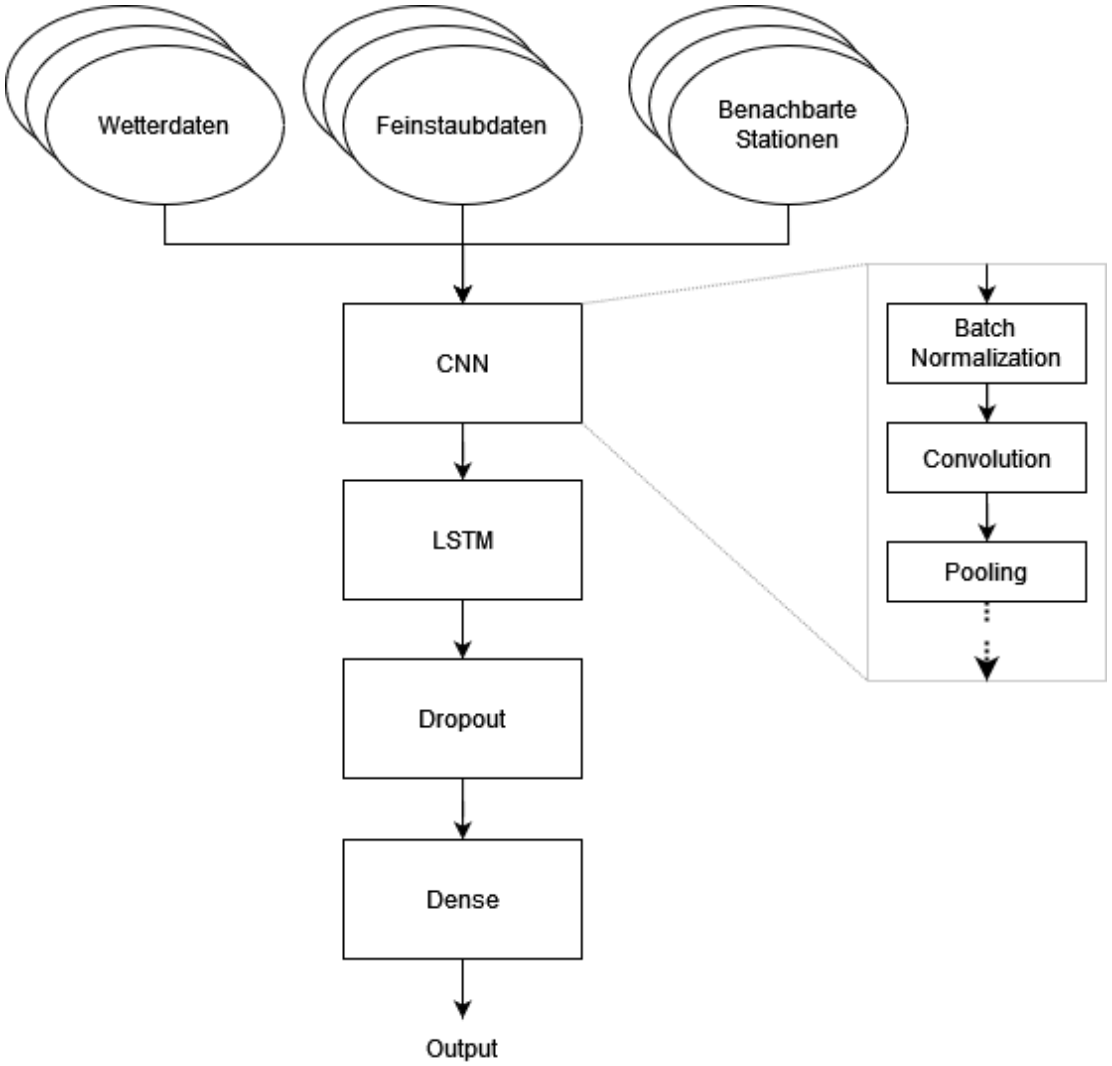
- Neuronale Netze mit Rückkopplungen
  - ➢ Erzeugung von Gehirn im Netz
- Besonders geeignet für zeitliche und sequentielle Daten
- Bekannteste Varianten GRU und LSTM

- LSTM (Long short-term memory network)
  - 1997 von Hochreiter & Schmidhuber entwickelt
  - Lernen von Langzeitabhängigkeiten in sequentiellen Daten
  - Verarbeitet Sequenz von Daten
  - Besteht aus Memory Cells
    - Fungieren als Gedächtnis → Speichern Informationen der vorrangegangenen Daten
  - ➢ Besonders gut geeignet für Feinstaubprädiktion

Modell

- Supervised ML Problem
- CNN-LSTM Kombination zur Lösung des Regressionsproblems
  - Multi-Step Modell
- Multi-Step Forecasting (PM_{10})
  - Input: Merkmalsvektor
  - Output: PM_{10} Wert
- Modell mit MSE als Fehlerfunktion trainieren
- Bewertung des Modells mittels RMSE, MSE und MAE
- Einbeziehen von 3 benachbarten Stationen
  - Gewichtung nach Entfernung
- Trainieren und erstellen des Netzes mit Tensorflow
- Trainingsdaten und Testdaten jeweils 1 Jahr

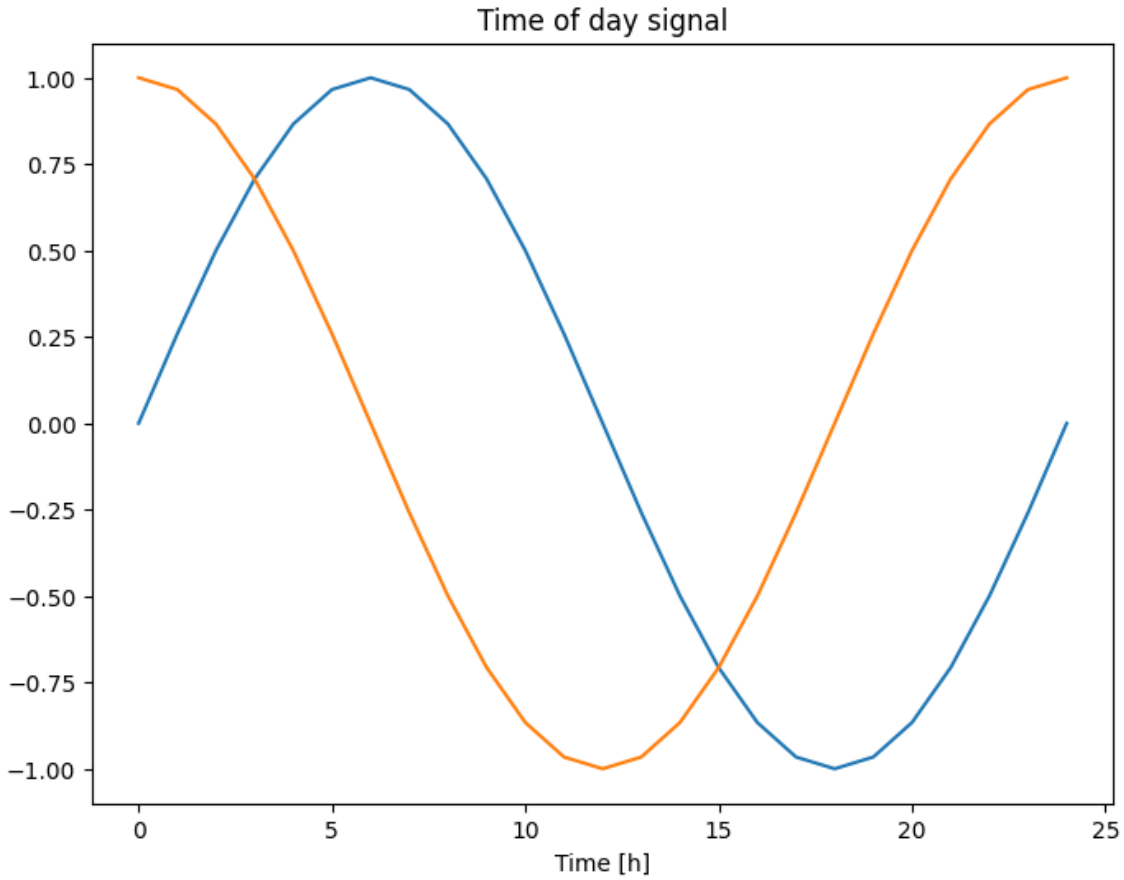
Modell


Wetterdaten
Feinstaubdaten
Benachbarte Stationen
CNN
LSTM
Dropout
Dense
Output
Batch Normalization
Convolution
Pooling

Feature Engineering - Zeit

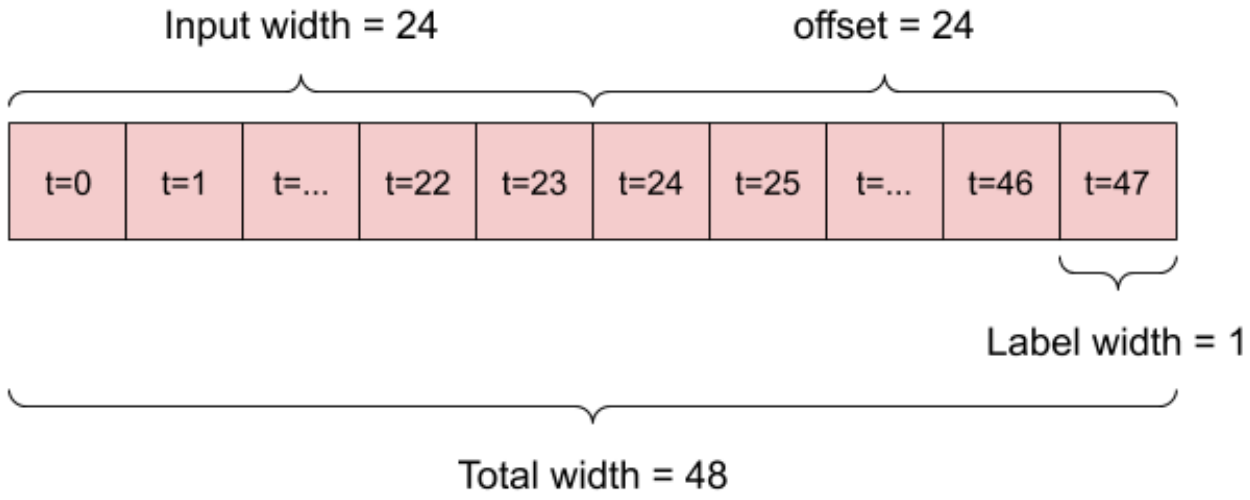
- Zeit periodisch angeben statt absolut
- Perioden für Tag und Jahr
- Für Modell besser verwendbar
- Periode durch Sin und Cos
  - Eindeutigkeit der Werte durch Kombination

- Gleiches auch für Windrichtung relevant
  - Grad in Radiant umwandeln
  - ➢ Abstand zwischen 360° und 0° nicht vorhanden in Radiant



Data Windowing

- Definition des Offset der Vorhersage
- Definition der Inputdaten-Zeitspanne
- Beispiel:
  - Ergebnis, das 24 Stunden in der Zukunft liegt
  - Historie an 24 Stunden Daten



- In unserem Fall:
  - 14-tägige Frames

Multi-step model

1. Single shot predictions
   - Gesamte Vorhersage durch einmalige prediction
2. Autoregressive predictions
   - Single step prediction
   - Ergebnis zur Bestimmung der nächsten prediction
   - Verwendet bei Korrelationen der vorherigen Werte