Hochschule Bielefeld
Fachbereich Campus Minden
Studiengang Informatik

Prädiktion von Feinstaubdaten in Polen

Abschlussbericht

Sommersemester 2023

Vorgelegt von:
Tobias Kaps 1187517
Niklas Lange 1186581
Sarah Flohr 1326721
Svea Worms 1175284
Niklas Theis 1179422

Abgabe am: 14. Juli 2023

Inhaltsverzeichnis

Abbildungsverzeichnis

Tabellenverzeichnis

Abkürzungsverzeichnis

API Application Programming Interface

CNN Convolutional Neural Network

CSV Comma-Separated Values

EU Europäische Union

JSON JavaScript Object Notation

LSTM Long Short-Term Memory

MAE Mean Absolute Error

ML Machine Learning

NRW Nordrhein-Westfalen

PM Particulate Matter

REST Representational State Transfer

UTC Universal Time Coordinated

WHO Weltgesundheitsorganisation

1 Einleitung

Das Projekt im Rahmen des Moduls *Data Mining* hat zum Ziel, zukünftige Feinstaubwerte vorauszusagen. Aufgrund von (zu) hohen Feinstaubwerten in der Europäische Union (EU) und Deutschland ist es sinnvoll, Feinstaubwerte vorherzusagen, um Informationen zur Verfügung zu stellen, damit Menschen sich und ihre Gesundheit schützen können.
Ziel des Projektes ist es, Feinstaubdaten zu sammeln, diese mit Wetterdaten zu konsolidieren und basierend auf diesen Informationen Voraussagen über den Feinstaubwert mithilfe eines neuronalen Netzes zu treffen. Dazu wurden zunächst die Feinstaubdaten des polnischen Umweltamtes und die Wetterdaten von einer Open-Source Wetter Application Programming Interface (API) abgerufen. Diese bilden die Lerngrundlage für das neuronale Netz. Dieses besteht aus einer Kombination aus Convolutional Neural Network (CNN) und Long Short-Term Memory (LSTM) damit die Feinstaubdaten als Zeitreihe gelernt werden können. Die Annahme ist, dass aufgrund von vorausgegangenen zusammenhängenden Wetter- und Feinstaubdaten zukünftige Feinstaubwerte prädiziert werden können. Dabei wurde untersucht, wie weit in die Zukunft eine Vorhersage möglich ist und mit welcher Präzision eine Prädiktion gemacht werden kann.

In dem folgenden Projektbericht wird in Kapitel 2 zunächst näher auf die Definition und Grenzwerte von Feinstaub eingegangen. Des Weiteren wird die Schädlichkeit für die Gesundheit und die Quellen von Feinstaub dargestellt. Kapitel 3 stellt die untersuchten Forschungsfragen vor. Auf das Untersuchungsgebiet und die Datenbeschaffung wird in Kapitel 4 eingegangen. Anschließend werden die Daten in Kapitel 5 untersucht und Analysen hinsichtlich Korrelationen durchgeführt. Kapitel 6 befasst sich mit der Vorverarbeitung der Daten. Dabei werden die Sensordaten untersucht, fehlende Daten identifiziert und mithilfe von Interpolation und Outlier Detection beseitigt. In Kapitel 7 werden spezifische Merkmale der vorverarbeiteten Daten betrachtet und ein Feature Engineering durchgeführt, sodass die Daten ein für das neuronale Netz passendes Format haben. Zudem wird auch der finale Merkmalsvektor beschrieben, der für das neuronale Netz verwendet wird. Der Aufbau des neuronalen Netzes sowie die verschiedenen entwickelten Variationen des Netzes sind in Kapitel 8 näher beschrieben. Kapitel 9 stellt die Ergebnisse der Machine Learning (ML) Experimente dar. Zudem werden die aufgestellten Forschungsfragen beantwortet. Abschließend wird in Kapitel 10 ein Fazit über das Projekt gezogen und ein Ausblick gegeben.

2 Interdisziplinäre Grundlagen

In diesem Kapitel wird genauer auf die Definition von Feinstaub eingegangen. Die Gefahr von Feinstaub für die menschliche Gesundheit nimmt mit sinkender Partikelgröße zu. Des Weiteren werden die Grenzwerte für Feinstaub erläutert, sowie die Gefahren für die Gesundheit weiter betrachtet und die Hauptquellen der Emission von Feinstaub identifiziert.

2.1 Definition und Grenzwerte von Feinstaub

Feinstaub zählt zu den bedeutendsten Umweltfaktoren, die ein erhebliches Risiko für die menschliche Gesundheit darstellen. Die Stärke der Luftverschmutzung steht in direktem Zusammenhang mit Herz-Kreislauf-Erkrankungen, Lungenkrebs und Atemwegserkrankungen. Laut Schätzungen der Weltgesundheitsorganisation (WHO) sind im Jahr 2016 weltweit rund 4,2 Millionen vorzeitige Todesfälle auf Luftverschmutzung zurückzuführen [12]. Die gesundheitlichen Auswirkungen von Feinstaubpartikeln hängen maßgeblich von ihren physikalischen und chemischen Eigenschaften ab. Zum Beispiel können Schwermetalle oder andere krebserregende Stoffe an der Oberfläche haften bleiben und über die Lunge in den Blutkreislauf gelangen. Darüber hinaus stellen die Partikel selbst ein Gesundheitsrisiko dar. Je kleiner die Partikel sind, desto tiefer können sie in die Lunge und das Lymphsystem eindringen. Dadurch wird es unwahrscheinlicher, dass sie wieder ausgeatmet werden oder dass der Körper sie als Fremdkörper erkennt und bekämpft [4].
Feinstaub bezieht sich auf Partikel (Particulate Matter (PM)) in der Luft, die nicht sofort zu Boden sinken, sondern eine gewisse Zeit in der Atmosphäre verweilen. Die Staubpartikel werden je nach ihrem maximalen Durchmesser in verschiedene Kategorien unterteilt. Partikel mit einem Durchmesser kleiner als 10 µm gehören zur Kategorie PM_{10}. Eine Unterkategorie von PM_{10} ist $PM_{2.5}$, in der Partikel mit einem Durchmesser kleiner als 2.5 µm fallen [11]. Feinstaub der Kategorie $PM_{2.5}$ ist im besonderem Maße gesundheitsschädlich. Kleine Partikel können schon in geringer Konzentration die Gesundheit negativ beeinträchtigen. Laut WHO ist Feinstaub immer schädlich - egal bei welcher Konzentration [12]. Im Jahr 2021 wurden die Empfehlungen der WHO für die Grenzwerte von Luftschadstoffkonzentrationen aktualisiert. Die neuen Empfehlungen liegen deutlich unter den bisherigen Grenzwerten. Die Tabelle 2.1 stellt die aktuellen Grenzwerte gemäß der europäischen Luftqualitätsrichtlinie 2008/50/EG den Zielen der WHO gegenüber. Die Tabelle verdeutlicht, dass die WHO-Empfehlungen deutlich niedriger sind als die derzeit gültigen Grenzwerte.

Das Land Nordrhein-Westfalen (NRW) erfasst an verschiedenen Standorten die Luftquali-

tät. Laut dem Luftqualitätsbericht 2021 aus NRW wurde der Jahresmittelwert des PM_{10}-Grenzwertes von 40 μm^3 an allen 66 Messstationen eingehalten. Auch die Anzahl der Tagesüberschreitungen liegt überall unter dem Grenzwert von 35 Tagen pro Jahr. Wenn jedoch die WHO-Grenzwerte betrachtet werden, überschreiten mehr als die Hälfte der PM_{10}-Proben den Jahresmittelwert von 15 μm^3. Nur an 19 Messstellen liegt der Wert bei 15 μm^3 oder darunter. Der Grenzwert gemäß der Europäischen Union für den Jahresmittelwert von $PM_{2.5}$ beträgt 25 μm^3. In NRW werden an 27 Messstationen $PM_{2.5}$-Messungen durchgeführt und alle Messwerte liegen unter dem entsprechenden Grenzwert. Allerdings überschreiten die Messwerte deutlich die Empfehlungen der Weltgesundheitsorganisation (WHO). An allen Messstationen wird ein Jahresmittelwert von mehr als 5 μm^3 gemessen. Bisher erfasst NRW keine Überschreitungen des Tagesgrenzwertes für $PM_{2.5}$ [5].

Schadstoff	Mittelungszeitraum	EU	WHO
Feinstaub PM_{10} in μm^3	1 Jahr	40	15
	24 Stunden	50	45
	Erlaubte Überschreitung	35 Tage/Jahr	3-4 Tage/Jahr
Feinstaub $PM_{2.5}$ in μm^3	1 Jahr	25	5
	24 Stunden	-	15
	Erlaubte Überschreitung	-	3-4 Tage/Jahr

Tabelle 2.1: Gegenüberstellung der geltenden, rechtlich verbindlichen Grenzwerte der europäischen Luftqualitätsrichtlinie 2008/50/EG mit den WHO 2021 Empfehlungen [5].

2.2 Auswirkungen der Feinstaubbelastung

Die Feinstaubbelastung ist in Städten der EU besonders hoch und beeinträchtigt damit einen Großteil der Menschen in Europa. Laut Schätzungen der Europäischen Umweltagentur verstarben im Jahr 2020 in der EU mindestens 238.000 Menschen vorzeitig, weil sie $PM_{2.5}$-Konzentrationen von über 5 $\mu g/m^3$ ausgesetzt waren.
Die hohe Luftverschmutzung hat nicht nur vorzeitige Todesfälle zur Folge, sondern verursacht auch hohe Kosten im Gesundheitswesen aufgrund von massiven Gesundheitsproblemen. Die hohe Exposition gegenüber $PM_{2.5}$ führte im Jahr 2019 in 30 europäischen Ländern zu insgesamt 175.702 verlorenen gesunden Lebensjahren. Die gesundheitlichen Probleme sind dabei vor allem auf chronisch obstruktive Lungenerkrankungen zurückzuführen.
Trotzdem ist zu erwähnen, dass die Zahl der $PM_{2.5}$-bedingten vorzeitigen Todesfälle in der EU im Zeitraum von 2005 bis 2020 um 45 % zurückgegangen sind. Die EU hat das Ziel,

den Schadstoffausstoß bis zum Jahr 2050 auf ein Niveau zu senken, das nicht mehr als gesundheitsgefährdend einzustufen ist. Wenn der negative Trend weiter beibehalten wird, könnte schon im Jahr 2030 eine $PM_{2.5}$-Reduktion von 55 % erreicht werden [1].

2.3 Hauptquellen der Feinstaubbelastung

Die Hauptquelle der Feinstaubbelastung in Europa ist die Verbrennung von Brennstoffen im Wohn-, Gewerbe- und institutionellen Bereich. Diese Emissionen entstehen vor allem durch Verbrennungen für Heizzwecke. Im Jahr 2020 waren 44 % der PM_{10}- und 58 % der $PM_{2.5}$-Emissionen auf das Heizen zurückzuführen [1].
Für Deutschland entfallen im Jahr 2021 rund 41,5 % der PM_{10}-Emissionen auf Produktionsprozesse, besonders bei der Herstellung von Metallen und mineralischen Produkten. Auf den Straßenverkehr und Verbrennungsprozessen aus Haushalten sind 36,5 % der Emissionen zurückzuführen. In der Landwirtschaft entstehen besonders viele Emissionen bei der Tierhaltung und der Bearbeitung von landwirtschaftlichen Böden. So sind etwa 37 % der PM_{10}-Emissionen auf die Tierhaltung, vor allem von Rindern und Milchkühen, zurückzuführen. Ganze 63 % der landwirtschaftlichen Emissionen gehen auf die Bodenbearbeitung zurück [3].

3 Forschungsfragen

Wie bereits in den interdisziplinären Grundlagen erläutert, stellt Feinstaub ein großes Risiko für die menschliche Gesundheit dar. Dieses Projekt hat daher das Ziel, zukünftige Feinstaubwerte vorherzusagen. Um dieses Ziel zu erreichen, ist es von entscheidender Bedeutung, präzise Forschungsfragen zu formulieren, die den Fokus und die Ausrichtung der Untersuchungen definieren.
Im Rahmen des Projektes wurde eine primäre Forschungsfrage erarbeitet, die in eine Reihe von weiteren Forschungsfragen untergegliedert ist. Die primäre Forschungsfrage lautet wie folgt:

Forschungsfrage 3.1 *Lässt sich mithilfe eines neuronalen Netzes, unter Verwendung einer CNN-LSTM Kombination, eine stündliche Prognose von Feinstaubdaten für die nächsten 14 Tage realisieren?*

Es wird eine CNN-LSTM Kombination gewählt, da bereits andere Forschergruppen mit dieser Architektur gute Ergebnisse bei der Prädiktion von Feinstaubdaten erzielt haben [2]. Damit die Performanz der Prognosen des neuronalen Netzes analysiert und in Relation zu Prognosen anderer Forschergruppen gesetzt werden kann, wird als Bewertungskriterium der Mean Absolute Error (MAE) herangezogen. Es werden zwei weitere Forschungsfragen gestellt, die als Ziel haben den Feinstaubwert mit einem MAE unter 10 vorherzusagen. Dabei wird sowohl der PM_{10}- als auch der $PM_{2.5}$-Wert untersucht.

Forschungsfrage 3.1.1 *Ist es möglich den PM_{10}-Wert mit einem MAE unter 10 vorherzusagen?*

Forschungsfrage 3.1.2 *Ist es möglich den $PM_{2.5}$-Wert mit einem MAE unter 10 vorherzusagen?*

Des Weiteren wird der Zusammenhang zwischen PM_{10} und $PM_{2.5}$ untersucht. Dabei liegt der Fokus darauf, beide Feinstaubwerte mit dem gleichen Modell zu prädizieren. Die konkrete Forschungsfrage lautet wie folgt:

Forschungsfrage 3.1.3 *Gibt es einen Zusammenhang zwischen PM_{10} und $PM_{2.5}$, sodass $PM_{2.5}$ mit dem Modell für PM_{10} vorhergesagt werden kann?*

In diesem Projekt wird mit Feinstaubdaten aus Polen gearbeitet. Das polnische Umweltamt, welche die historischen Feinstaubdaten bereitstellt, stellt auch aktuelle Prognosen für die nächsten 24 Stunden zur Verfügung. Daher wird untersucht, wie gut die Prognose des neuronalen Netzes gegenüber der Prognose des polnischen Umweltamtes abschneidet.

Forschungsfrage 3.1.4 *Wie sehen unsere Prognosen für einen Tag im Vergleich mit denen des polnischen Umweltamtes aus?*

Das polnische Umweltamt erfasst Feinstaubwerte an vielen verschiedenen Orten des Landes. Daher bietet sich die Möglichkeit, konkrete Modelle nur für spezifische Stationen zu trainieren. Besser wäre es jedoch ein Modell zu erstellen, welches in der Lage ist Vorhersagen für einen größeren Bereich mit mehreren Stationen zu treffen. Die Fragestellung, ob dies sinnvoll ist, wird mit der letzten Forschungsfrage untersucht.

Forschungsfrage 3.1.5 *Ist es sinnvoll, Stationen zu Gebieten zusammenzufassen, sodass die Aussagekraft der Prädiktion im Vergleich zu den einzelnen Stationen gleich bleibt oder verbessert wird?*

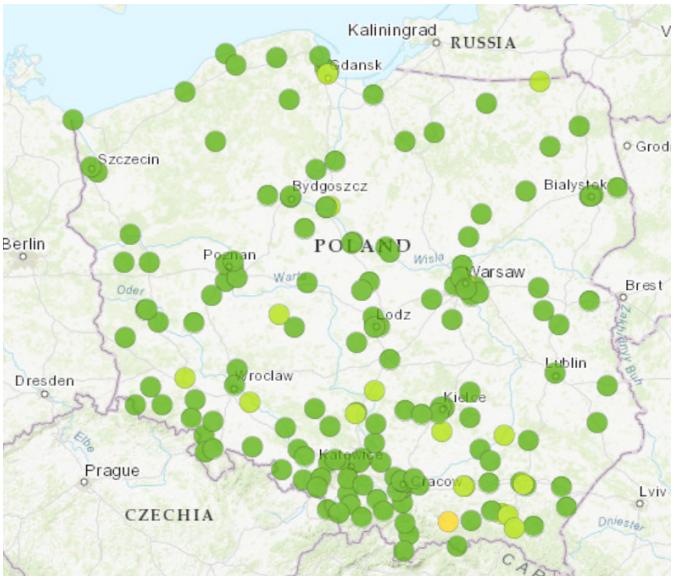
Die aufgestellten Forschungsfragen werden im weiteren Verlauf der Arbeit aufgegriffen und untersucht. Abschließend werden die Forschungsfragen in Kapitel 9 beantwortet.

4 Datenbeschaffung

In diesem Kapitel wird das Untersuchungsgebiet für Feinstaub in Polen vorgestellt. Des Weiteren wird erklärt, wie die Feinstaubdaten beschafft werden. Außerdem werden neben den Feinstaubdaten auch Wetterdaten abgerufen.

4.1 Untersuchungsgebiet

Für das Untersuchungsgebiet wird das Land Polen gewählt. Das polnische Umweltamt betreibt mehrere Messstationen für verschiedenen Emissionen. Neben den PM_{10} und $PM_{2.5}$-Werten, werden auch Schwefeldioxid, Stickstoffoxid und Ozon erfasst. Das polnische Umweltamt stellt die entsprechenden Daten auf ihrer Website [7] zur Verfügung. Für das Projekt werden nur PM_{10}- und $PM_{2.5}$-Werte betrachtet.
Abbildung 4.1 zeigt alle PM_{10}-Messstationen sowie die Kategorisierung der Gesundheitsschädlichkeit. In Abbildung 4.1a sind die PM_{10}-Stationen und ihr jeweiliger Standort zu sehen. Die Legende in Abbildung 4.1b zeigt die zugehörige Kategorisierung der PM_{10}-Werte. Die einzelnen Kategorien werden vom polnischen Umweltamt getroffen. Sie sind nicht am WHO-Standard orientiert, jedoch strenger als die EU-Richtlinien.



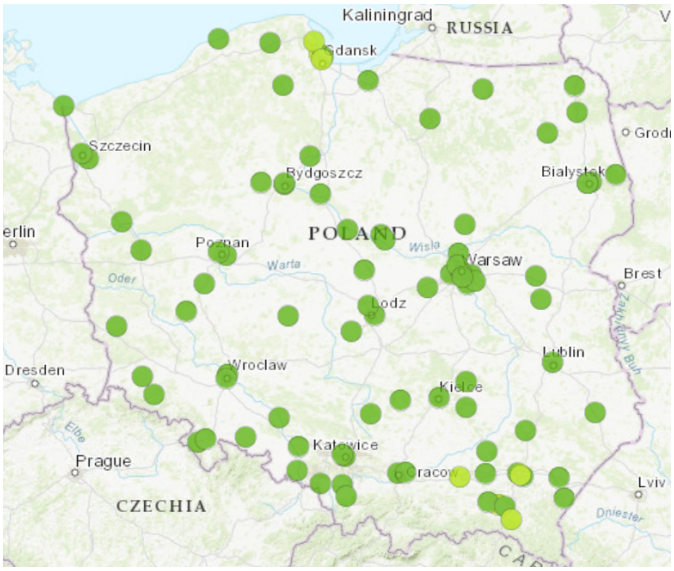
(a) PM_{10}-Messstationen.

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

(b) Legende PM_{10}-Werte.

Abbildung 4.1: Alle Stationen die den PM_{10}-Wert in Polen messen inklusive der entsprechenden Kategorisierung der Gesundheitsschädlichkeit [7].

Abbildung 4.2 zeigt alle $PM_{2.5}$-Messstationen sowie die zugehörige Kategorisierung und ist gleich zu interpretieren.



(a) $PM_{2.5}$-Messstationen.

0 - 13 µg/m³	Very good
13.1 - 35 µg/m³	Good
35.1 - 55 µg/m³	Moderate
55.1 - 75 µg/m³	Sufficient
75.1 - 110 µg/m³	Bad
> 110 µg/m³	Very bad

(b) Legende $PM_{2.5}$-Werte.

Abbildung 4.2: Alle Stationen die den $PM_{2.5}$-Wert in Polen messen inklusive der entsprechenden Kategorisierung der Gesundheitsschädlichkeit [7].

4.2 Daten-Crawling

Um die Forschungsfragen, die in Kapitel 3 aufgeworfen wurden, umfassend zu beantworten, ist die Beschaffung von hochwertigen Feinstaub- und Wetterdaten von entscheidender Bedeutung. In diesem Abschnitt werden die verschiedenen Schritte und Herausforderungen bei der Datenbeschaffung erläutert. Das Forschungsprojekt konzentriert sich geografisch auf das Land Polen, weshalb die Beschaffung der Feinstaubdaten vorrangig über die Representational State Transfer (REST)-API des polnischen Umweltamtes erfolgt. Die Webseite des Umweltamtes stellt neben der API auch andere Möglichkeiten zur Verfügung, um historische Daten abzufragen, die für die Analyse von Bedeutung sind. Ältere Daten, insbesondere aus den Jahren 2019 bis 2021, werden als Excel-Dateien zum Download angeboten.
Für neuere Daten ist jedoch ausschließlich der Zugriff über die REST-API-Schnittstelle möglich. Im Gegensatz zu den Excel-Dateien stellt die API die Daten nicht als Gesamtdatensatz aller Stationen und Sensoren zur Verfügung. Stattdessen werden die Daten pro Sensor zur Verfügung gestellt. Dies stellt eine gewisse Herausforderung bei der Datenbeschaffung für das Jahr 2022 dar. Zunächst werden alle verfügbaren Stationen ermittelt und die dazugehörigen

Sensoren inklusive der Metadaten, wie Koordinaten und Positionierung, erfasst. Anhand dieser Informationen können mithilfe der Sensor-ID die Daten für jeden einzelnen Sensor abgerufen werden.
Die Daten werden paginiert und als JavaScript Object Notation (JSON)-Response geliefert. Die Response enthält mehrere Links, die einerseits auf den nächsten Datensatz (Page) verweisen und andererseits auf die erste oder letzte Seite des Datensatzes. Basierend auf diesen Daten und Informationen wird ein automatisierter Crawling-Lauf gestartet, der die vom Umweltamt festgelegte Einschränkung von maximal zwei Anfragen pro Minute berücksichtigt [8]. Die erhaltenen Daten werden pro Station in Comma-Separated Values (CSV)-Dateien gespeichert, um sie später im Verlauf des Projekts für weitere Analysen nutzen zu können.
Die Wetterdaten werden von der Open-Source Wetter-API Open-Meteo [6] bezogen. Open-Meteo stellt dabei Wetterdaten bereit, welche anhand von Wettermodellen berechnet werden. Für die Berechnungen werden als Datenquellen Wetterstation, Sensoren an Flugzeugen, Radarmessungen und Messungen von Satelliten genutzt. Dadurch stehen für jede Koordinate der jeweiligen Feinstaubmessstation präzise Wetterdaten zur Verfügung. Dies ist wichtig, da die Wetterstationen meist nicht an denselben Orten stehen, wie die Feinstaubmessstationen.

5 Datenanalyse

Abbildung 5.1 zeigt den durchschnittlichen PM_{10}-Wert pro Tag für die Jahre 2019 bis 2022. Die Werte wurden an Station 814 gemessen, welche in Kattowitz liegt.
Es ist deutlich ersichtlich, dass während der Wintermonate eine Abnahme der Luftqualität festzustellen ist. Diese Tatsache kann vermutlich auf die Heizperiode zurückgeführt werden. Zusätzlich werden durch grau hinterlegte Bereiche deutliche Lücken in den Daten sichtbar. Ein solcher Fall tritt beispielsweise im Oktober 2019 auf, wo nahezu der gesamte Monat fehlt. Des Weiteren fehlen auch Daten für das erste Wochenende im August 2021. Fehlende Daten sind für die Vorhersage zukünftiger Feinstaubdaten problematisch, vor allem wenn größere Zeitabschnitte fehlen, da dem neuronalen Netz die Lerngrundlage genommen wird. Daher wird in Kapitel 6 näher auf die Interpolation und in Kapitel 8 auf den Umgang mit fehlender Kontinuität in den Datensets eingegangen.

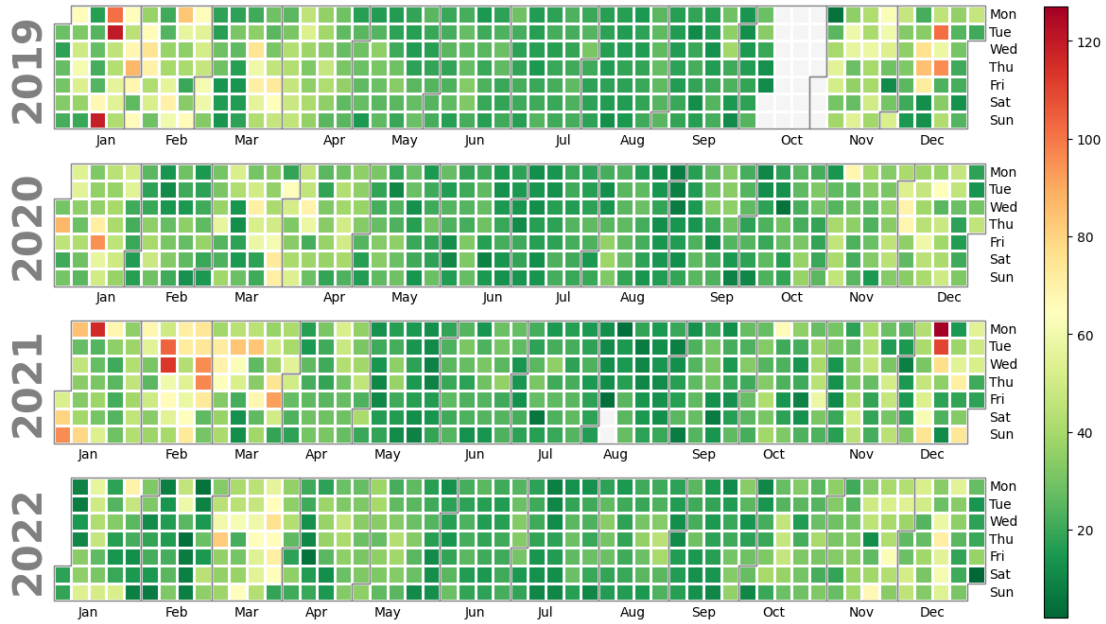
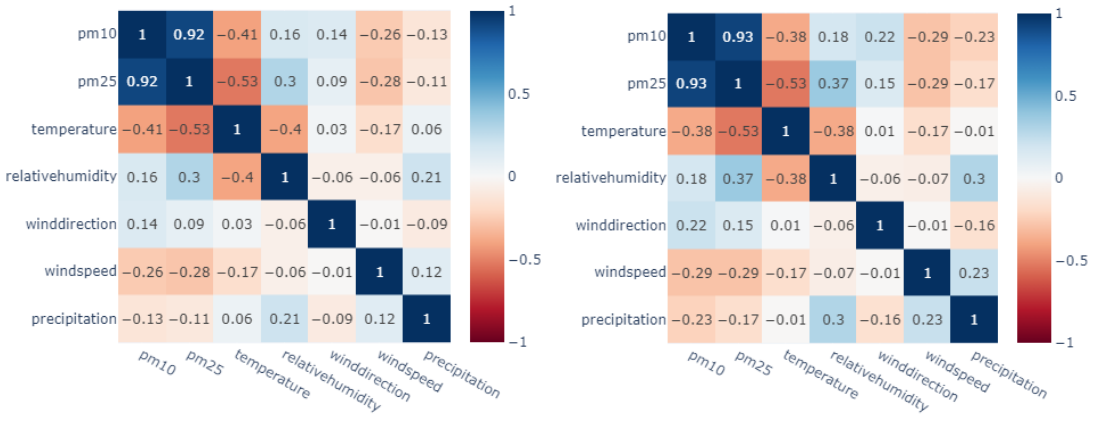


Abbildung 5.1: Durchschnittlicher PM_{10}-Wert pro Tag für die Jahre 2019 bis 2022, gemessen an Station 814 in Kattowitz.

Neben den Feinstaubdaten werden auch die entsprechenden Wetterdaten für die jeweilige Station betrachtet. Daraus ergibt sich die Frage, ob ein oder mehrere Merkmale zur Vorhersage anderer Merkmale verwendet werden können. In diesem Fall ist das Ziel zu untersuchen, ob anhand einzelner Wetterdaten die Feinstaubwerte vorhergesagt werden können. Dazu wurde eine Korrelationsanalyse auf Basis der verschiedenen Werte durchgeführt, um einen

Zusammenhang zwischen den Variablen zu erkennen.
Abbildung 5.2 zeigt die Korrelationsanalyse für die Station 813 für das gesamte Jahr 2022. Betrachtet werden der PM_{10}- und der $PM_{2.5}$-Wert, sowie Wetterdaten mit Temperatur, relativer Feuchtigkeit, Windrichtung, Windgeschwindigkeit und Niederschlag. Abbildung 5.2a zeigt die Korrelationsanalyse nach Pearson und Abbildung 5.2b zeigt die Korrelationsanalyse nach Spearman auf den gleichen Daten. Den Abbildungen ist zu entnehmen, dass es keine (signifikante) Korrelationen zwischen den Variablen gibt, bis auf die Korrelation zwischen den PM_{10}- und den $PM_{2.5}$-Werten. Das bedeutet im Umkehrschluss, dass die Wetterdaten wahrscheinlich wenig Einfluss auf die Vorhersage der Feinstaubwerte haben.



(a) Pearson-Korrelationsanalyse.

(b) Spearman-Korrelationsanalyse.

Abbildung 5.2: Korrelationsanalyse: Wetterdaten und Feinstaubwerte für Station 813 in Kattowitz im Jahr 2022.

Für andere Stationen und Zeitabschnitte ergibt sich das gleiche Ergebnis der Korrelationsanalyse. In keinem Fall wird eine signifikante Korrelation zwischen Feinstaub- und Wetterdaten gefunden.

6 Vorverarbeitung

Die vom polnischen Umweltamt bezogenen Feinstaubdaten müssen einer Untersuchung auf Ausreißer und fehlende Werte unterzogen werden. Hierbei wird analysiert, welche Stationen die meisten Daten zu verschiedenen Zeitpunkten liefern und wie die Ausfallzeiten verteilt sind. In Abbildung 6.1 werden alle Stationen zusammen mit dem zugehörigen Prozentsatz dargestellt, der angibt, für wie viele Stunden Daten verfügbar sind. Die Schwankungen in den Prozentsätzen resultieren aus der Tatsache, dass einige Stationen im Laufe der Jahre errichtet oder abgebaut wurden. Die Ausfallzeiten betragen in 86,4 % der Fälle maximal 5 Stunden, während die restlichen Ausfälle bis zu 119 Tage andauern.

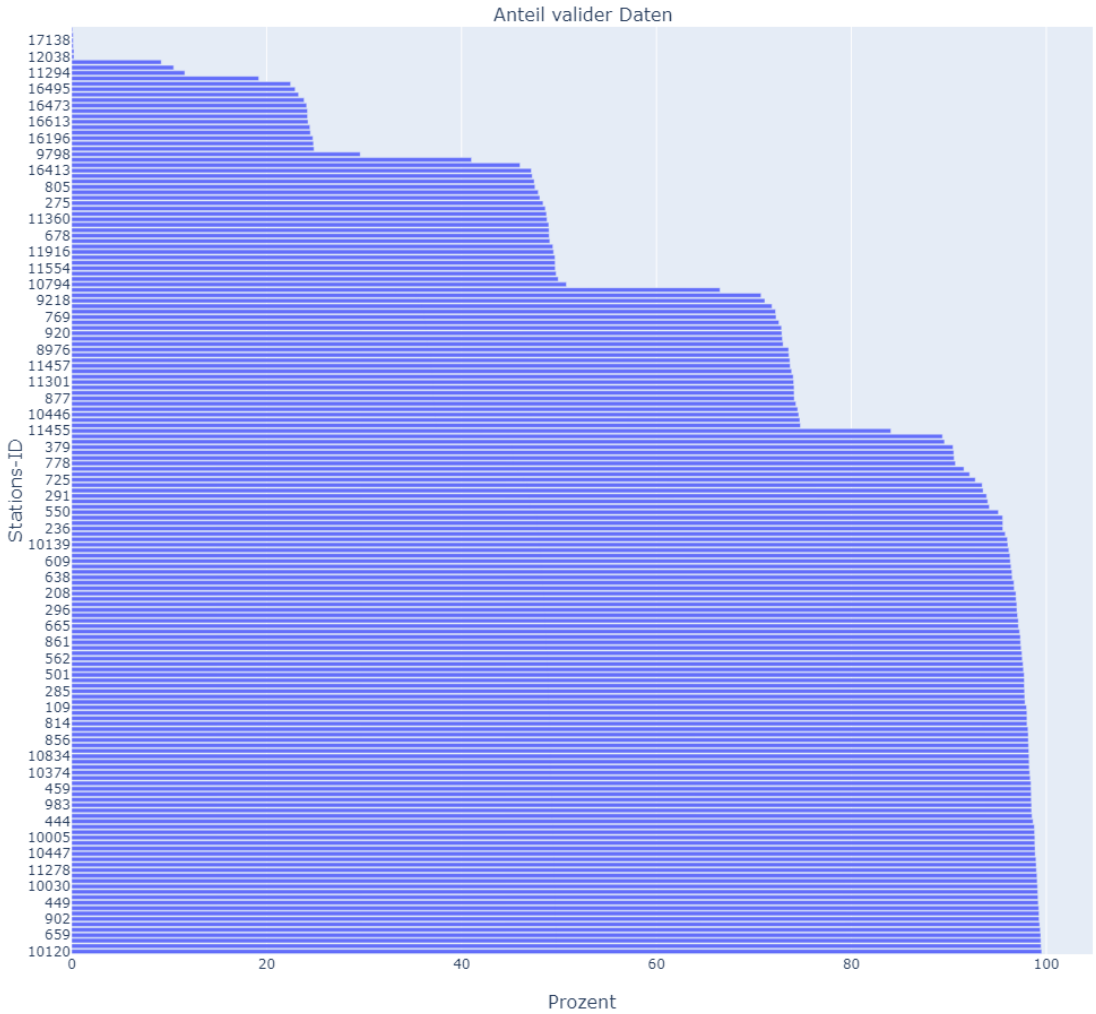


Abbildung 6.1: Für alle Stationen ist dargestellt, für wie viel Prozent der Stunden von 2019 bis 2022 Werte vorhanden sind.

6.1 Interpolation

Um ausreichend große, valide Zeiträume für das Training der Modelle zur Verfügung zu haben, werden fehlende Daten interpoliert. Da die PM_{10}-Werte bereits innerhalb weniger Stunden stark schwanken können, wird beschlossen, nur Lücken von höchstens 5 Stunden zu interpolieren. Mit dieser Begrenzung können immer noch 86,4 % der Lücken gefüllt werden. Für die Schätzung der Werte wird eine lineare Interpolation verwendet. In Abbildung 6.2 wird ein Beispielverlauf der PM_{10}-Daten der Station 814 in Kattowitz im Juli 2022 dargestellt. In dem Diagramm ist der Verlauf der Feinstaubmessungen in blau dargestellt, während der interpolierte Verlauf in rot dargestellt ist. Es ist zu sehen, dass die lineare Interpolation für die ein bis zwei Stunden langen Lücken in diesem Beispiel ausreichend ist. Der Verlauf wird durch die Interpolation nicht verfälscht, da diese lediglich vorhandene Tendenzen fortsetzt. Alle Zeitpunkte, für die nach der Interpolation keine PM_{10}-Werte vorliegen, werden aus dem Datensatz entfernt. Für das Training der ML-Modelle werden nur Daten von Stationen verwendet, bei denen für mindestens 50 % der Zeitpunkte zwischen Anfang 2019 und Ende 2022 PM_{10}-Werte vorliegen.

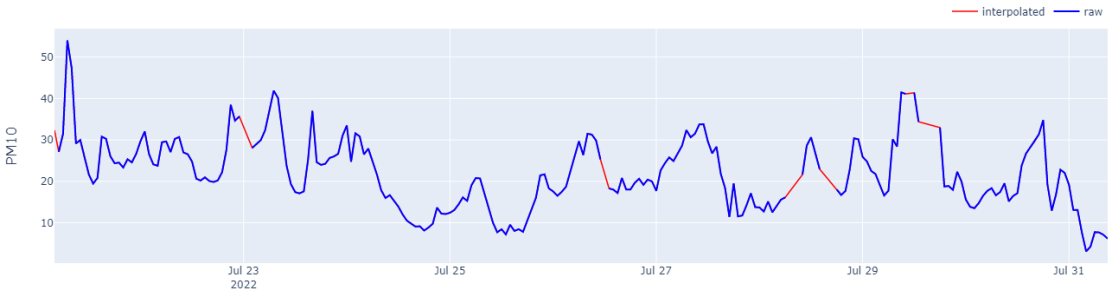


Abbildung 6.2: Beispielausschnitt aus den Daten der Station 814 in Kattowitz. In Blau ist der originale Verlauf dargestellt und in Rot der interpolierte Verlauf.

6.2 Outlier Detection

Zur Untersuchung von Ausreißern (Outliers) in den Daten werden zunächst die Maximalwerte der einzelnen Stationen analysiert. Ein Ausschnitt aus den Maximalwerten für PM_{10} ist in Abbildung 6.3 dargestellt. Dieser Ausschnitt umfasst 21 Stationen, einschließlich der Station mit dem insgesamt höchsten Wert. Die Station mit dem Index 88 weist einen Maximalwert von 1.019,7 $\mu g/m^3$. Da dieser Wert immer noch im Bereich realistisch zu erwartender Werte liegt und die umliegenden Werte ähnlich hoch sind, wird beschlossen, keine weitere Ausreißererkennung nach oben durchzuführen. Bei dieser Entscheidung wird

auch berücksichtigt, dass bereits seitens des polnischen Umweltamtes bei der Aggregation der Daten auf stündliche Werte eine Vorverarbeitung durchgeführt wird.

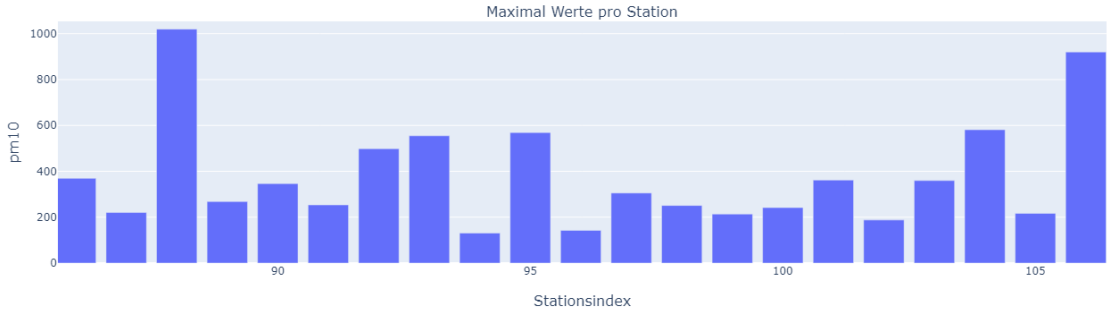


Abbildung 6.3: Ausschnitt aus den Maximalwerten für PM_{10} pro Station. Die Stationsindizes sind nicht identisch mit den Station-IDs.

Im Rahmen der Outlier Detection werden auch die Minimalwerte für PM_{10} untersucht. Ein Ausschnitt dieser Werte ist in Abbildung 6.4 dargestellt. Dabei fällt der negative Wert der Station mit dem Index 15 auf (roter Kreis). Solche negativen Werte treten nur vereinzelt an wenigen Stationen auf, wobei die benachbarten Werte alle gültig sind. Daher wird entschieden die negativen Werte zu entfernen und gemäß dem in Abschnitt 6.1 beschriebenen Verfahren zu interpolieren.
Bei einigen Stationen sind Nullwerte vorhanden. Da das Fehlen von Feinstaub in bewohnten Gebieten nicht realistisch ist, werden auch diese Werte entfernt und entsprechend interpoliert.

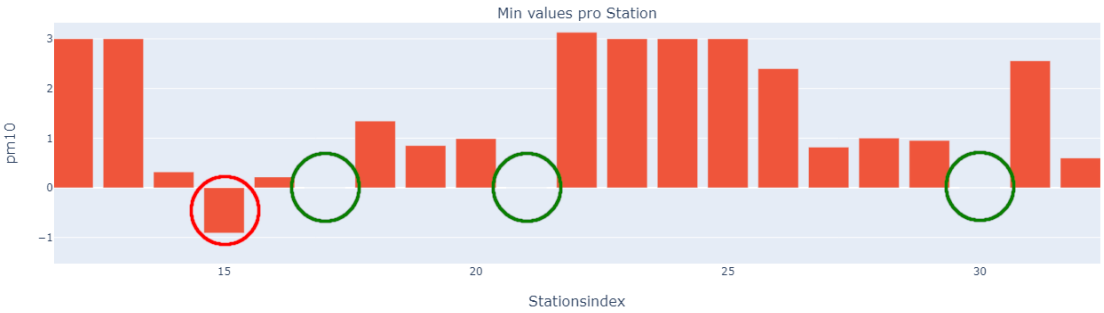


Abbildung 6.4: Ausschnitt aus den Minimalwerten für PM_{10} pro Station. Die Stationsindizes sind nicht identisch mit den Station-IDs.

7 Merkmalsextraktion

Das vorliegende Kapitel behandelt die Merkmalsextraktion, die auf den in Kapitel 6 aufbereiteten Daten durchgeführt wird. Eine Merkmalsextraktion ist erforderlich, da neuronale Netze Daten nicht in jedem Format verarbeiten können. Zum Beispiel müssen kategoriale Variablen zunächst kodiert werden, bevor sie dem neuronalen Netz zur Verfügung gestellt werden können. Die Werte für Temperatur, Luftfeuchtigkeit, Windgeschwindigkeit, Niederschlag, sowie der PM_{10}- und der $PM_{2.5}$-Wert sind metrische, kontinuierliche Variablen, die direkt vom neuronalen Netz verarbeitet werden können. In den folgenden Unterkapiteln wird das Vorgehen der Merkmalsextraktion für die verbleibenden Merkmale beschrieben.

7.1 Zeit

Der zeitliche Zusammenhang zwischen verschiedenen Datensätzen ist für die spätere Vorhersage des Feinstaubwerts relevant. Der Zeitstempel eines Datensatzes folgt dem Format der ISO8601. Beispielsweise hat der Zeitstempel für den 11.07.2023 um 12 Uhr nach ISO8601 das Format 2023-07-11 12:00:00. Dieses Format ist jedoch nicht für die Verwendung im neuronalen Netz geeignet. Durch die periodische Natur der Zeit können die Zeitstempel mithilfe einer Kombination von Sinus und Kosinus dargestellt werden.
Insbesondere der Verlauf der Messwerte über einen Tag ist von Interesse, da dieser zum Beispiel den Verlauf der Temperatur oder des Verkehrsaufkommens (Pendlerverkehr) widerspiegelt. Da auch eine jährliche Periodizität, einschließlich Sommer und Winter mit der Heizperiode, vorhanden ist, erfolgt das Feature Engineering für den Tages- und Jahresverlauf der Zeit.
Die Umrechnung des Zeitstempels in Sinus- und Kosinus-Werte für den Tag und das Jahr erfolgt gemäß den Formeln (7.1) bis (7.4). Dabei entspricht der genannte $timestamp$ den vergangenen Sekunden seit dem 1. Januar 1970 um 0 Uhr Universal Time Coordinated (UTC). Die Zahl 86400 entspricht der Anzahl der Sekunden an einem Tag. Diese lässt sich aus 24 Stunden, multipliziert mit 60 Minuten pro Stunde, multipliziert mit 60 Sekunden pro Stunde errechnen. Für den Jahres-Sinus- und Kosinus-Wert werden die Sekunden pro Tag mit 365.2425 Tagen pro Jahr multipliziert.

$$day_{sin} = sin(timestamp * (2 * \pi/86400)) \tag{7.1}$$

$$day_{cos} = cos(timestamp * (2 * \pi/86400)) \tag{7.2}$$

$$year_{cos} = sin(timestamp * (2 * \pi/(365,2425 * 86400))) \tag{7.3}$$

$$year_{cos} = cos(timestamp * (2 * \pi/(365,2425 * 86400))) \tag{7.4}$$

Abbildung 7.1 zeigt den Verlauf des Tages-Sinus- und Kosinus-Wertes für den ersten Januar des Jahres 2019. Dabei ist zu sehen, dass nur über die Kombination der beiden Kurven ein Rückschluss auf den exakten Zeitstempel möglich ist. Würde nur die Kosinuskurve verwendet werden, würde $cos(0) = 0$ den Zeitpunkt um 6 Uhr und den Zeitpunkt um 18 Uhr repräsentieren.

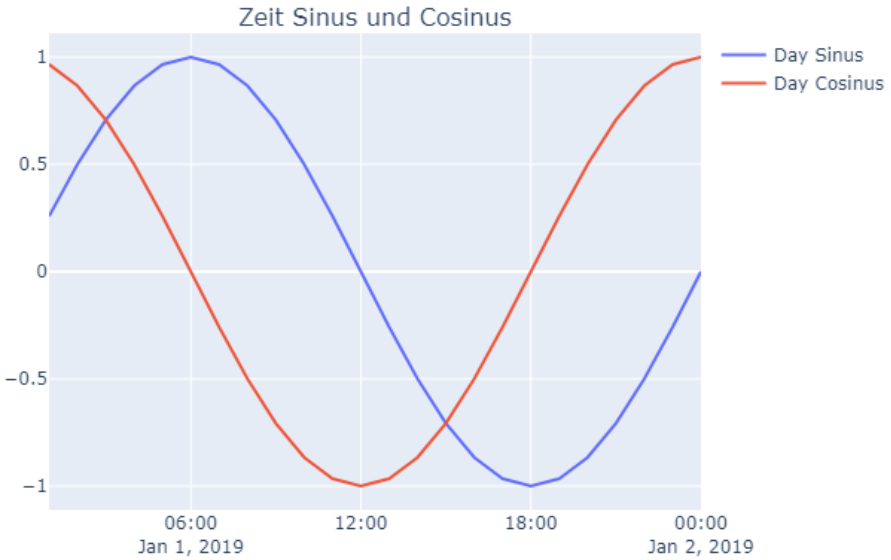


Abbildung 7.1: Feature Engineering der Zeitdarstellung mit Sinus und Kosinus.

7.2 Windrichtung

Die Wetter-API liefert auch Informationen über die Windrichtung an den jeweiligen Standorten. Die zurückgegebenen Werte liegen zwischen 0° und 360° und werden in Grad angegeben. Für neuronale Netze stellt der Übergang von 0° zu 360° ein Problem dar. Obwohl die Änderung der Windrichtung bei diesem Übergang nur minimal ist, ist die Veränderung im absoluten Wert groß. Da neuronale Netze besonders stark auf große Änderungen reagieren und dieser Sprung dem Betrag der Änderung der Windrichtung nicht entspricht, wird auch für die Windrichtung eine Kodierung durchgeführt. Im ersten Schritt erfolgt die Umrechnung der Windrichtung von Grad in Radiant. Anschließend erfolgt die Kodierung der Windrichtung mithilfe von Sinus und Kosinus. Dadurch gehen in die Sinus- und Kosinus-Darstellung die Werte von 360° und 0° nahtlos ineinander über.
Abbildung 7.2 zeigt beispielhaft das Feature Engineering für die Windrichtung im Zeitraum vom 05. Januar 2019 bis zum 16. Januar 2019. Die beiden roten Kästen markieren den Übergang

zwischen 0° und 360°. In der obersten Grafik ist der Sprung im Verlauf der Windrichtung in Radiant deutlich zu sehen. Dieser künstlich erzeugte Sprung ist in der kodierten Darstellung durch Sinus und Kosinus nicht mehr vorhanden. Das neuronale Netz kann somit besser mit dieser kodierten Darstellung der Windrichtung umgehen. Ein weiterer Vorteil besteht darin, dass der Wertebereich durch die Sinus- und Kosinus-Darstellung auf -1 bis 1 begrenzt ist. Dieser Wertebereich eignet sich deutlich besser als der ursprüngliche Bereich von 0° bis 360°.

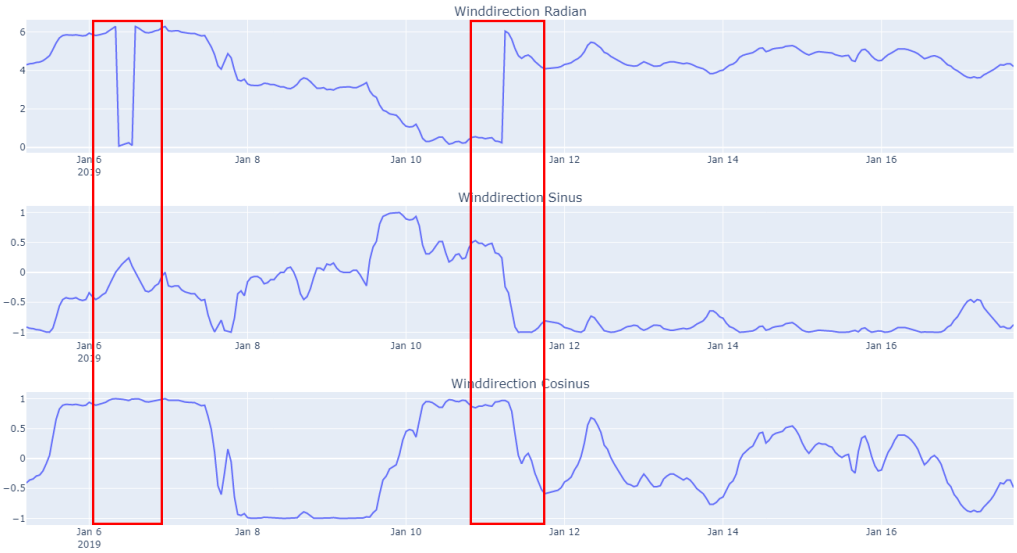


Abbildung 7.2: Feature Engineering der Windrichtung mit Sinus und Kosinus.

7.3 Merkmalsvektor

Der Merkmalsvektor setzt sich aus den Wetterdaten sowie dem Feinstaubwert zusammen. Die nachfolgende Liste zeigt den endgültigen Merkmalsvektor mit Angabe der jeweiligen Einheit des Merkmals:

- Temperatur (°C)
- Luftfeuchtigkeit (%)
- Windgeschwindigkeit (m/s)
- Niederschlag (l/m^2)
- Windrichtung (sin & cos)
- Tag (sin & cos)
- Jahr (sin & cos)
- PM_{10} (μm^3)

Alle analysierten Wetterdaten sowie der Feinstaubwert werden als Merkmale in den Featurevektor aufgenommen, da bei der Analyse keine deutlichen Korrelationen zwischen den verschiedenen Parametern erkennbar sind. Dies bedeutet, dass keine signifikanten Abhängigkeiten oder Beziehungen zwischen den einzelnen Datenpunkten vorliegen, die eine Reduzierung des Merkmalsraums gerechtfertigt hätten. Daher wird beschlossen, alle verfügbaren Informationen beizubehalten, um dem neuronalen Netz die größtmögliche Menge an Informationen zur Verfügung zu stellen.

Um eine optimale Verarbeitung der Eingaben durch das neuronale Netz zu gewährleisten, ist es erforderlich, die Merkmale des Merkmalsvektors zu normalisieren. Es ist üblich, die Daten vor der Eingabe in das neuronale Netz auf einen bestimmten Bereich oder eine bestimmte Skala zu normalisieren. Diese Normalisierung kann dazu beitragen, eine bessere Konvergenz während des Trainingsprozesses zu erreichen und die Leistung des Modells zu verbessern.
In diesem Fall wird beschlossen, auf eine vorherige Normalisierung der Daten zu verzichten. Stattdessen wird die Normalisierung mithilfe eines Normalisierungslayers innerhalb des neuronalen Netzes selbst durchgeführt. Dieser Ansatz ermöglicht es dem Modell, die Verteilung der Daten eigenständig zu erlernen und die Gewichtungen entsprechend anzupassen. Durch die Integration der Normalisierung innerhalb des Modells wird eine verbesserte Flexibilität und Anpassungsfähigkeit erreicht, da das Netz die Daten während des Trainings basierend auf dem zugrunde liegenden Muster normalisieren kann. Der Normalisierungslayer führt die Normalisierung gemäß der Gleichung (7.5) durch. Die normalisierten Werte werden um null verteilt und weisen eine Standardabweichung von eins auf.

$$norm_{value} = (value - mean)/\sqrt{var} \tag{7.5}$$

Der Mittelwert sowie die Varianz werden dann während des Trainings pro Feature durch den Normalisierungslayer erlernt.
Der endgültige Merkmalsvektor bietet eine umfassende Darstellung der Wetterbedingungen, indem er verschiedene relevante meteorologische Parameter kombiniert. Diese werden anschließend mit dem Feinstaubwert zusammengeführt. Durch die Integration aller verfügbaren Daten und die Durchführung der Normalisierung innerhalb des neuronalen Netzes wird eine optimale Grundlage für die Vorhersage der Feinstaubwerte geschaffen.

8 Maschinelles Lernen

Diese Kapitel beschreibt die Strukturierung der Daten und die verwendeten neuronalen Netze. Die Strukturierung der Daten ist erforderlich, um die zeitlichen Daten in ein für das Netz verwendbares und erlernbares Eingabeformat umzuwandeln.

8.1 Window Generator

Um den Daten eine zeitliche Struktur zu verleihen, wird ein Data Windowing-Verfahren angewendet. Die resultierenden Fenster werden als Eingabe für das neuronale Netz verwendet. Ein Data Window wird durch Inputs und Labels definiert. Im Input-Bereich sind die Merkmalsvektoren der Stunden enthalten, die als Historie dienen. Im Label-Bereich sind die stündlichen PM_{10}-Daten definiert, die prognostiziert werden sollen. Dieser Bereich dient als Referenz für das Modell, um das Training durchzuführen [10].
Die verwendete Struktur der Data Windows ist in Abbildung 8.1 dargestellt. Als Historie werden die Merkmalsvektoren eines Zeitfensters von 168 Stunden, also sieben Tagen, verwendet. Die zu prognostizierenden PM_{10}-Werte beziehen sich auf die darauf folgenden 24 Stunden, also einen Tag. Somit besteht ein Data Window aus aufeinanderfolgenden stündlichen Daten über einen Zeitraum von 192 Stunden, also acht Tagen.

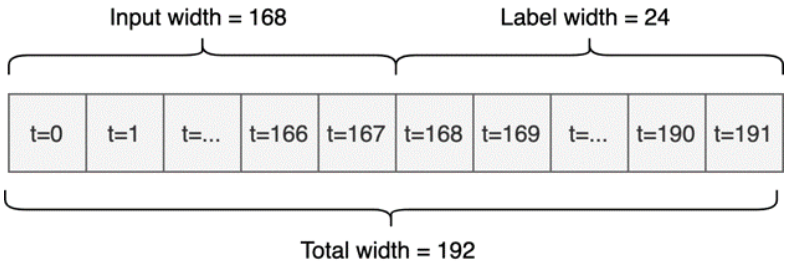


Abbildung 8.1: Aufbau eines Data Windows.

Um die Gültigkeit der Werte bzw. Windows sicherzustellen, werden nur Windows über vollständigen und aufeinander folgenden stündlichen Daten erstellt. Der Versatz der Windows wird minimal gehalten, um eine möglichst große Anzahl unterschiedlicher Windows zu erhalten. Wie in Abbildung 8.2 zu sehen, werden die Windows mit einem Versatz von einer Stunde generiert.
Zur Veranschaulichung der im Window enthaltenen Daten, ist in Abbildung 8.3 ein beispielhaftes Window dargestellt.

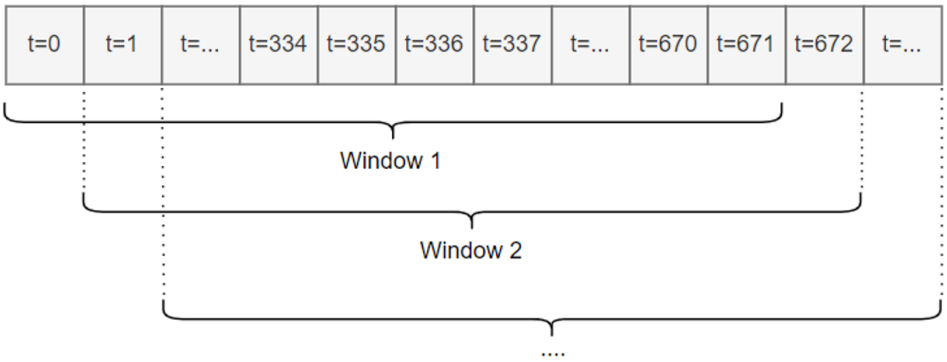


Abbildung 8.2: Versatz der Windows in den Daten.

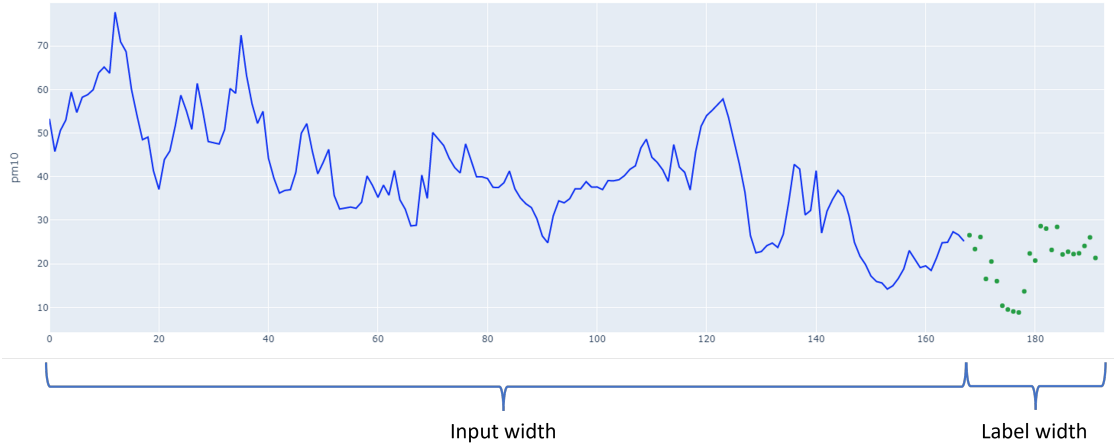


Abbildung 8.3: Beispielhafter Verlauf der PM_{10}-Werte eines Windows.

In der Grafik wird der Verlauf der PM_{10}-Werte betrachtet. Die Werte, die durch die blaue Linie abgebildet sind, repräsentieren die PM_{10}-Werte der 168 Stunden, die als Historie dienen. Die grünen Datenpunkte stellen die zu prognostizierenden PM_{10}-Werte dar. Dies sind die Label und definieren die tatsächlichen Werte, die 24 Stunden nach den Inputdaten aufgezeichnet wurden.

8.2 Neuronales Netz

Für die Prognose der Daten wird ein neuronales Netzwerk verwendet. In diesem Projekt wird eine Kombination aus CNN und LSTM als single-shot multi-step Modell verwendet.
Das neuronale Netz besteht aus einer Reihe von Schichten und Operationen zur Datenverarbeitung. Zunächst werden die Eingangsdaten normalisiert, um eine einheitliche Skalierung

sicherzustellen.
Anschließend wird eine Convolutional-Schicht mit einer Kernel-Größe von 24 verwendet. Danach folgt eine Pooling-Schicht, um die räumliche Dimension der Daten zu reduzieren. Darauf folgend wird eine BatchNormalization-Schicht verwendet, um die Daten zu normalisieren und das Training zu stabilisieren. Dieser Block, bestehend aus Convolutional-, MaxPooling- und BatchNormalization-Schicht, wird zweimal wiederholt.
Durch weitere MaxPooling1D- und BatchNormalization-Schichten wird die Dimensionalität der Daten reduziert und die Daten normalisiert.
Nach diesen Blöcken folgt eine LSTM-Schicht zur Verarbeitung von sequenziellen Daten. Dense-Schichten werden verwendet, um lineare Transformationen auf die Daten anzuwenden. Die darauffolgende Dropout-Schicht vermeidet eine Überanpassung/Overfitting des Modells an die Trainingsdaten.
Die Convolutional-Schichten besitzen insgesamt eine Kernelgrößen von 24, 12 und 6. Diese Größen wird gewählt, um mithilfe der Kernel den zeitlichen Verlauf eines Tages zu repräsentieren.
In Abbildung 8.4 ist die beschriebene Struktur des Modells abgebildet.

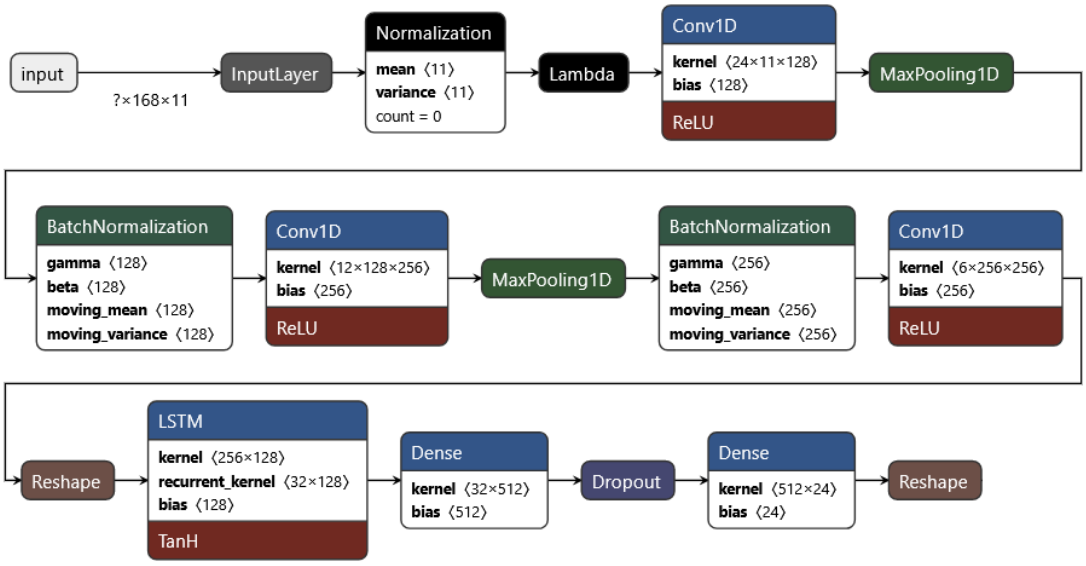


Abbildung 8.4: Grundstruktur des neuronalen Netzes.

8.2.1 Variationen

Basierend auf dem in Abschnitt 8.2 beschrieben Modell werden verschiedene Änderungen vorgenommen, um die Auswirkungen der Modellarchitektur auf die Prognosen zu analysieren.

Jede beschriebene Änderung basiert auf dem Basismodell.
Die erste Variation erhöht die Größe des LSTM von 32 auf 64 Einheiten. Durch Erhöhung der LSTM-Größe wird die Modellkapazität gesteigert und die Fähigkeit, komplexere Muster in den Daten zu erfassen, verbessert.
Als weitere Variante wird die Auswirkung von der Hinzunahme einer weiteren LSTM-Schicht betrachtet.
Die dritte Variante stellt eine Kombination der ersten beiden Änderungen dar. Es wird ein Modell mit zwei LSTM-Schichten und einer LSTM-Größe von jeweils 64 Einheiten betrachten. Hierbei wird erwartet, dass eine erhöhte Größe der LSTM-Schichten und eine Hinzunahme einer weiteren LSTM-Schicht zu einer höheren Modellkapazität und einer besseren Erfassung von langfristigen Abhängigkeiten führen.
Eine weitere Variation betrifft die Verwendung einer bidirektionalen LSTM-Schicht. Um die Leistungsfähigkeit weiter zu verbessern, wird auch die Verwendung von zwei bidirektionalen LSTM-Schichten im Modell untersucht.
Darüber hinaus wird mit der Verwendung von Average Pooling anstelle von Max Pooling experimentiert. Durch diese Änderungen wird in den Daten nicht nach dem Maximum der Daten eines Filters gesucht, sondern der Mittelwert der Daten eines Filters bestimmt.
Außerdem wird die Auswirkung der Hinzunahme einer weiteren Dense-Schicht betrachtet.
Um das Overfitting des Modells zu regulieren, wird der Wert des Dropouts im Intervall [0,2; 0,6] in 0,1er Schritten erhöht.
Abschließend werden die Auswirkung der Hinzunahme benachbarter Stationen zu dem Featurevektor betrachtet. Dadurch wird die Menge an Daten, die für eine Vorhersage verwendet werden, erhöht.
Durch die systematische Untersuchung dieser Variationen sollen Erkenntnisse über die Auswirkungen der Modellarchitektur auf die Leistung des CNN-LSTM-Modells gewonnen und die optimale Konfiguration für das beschriebene Prognoseproblem abgeleitet werden.

8.2.2 Klassifikation

Das polnische Umweltamt stellt auf seiner offiziellen Webseite [7] Prognosen für PM_{10}-Werte bereit. Diese Prognosen basieren auf einer Kategorisierungsmethode (vgl. Tabelle 8.1).
Nach diesem Vorbild wird ein neues Modell trainiert, bei dem die konkreten PM_{10}-Werte in die Kategorien aus Tabelle 8.1 umgewandelt werden. Zu diesem Zweck wurde der Merkmalsvektor um die entsprechenden Kategorien erweitert. Die Abbildung 8.5 veranschaulicht, wie die konkreten PM_{10}-Werte (blau) nach der Kategorisierung (rot) aussehen. Die Grundidee besteht darin, dass das Modell im Verlauf weniger Fluktuationen erlernen muss, was zu einer verbesserten Leistung führen könnte.

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

Tabelle 8.1: Zuordnung der PM_{10}-Werte zu den Kategorien [7].

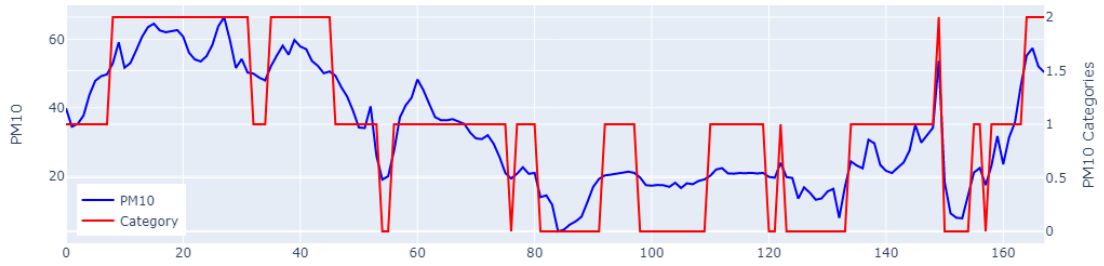


Abbildung 8.5: Beispielhafte Klassifikation der PM_{10}-Daten.

Die Anwendung einer Kategorisierungsmethode für PM_{10}-Werte ermöglicht eine einfachere Verarbeitung und Interpretation der Daten. Durch die Umwandlung der kontinuierlichen PM_{10}-Werte in diskrete Kategorien wird der Datenraum reduziert und das Modell kann sich auf das Erlernen der Zusammenhänge innerhalb jeder Kategorie konzentrieren. Dies kann dazu beitragen, die Komplexität der Lernprozesse zu verringern und die Leistung des Modells zu verbessern.
Die Verwendung der Kategorisierungsmethode erleichtert auch den Vergleich der Vorhersagen des Modells mit den PM_{10}-Vorhersagen des Umweltamtes. Die Kategorien bieten eine gemeinsame Grundlage für die Bewertung der PM_{10}-Belastung und ermöglichen einen direkten Vergleich zwischen den Vorhersagen des Modells und den offiziellen Vorhersagen.
Es ist jedoch zu beachten, dass die Umwandlung von PM_{10}-Werten in Kategorien auch gewisse Einschränkungen mit sich bringt. Die Granularität der Kategorien führt beispielsweise dazu, dass feinere Unterschiede in den PM_{10}-Werten nicht erfasst werden. Darüber führt die Kategorisierungsmethode zu einem Informationsverlust, da die kontinuierliche Natur der PM_{10}-Werte in diskrete Kategorien überführt wird.

9 Ergebnisse

Um die generelle Funktionsfähigkeit der Feinstaubdatenprädiktion zu überprüfen, wird zunächst ein Modell ausschließlich unter Verwendung von Daten einer einzigen Station trainiert. Dazu werden die historischen Feinstaub- und Wetterdaten der Jahre 2019 bis 2022 verwendet. Das Training wird auf den Daten der Station 814 durchgeführt, die sich in Kattowitz im Süden von Polen befindet. Die Station liegt am Stadtrand in einem Wohngebiet.
Für das Training wird ein Daten-Split auf den generierten Windows durchgeführt. 70 % der Windows werden als Trainingsdaten, 10 % als Validierungsdaten und 20 % als Testdaten verwendet. Die Test-Windows befinden sich zeitlich am Ende der Datenzeitspanne und werden nicht geshuffelt. Der Bereich der Trainings- und Validierungs-Windows liegt zeitlich vor den Test-Windows. Die Windows aus diesem Bereich werden erst zufällig geshuffelt und anschließend in Trainings- und Validierungs-Windows aufgeteilt. Es werden alle in Abschnitt 8.2.1 vorgestellten Variationen des Basismodells trainiert. Abbildung 9.1 zeigt das Training dieser Modelle über einen Zeitraum von 200 Epochen. Als Loss-Funktion für das Training wird der MAE verwendet.

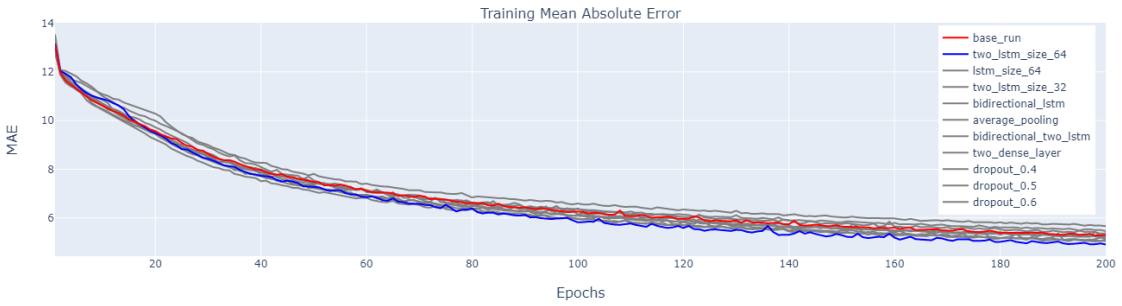


Abbildung 9.1: MAE der in Abschnitt 8.2.1 beschriebenen Modellvariationen über die Trainingsdauer von 200 Epochen. In Rot ist der Verlauf des Basismodells dargestellt. Die blaue Kurve zeigt den Verlauf des besten Modells.

Die Grafik zeigt, dass das neuronale Netz in der Lage ist, einen Zusammenhang zu erlernen. Zudem ist ersichtlich, dass alle Modelle einen ähnlichen Verlauf der Fehlerwerte aufweisen. Der Verlauf des Basismodells ist in der Grafik in Rot hervorgehoben. Da sich der Verlauf der Loss-Funktion bei allen Modellvariationen gleicht, wird für weitere Tests die Konfiguration des besten Modells verwendet. Der Verlauf dieses Modells ist in Abbildung 9.1 in Blau dargestellt. Das beste Modell besitzt die in Tabelle 9.1 dargestellten Parameter. Es zeigt sich, dass ein möglichst langes Training und die Verwendung von zwei LSTM-Schichten der Größe 64 bessere Ergebnisse liefern.

Parameter	Wert
LSTM-Layer	2
LSTM-Größe	64
Epochen	200
Dropout	0.2
Eingabe	7 Tage
Vorhersage	24 Stunden

Tabelle 9.1: Parameter des als bestes ausgewählten Modells aus dem Trainingsdurchlauf auf Daten der Station 814.

Um die Leistung des besten Modells zu evaluieren, wird es anschließend auf dem Testdatensatz bewertet. Die Ergebnisse auf allen Datensätzen sind in Tabelle 9.2 auf zwei Nachkommastellen gerundet dargestellt. Dabei zeigt sich, dass das Modell ein ziemliches Overfitting auf den Trainingsdaten aufweist, da der Fehler auf den Validierungsdaten fast doppelt so groß ist. Der Fehler auf dem Testdatensatz beträgt 10,21 und ist etwas größer als der Validierungsfehler von 8,05, bewegt sich jedoch immer noch im angestrebten Bereich.

Metrik	Wert
Train MAE	4,91
Val MAE	8,05
Test MAE	10,21

Tabelle 9.2: MAE des besten Modells auf dem Trainings-, Validierungs- und Testdatenset der Station 814.

Abbildung 9.2 veranschaulicht den Verlauf der PM_{10}-Werte eines Training-Windows, sowie die anschließende Vorhersage des neuronalen Netzes. Die blaue Linie repräsentiert dabei die PM_{10}-Werte der vergangenen sieben Tage, die dem Netz zur Verfügung gestellt werden. Die grüne Kurve stellt den tatsächlichen Verlauf der PM_{10}-Werte der nächsten 24 Stunden dar, während die orange Kurve die Vorhersage des Modells zeigt.

Anhand der Abbildung ist zu erkennen, dass das Modell in der Lage ist, den Ausschlag im Feinstaubdatenverlauf vorherzusagen. Auch die Tendenz nach dem Ausschlag in der grünen Kurve wird durch die orange Kurve abgebildet.
Allerdings hat das Modell Schwierigkeiten, die genauen Feinstaubwerte präzise vorherzusagen. Es ist zu beachten, dass die Vorhersagen für die Validierungs- und Test-Windows eine geringere Präzision als für die Trainings-Windows aufweist und die Genauigkeit der Vorhersagen von

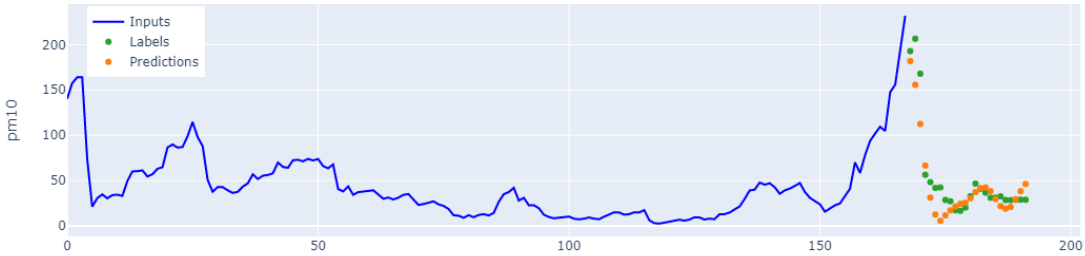


Abbildung 9.2: Beispielwindow aus dem Trainingsset der Station 814. In Blau ist der Verlauf der PM_{10}-Wert der vergangenen sieben Tage dargestellt. Die grünen Punkte zeigen den tatsächlichen Verlauf der Werte der nächsten 24 Stunden. Die orangen Punkte sind die durch das Modell prädizierten Werte.

dem spezifisch ausgewählten Window abhängt.
Es ist anzumerken, dass der MAE den durchschnittlichen Fehler über alle Windows des Datensatzes berechnet. Die Performance des Modells, wie in Tabelle 9.2 dargestellt, liegt im angestrebten Bereich für diese Station.

9.1 Benachbarte Stationen

Des Weiteren gilt es zu prüfen, ob die Entwicklung eines Modells für mehrere Stationen und damit eine Prognose für ein Gebiet möglich ist. Daher wird im nächsten Schritt untersucht, ob eine Verknüpfung von zwei benachbarten Stationen zur Generierung präziser Vorhersagen für dieses Gebiet realisierbar ist. Diese Fragestellung spiegelt sich auch in der Forschungsfrage 3.1.5 wider.
Für die Beantwortung dieser Frage werden zwei benachbarte Stationen ausgewählt. Die Stationen liegen ungefähr neun Kilometer entfernt voneinander. Die Lage der Stationen zueinander ist auf der Karte in Abbildung 9.3 dargestellt.
Um einen Vergleich durchzuführen, werden drei Modelle trainiert: ein Modell nur mit den Daten einer der beiden Stationen, ein Modell nur mit den Daten der anderen Station und ein Modell mit den Daten beider Stationen. Letzteres Modell wird im Folgenden als Kombinationsmodell bezeichnet.
Abbildung 9.4 zeigt die Vorhersagen der drei Modelle anhand eines beispielhaften Windows mit Daten der Station 530. Das Diagramm in Abbildung 9.4a stellt die Vorhersage des Modells dar, das nur mit den Daten von Station 530 trainiert wurde. Es ist zu erkennen, dass dieses Modell den Verlauf sehr gut vorhersagt, abgesehen von einem kleinen Ausschlag nach zwei Stunden. Im Vergleich dazu liefert das Kombinationsmodell eine bessere Vorhersage für die

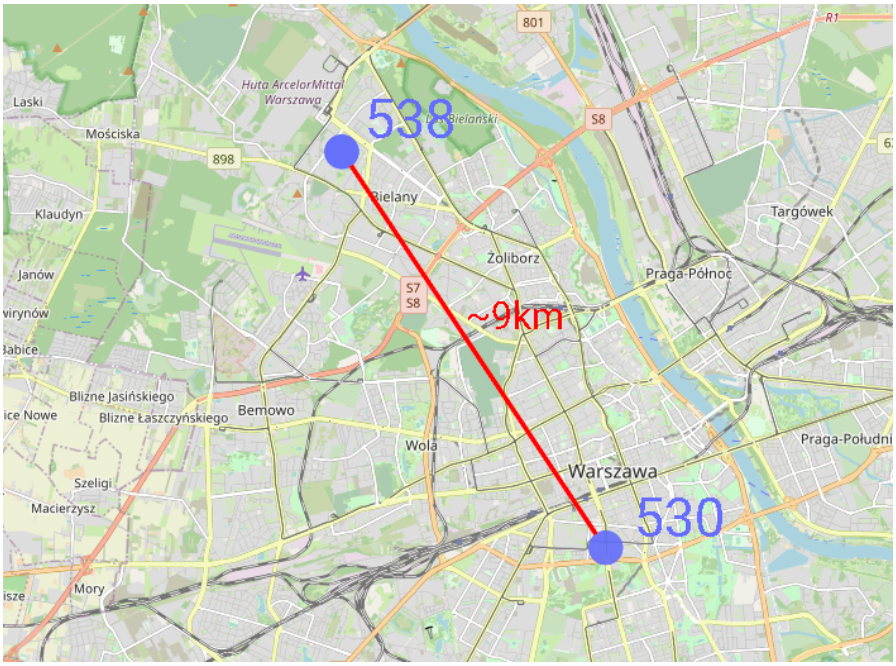
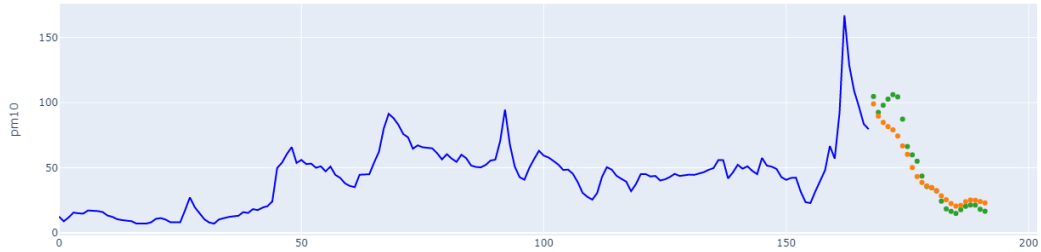


Abbildung 9.3: Karte mit der Lage der Stationen 530 und 538 im Stadtgebiet von Warschau. Die Station 530 liegt am Rand einer Straße mit Straßenbahnschienen und damit im Gebietstyp *Traffic*. Die Station 538 liegt neben einer Schule in einem Wohngebiet und damit im Gebietstyp *Background*.

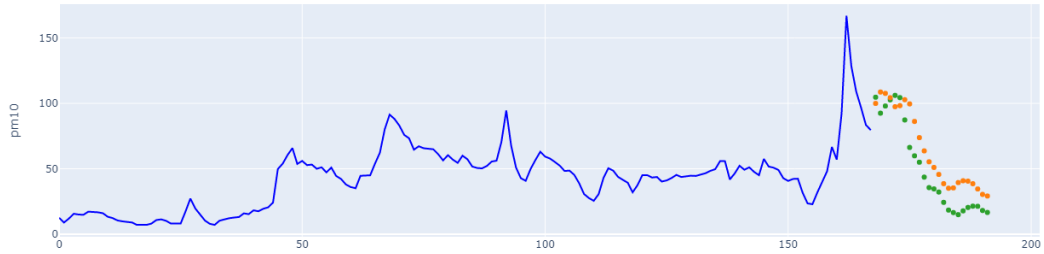
ersten sieben Stunden, wie in Abbildung 9.4b zu sehen ist. Allerdings weicht die weitere Vorhersage leicht nach oben von den tatsächlichen Werten ab. Dennoch wird der Verlauf auch mit dem erneuten leichten Anstieg in den letzten Stunden gut vorhergesagt. Im Gegensatz zu den beiden Modell, welche die Daten direkt gelernt haben, ist das Modell der Station 538 nicht in der Lage mehr als eine grobe Tendenz vorauszusagen. Dabei wird der tatsächliche Verlauf nicht getroffen.
Basierend auf diesen Ergebnissen muss festgestellt werden, dass es nicht möglich ist, mit dem Modell einer Station genaue Vorhersagen für eine andere Station zu treffen.

9.2 Klassifikation

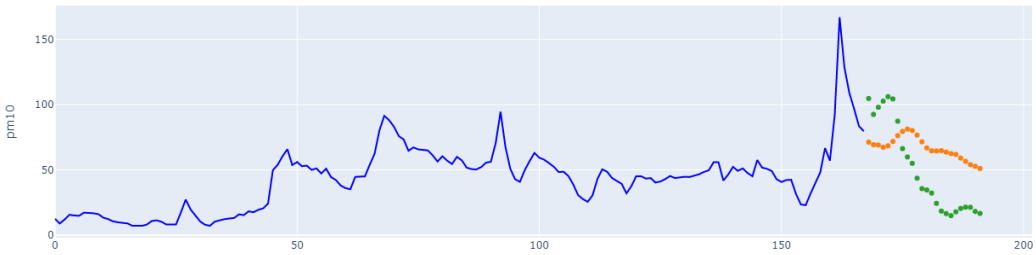
Da die präzise Vorhersage konkreter Werte nicht in allen Fällen und nur mit Ungenauigkeiten möglich ist, wird ein Klassifikationsmodell entwickelt.
Durch die Gruppierung der PM_{10}-Werte in sechs Kategorien wird dem Modell das Lernen erleichtert. Das verwendete Modell basiert auf dem besten Regressionsmodell, jedoch mit zwei Unterschieden: Die Dropout-Schicht wird von 0,2 auf 0,4 erhöht und die finale Dense-Schicht



(a) Prädiktion mit dem Modell für Station 530.



(b) Prädiktion mit dem Kombinationsmodell.



(c) Prädiktion mit dem Modell für Station 538.

Abbildung 9.4: Vergleich der Modelle für die Stationen 530 und 538, sowie das Kombinationsmodell aus beiden Stationen. Für den Vergleich wurde ein Fenster aus den Daten der Station 530 verwendet. Das Modell für Station 530 und das Kombinationsmodell haben diese Daten während des Trainings gelernt.

wird mit der Aktivierungsfunktion Softmax für die Klassifikation angepasst.
Beim Training erzielt das Modell eine Genauigkeit von 97,3 % auf den Trainingsdaten und 96,2 % auf den Validierungsdaten. Die Genauigkeit auf den Testdaten ist mit 58,9 % deutlich niedriger, was trotz des erhöhten Dropouts auf ein Overfitting schließen lässt. Ein Beispiel für eine Vorhersage auf den Testdaten ist in Abbildung 9.5 dargestellt.

In der Abbildung ist zu erkennen, dass die Vorhersage des Modells (orange) an einigen Stellen exakt mit dem tatsächlichen Wert (grün) übereinstimmt, da die Vorhersage den tatsächlichen Wert überlagert. Es ist erkennbar, dass die Tendenz der Vorhersage im Allgemeinen nicht

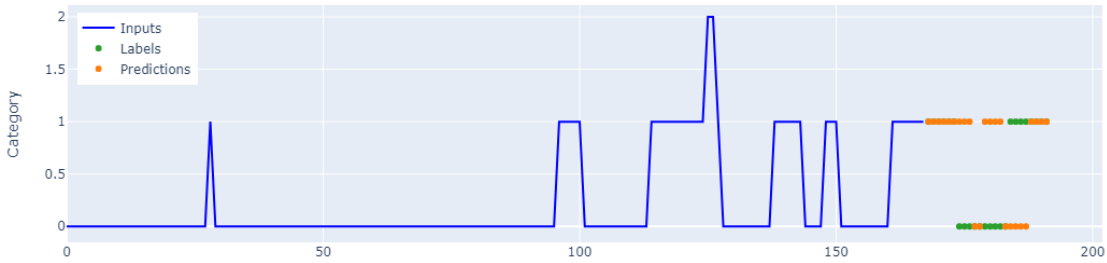


Abbildung 9.5: Beispielwindow aus dem Testset von den PM_{10} Daten der Station 814. In Blau ist der Verlauf der vergangenen 7 Tage, in Grün der nächsten 24 Stunden dargestellt. Die orange Punkte zeigen die prädizierten Werte.

grundlegend inkorrekt ist, jedoch werden Sprünge zwischen den Kategorien oft nicht richtig erfasst oder erst verzögert vorhergesagt. Daher liefert das Modell auch auf unbekannten Daten bessere Ergebnisse als die Genauigkeit vermuten lässt.

In der Forschungsfrage 3.1.4 wird die Frage aufgeworfen, wie das Klassifikationsmodell im Vergleich mit den Vorhersagen des polnischen Umweltamtes abschneidet. In Tabelle 9.3 wird der Vergleich des Klassifikationsmodells mit den Vorhersagen des polnischen Umweltamtes und den tatsächlichen Daten für den Zeitraum vom 28.06.2023, 12:00 Uhr, bis 29.06.2023, 11:00 Uhr, dargestellt. Alle Daten stammen von der Station 814 in Kattowitz. Die Vorhersagen des polnischen Umweltamtes werden aus der entsprechenden Karte abgelesen [9].

Im Vergleich zu den tatsächlichen Daten liefern sowohl die Vorhersagen des Klassifikationsmodells als auch die des polnischen Umweltamtes keine perfekten Ergebnisse. Im Zeitraum von 12:00 bis 19:00 am 28.06. prognostizieren das Klassifikationsmodell und das polnische Umweltamt beide die niedrigste Klasse. Damit liegen sie auch bis auf einen kurzen Spike knapp über die Grenze von 20 μm^3 um 15 Uhr richtig. Ab 20 Uhr am 28.06. bis zum Ende des Zeitraums um 11 Uhr am 29.06. sagt das polnische Umweltamt konstant die Klasse 1 voraus. Das Klassifikationsmodell prädiziert im Gegensatz dazu für den gesamten Zeitraum die Klasse 0.

Mit der Prädiktion der Klasse 0 liegt das Klassifikationsmodell nur für sieben Stunden zwischen 3 Uhr und 9 Uhr am 29.06. falsch. Die tatsächliche PM_{10}-Werte sind in diesem Zeitraum nur gering oberhalb der Klassengrenze von 20 μm^3. Gleichzeitig liegen die Werte weit entfernt von der nächsthöheren Klassengrenzen von 50 μm^3.

Zusammenfassend liefert das Klassifikationsmodell eine bessere Vorhersage als das polnische Umweltamt, beide Vorhersagen weisen allerdings noch Potenzial für Verbesserungen auf.

Uhrzeit	Modell	Umweltamt	Tatsächliche Daten	
			Kategorie	PM10-Messwert
28.06. 12:00	0	0	0	9,33
28.06. 13:00	0	0	0	9,25
28.06. 14:00	0	0	0	14,98
28.06. 15:00	0	0	1	20,57
28.06. 16:00	0	0	0	10,44
28.06. 17:00	0	0	0	9,76
28.06. 18:00	0	0	0	7,66
28.06. 19:00	0	0	0	8,73
28.06. 20:00	0	1	0	7,71
28.06. 21:00	0	1	0	9,07
28.06. 22:00	0	1	0	9,28
28.06. 23:00	0	1	0	10,49
29.06. 00:00	0	1	0	12,91
29.06. 01:00	0	1	0	12,95
29.06. 02:00	0	1	0	18,77
29.06. 03:00	0	1	1	22,48
29.06. 04:00	0	1	1	22,04
29.06. 05:00	0	1	1	22,65
29.06. 06:00	0	1	1	20,77
29.06. 07:00	0	1	1	22,37
29.06. 08:00	0	1	1	27,06
29.06. 09:00	0	1	1	24,95
29.06. 10:00	0	1	0	18,42
29.06. 11:00	0	1	0	16,73

Tabelle 9.3: Vergleich der vom Klassifikationsmodell vorhergesagten Werte mit den vorhergesagten Werten des polnischen Umweltamts sowie den tatsächlich gemessenen Werten. Verwendet werden Werte der Station 814 in Kattowitz im Zeitraum von 28.06.2023 12:00 bis 29.06.2023 11:00. Die Vorhersagen des polnischen Umweltamtes werden am 28.06.2023 um 12:00 aus der Vorhersagenkarte abgelesen [9].

9.3 Forschungsfragen

Die primäre Forschungsfrage 3.1 kann positiv beantwortet werden, wenn der Vorhersagezeitraum auf einen Tag verkürzt wird. Das bedeutet, dass mit der Kombination aus CNN und LSTM des neuronalen Netzes eine stündliche Vorhersage der Feinstaubdaten möglich ist. Eine Vorhersage von 14 Tagen ist jedoch nicht umsetzbar, da der Vorhersagezeitraum zu groß ist. Die Ergebnisse des besten Regressionsmodells haben gezeigt, das eine Vorhersage der PM_{10}-Wert für einen Tag mit einem MAE von 10 möglich ist (Forschungsfrage 3.1.1).
Die Modelle wurden hauptsächlich auf PM_{10}-Feinstaubdaten trainiert. Aufgrund der starken Korrelation ($> 0{,}9$) zwischen PM_{10} und $PM_{2.5}$, die in Kapitel 5 untersucht und in Abbildung 5.2 dargestellt worden sind, kann davon ausgegangen werden, dass eine Vorhersage von $PM_{2.5}$ mit dem Modell für PM_{10} möglich ist (Forschungsfrage 3.1.3). Es wurde darauf verzichtet, die $PM_{2.5}$-Vorhersage aufgrund der starken Korrelation mit PM_{10} und des zeitlich Rahmen des Projektes näher zu untersuchen.
Die Forschungsfrage 3.1.2 kann deshalb nicht beantwortet werden, da kein Modell für die $PM_{2.5}$-Vorhersage trainiert wurde. Aufgrund der starken Korrelation zwischen den beiden Werten, kann jedoch auch hier angenommen werden, dass die Vorhersage mit einem MAE unter 10 möglich wäre.
In Abschnitt 9.2 wurde ein Klassifikationsmodell trainiert, um die Leistung des Modells mit der Vorhersage des polnischen Umweltamts zu vergleichen. Für den Vergleich wurde ein aktueller Tag ausgewählt und die tatsächlichen PM_{10}-Werte wurden mit den Vorhersagen des Klassifikationsmodells und den Vorhersagen des Umweltamtes verglichen. In Bezug auf die Forschungsfrage 3.1.4 zeigt der Vergleich, dass beide Modelle keine optimalen Ergebnisse liefern, jedoch erzielt das Klassifikationsmodell leicht bessere Vorhersagen.
Des Weiteren wurde in Abschnitt 9.1 untersucht, ob es sinnvoll ist, Stationen zu Gebieten zusammenzufassen. Die Forschungsfrage 3.1.5 kann beantwortet werden, indem festgestellt wurde, dass eine Vorhersage von Daten einer benachbarten Station nicht sinnvoll ist. Dies liegt hauptsächlich daran, dass die Stationen in der Regel mehrere Kilometer voneinander entfernt sind.
Zudem weisen die Stationen unterschiedliche Eigenschaften auf. Zum Beispiel befinden sie sich entweder an verkehrsreichen Straßen (Traffic) oder in begrünten Gebieten (Background). Dennoch liefert ein Modell, das mit Daten von beiden benachbarten Stationen trainiert wurde, adäquate Ergebnisse, die jedoch schlechter sind als die Ergebnisse der stationsspezifischen Modelle.

10 Fazit und Ausblick

Dieses Projekt befasst sich mit der Analyse von Feinstaubdaten und der Entwicklung von ML-Vorhersagemodellen für Feinstaubwerte. Feinstaub gilt als schädlich für die Gesundheit, weshalb die Prognose zukünftiger Feinstaubwerte von hoher Relevanz ist. In einem ersten Schritt wurden die Daten analytisch-visuell betrachtet. Die Untersuchung beinhaltet eine Korrelationsanalyse zwischen verschiedenen Wetterdaten und Feinstaubwerten, um mögliche Zusammenhänge zu identifizieren. Allerdings zeigte sich, dass keine ausschlaggebende Korrelation zwischen den Feinstaubwerten und den betrachteten Wettervariablen besteht.

Im weiteren Verlauf wurden die Feinstaubdaten gefiltert, vorbereitet und interpoliert, um eine geeignete Datenbasis für die Modellentwicklung zu schaffen. Der Ansatz des Data-Windowing wurde verwendet, um Zeitreihendaten in Trainingsdaten umzuwandeln. Anschließend wurde ein neuronales Netz mit den vorverarbeiteten Daten trainiert. Die Architektur des verwendete neuronale Netz besteht aus einer Kombination aus CNN und LSTM. Für die Extraktion von Merkmalen wurde die CNN Struktur eingebaut. Der LSTM Block eignet sich besonders gut für Zeitreihendaten und wurde daher dem Netz hinzugefügt. Zunächst wurde ein Regressionsmodell pro Station trainiert, welches kontinuierliche Werte vorhersagt.
Um die PM_{10}-Werte in Kategorien zu gliedern und beruhend auf diesen eine Prognose durchzuführen, wurde ein Klassifikationsmodell entwickelt, das die Werte auf Basis der Kategorien des Umweltamtes klassifiziert. Durch diese Variation des Modells konnte eine Verbesserung der Prognosen erzielt werden.
Die Vorhersage für mehrere benachbarte Stationen mit einem neuronalen Netz resultierte in einer Verschlechterung der Ergebnisse. Eine mögliche Ursache hierfür könnte in der räumlichen Disparität der Stationen liegen, da diese mehrere Kilometer voneinander entfernt sind und sich oft in unterschiedlichen Umgebungen befinden (z.B. Hauptstraße, Sportplatz, Park).
Allgemein wurde eine deutliche Diskrepanz zwischen den Ergebnissen des Validierungs- und Testdatensatzes festgestellt, was auf Overfitting hindeutet. Die Ursache könnte durch die Zeiträume der Trainings- und Testdaten verursacht worden seien. Da die Trainingsdaten während der Corona-Pandemie und die Testdaten in dem Jahr nach Ende dieser erfasst wurden, könnte die Differenz in der Qualität der Vorhersage durch die Auswirkungen der Corona-Pandemie bedingt sein.
Zusammenfassend zeigt sich, dass das beste Klassifikationsmodell eine gute Leistung erzielt und im direkten Vergleich mit dem Modell des polnischen Umweltamtes eine geringfügig bessere Prognose aufweist. Das Klassifikationsmodell liefer in 67 % der Fälle das richtige Ergebnisse, das Modell des Umweltamtes in 58% der Fälle. Das ist eine Verbesserung um 9

Prozentpunkte. Außerdem liegt das Klassifikationsmodell nur um maximal 7,06 μm^3 falsch, während das Modell des Umweltamtes sich um bis zu 12,29 μm^3 irrt.

Basierend auf den vorliegenden Ergebnissen ergeben sich interessante Perspektiven für zukünftige Forschungen im Bereich der Feinstaubanalyse. Die Einteilung von Feinstaubdaten in sinnvolle Kategorien hat sich als vielversprechend erwiesen und könnte weiterhin untersucht werden. Es bietet sich an, zusätzliche Experimente mit Klassifikationsmodellen durchzuführen, um die Genauigkeit und Zuverlässigkeit der Kategorisierung weiter zu verbessern.
Ein weiterer Aspekt ist die Vorhersage von $PM_{2.5}$-Werten mithilfe neuronaler Netze, die in diesem Projekt aufgrund des zeitlichen Rahmens nicht durchgeführt wurde. Durch die hohe Korrelation zwischen den PM_{10}- und $PM_{2.5}$-Werten ist davon auszugehen, dass die Ergebnisse einer $PM_{2.5}$-Prognose eine ähnliche Genauigkeit aufweisen wie die PM_{10}-Prognosen.
Außerdem könnte der Zusammenhang zwischen der Feinstaubbelastung und der Corona-Pandemie untersucht werden. Hierbei könnten die vorhandenen Feinstaubdaten mit den Daten zu COVID-19 in Verbindung gebracht werden, um mögliche Korrelationen oder Einflüsse zu identifizieren. Dies könnte Aufschluss darüber geben, inwiefern sich die Pandemie in der Feinstaubbelastung widerspiegelt und sich die Differenz der Trainings- und Testdaten durch die Pandemie erklären lässt.
Eine weitere Fragestellung betrifft die Möglichkeit der Erstellung umgebungsspezifischer Modelle. Diese Modelle könnten mehrere Stationen mit ähnlichen Umgebungen, wie beispielsweise Straßen, Wohngebiete oder Parks, in einem einzigen Modell zusammenfassen. Durch diese Vorgehensweise könnten allgemeingültige neuronale Netze entstehen, die jeweils auf Daten einer bestimmten Umgebungen anwendbar sind.
Eine Optimierung der Vorhersagen könnte durch die Einbeziehung der Wettervorhersage für den entsprechenden Vorhersagezeitraum erreicht werden. Hierfür ist eine Erweiterung des Merkmalsvektors um die Wettervorhersagevariablen notwendig.
Schließlich ist eine Erhöhung der Anzahl von Sensoren erforderlich, um genauere Vorhersagen zu ermöglichen. Eine engere Platzierung benachbarter Stationen und die Aufstellung weiterer Sensoren würden eine höhere räumliche Auflösung und eine größere Datenmenge ermöglichen, was zur weiteren Verbesserung der Modelle beitragen würde.

Insgesamt hat das Projekt gezeigt, dass sich mithilfe von neuronalen Netzen Feinstaubdaten vorhersagen lassen. Es resultiert, dass eine Klassifikation der Feinstaubwerte in Kategorien zu bevorzugen ist, da Feinstaubwerte starken Schwankungen unterliegen und örtlich stark begrenzt ist. Eine Vorhersage des exakten Wertes übersteigt die Komplexität des vorgestellten Modells. Um ein Modell für ein Gebiet erstellen zu können, ist es zunächst notwendig die

Anzahl der Sensoren zu erhöhen, um sowohl die verfügbaren Daten als auch die Gebietsabdeckung zu erhöhen. Des Weiteren wäre es sinnvoll weitere Merkmale, die mit Feinstaub in Verbindung stehen, hinzuzuziehen. Hierbei könnten Faktoren wie das Verkehrsaufkommen oder Wettervorhersagen von Relevanz sein. Mithilfe dieser Ansätze wäre es möglich, die Vorhersagegenauigkeit der Modelle noch zu verbessern.

Literaturverzeichnis

[1] European Environment Agency. *Vorzeitige Todesfälle aufgrund von Luftverschmutzung in der EU weiter rückläufig – mehr Anstrengungen für eine schadstofffreie Umwelt nötig*. 2023. URL: https://www.eea.europa.eu/de/highlights/vorzeitige-todesfaelle-aufgrund-von-luftverschmutzung#:~:text=Die%20schlechte%20Luftqualit%C3%A4t%2C%20vor%20allem,%C2%B5g%2Fm3%20ausgesetzt%20waren. (besucht am 11. 07. 2023).

[2] Abdellatif Bekkar u. a. „Air-pollution prediction in smart city, deep learning approach“. In: *Journal of Big Data* 8 (Dez. 2021). DOI: 10.1186/s40537-021-00548-1.

[3] Umwelt Bundesamt. *Emission von Feinstaub der Partikelgröße PM 10*. 2023. URL: https://www.umweltbundesamt.de/daten/luft/luftschadstoff-emissionen-in-deutschland/emission-von-feinstaub-der-partikelgroesse-pm10#emissionsentwicklung (besucht am 11. 07. 2023).

[4] Umwelt Bundesamt. *Warum is Feisntaub schädliche für den Menschen?* 2021. URL: https://www.umweltbundesamt.de/service/uba-fragen/warum-ist-feinstaub-schaedlich-fuer-den-menschen (besucht am 11. 07. 2023).

[5] Landesamt für Natur Umwelt und Verbraucherschutz Nordrhein-Westfalen. *Bericht über die Luftqualität im Jahr 2021*. 2022. URL: https://www.lanuv.nrw.de/fileadmin/lanuv/luft/immissionen/ber_trend/Bericht_ueber_die_Luftqualitaet_im_Jahr_2021.pdf (besucht am 11. 07. 2023).

[6] Open-Meteo. *Open-Meteo - Open-source weather API*. 2023. URL: https://open-meteo.com/ (besucht am 11. 07. 2023).

[7] Główny Inspektorat Ochrony Środowiska. 2023. URL: https://www.gios.gov.pl/pl/ (besucht am 12. 07. 2023).

[8] Główny Inspektorat Ochrony Środowiska. *Api-Schnittstelle*. 2023. URL: https://powietrze.gios.gov.pl/pjp/content/api (besucht am 12. 07. 2023).

[9] Główny Inspektorat Ochrony Środowiska. *Forecast maps*. 2023. URL: https://powietrze.gios.gov.pl/pjp/airPollution?lang=en (besucht am 12. 07. 2023).

[10] TensorFlow. *Time series forecasting*. 27. Mai 2023. URL: https://www.tensorflow.org/tutorials/structured_data/time_series (besucht am 11. 07. 2023).

[11] Bundesministerium für Umwelt Naturschutz nukleare Sicherheit Verbraucherschutz. *Feinstaub*. 2022. URL: https://www.bmuv.de/themen/gesundheit-chemikalien/gesundheit/luftreinhaltung/feinstaub (besucht am 11. 07. 2023).

[12] WHO. *Ambient (outdoor) air pollution*. 2021. URL: https://www.who.int/news-room/fact-sheets/detail/ambient-(outdoor)-air-quality-and-health (besucht am 11. 07. 2023).