Literatur-Seminar-Arbeit

Vorhersagen von Verkehrsunfällen mithilfe künstlicher neuronaler Netze

Erik Rohr

Fachbereich Informatik (02) - Hochschule Bonn-Rhein-Sieg

3. Juli 2024

Betreuerin: Doerthe Vieten

Abstrakt

≪ Kurz beschreiben ≫

Inhaltsverzeichnis

1 Einleitung

Jedes Jahr sterben weltweit ca. 1,19 Millionen Menschen in Verkehrsunfällen (RTA, engl.: „road traffic accident“) [1]. Eine umfassende Analyse vorliegender Verkehrsdaten im Hinblick auf potentielle Bedrohungen kann dazu beitragen, Baumaßnahmen unfallminimierend zu gestalten.

Vorangehende Analysen von Verkehrsdaten, die eine Vielzahl an „Machine Learning“ -Modelle nutzten, haben ergeben, dass eine Vielzahl an Faktoren existieren, die das Risiko auf RTA erhöhen, wie Wetterbedingungen, Straßenkonditionen, Zustand des Fahrers, Lichtverhältnisse, Tageszeit und die Verkehrsdichte [**ml**, **predict**]. „Machine Learning“-Modelle (ML) sind darauf ausgelegt, nicht anhand von spezifischen Anweisungen, sondern allein durch das Erkennen von Mustern und Abhängigkeiten mithilfe komplexer Funktionen [**predict**] in Daten Erkenntnisse zu ziehen und Entscheidungen sowie Vorhersagen treffen zu können [2].

Unter den ML-Modellen werden künstliche neuronale Netze (KNN) für diesen Anwendungsfall bevorzugt, da diese keine zugrundeliegenden Beziehungen zwischen den Eingangsvariablen benötigen und darauf ausgelegt sind, mit historischen Daten Aussagen über die Zukunft treffen zu können [**predict**].

In dieser Arbeit wird eine systematische Literaturrecherche (SLR) zur Vorhersage von RTAs mithilfe KNNs dargestellt. Die Ergebnisse werden im Hinblick auf deren Mehrwert in der Minimierung von Straßenverkehrsunfällen durch unsichere Planung der Architektur untersucht und eingeordnet.

1.1 Theorie

1.1.1 Künstliche neuronale Netze (KNN)

Künstliche neuronale Netze (KNN) sind ein Teilbereich der ML-Modelle. Sie können Entscheidungen auf ähnlicher Art und Weise treffen, wie das menschliche Gehirn [3]. Das Modell besteht aus einer Vielzahl von Schichten mit unterschiedlich vielen Knoten. Diese Knoten sind mit anderen Knoten aus den benachbarten Schichten verbunden und besitzen sogenannte „weights“ und „bias“ [3]. Eingabedaten werden von der „input layer“ (Deutsch: „Eingangsschicht“) entlang aller hidden layer (Deutsch: „verborgene Schicht“) durchgereicht, verarbeitet und schließlich in der „output layer“ (Deutsch: „Ausgangsschicht“) aggregiert ausgegeben [3].

Zur Konfigurierung von KNN können unter anderem die Anzahl der Schichten, die Optimierungsverfahren (engl.: „optimizers“), die Aktivierungsfunktion (engl.: „activation

function“) (AF) und die Verlustsfunktion (engl.: „loss function“) frei gewählt werden [4].

Ein Optimierungsverfahren ist eine Funktion oder ein Algorithmus, welches Parameter der Knoten wie die Gewichte und die Lernrate (engl.: „learning rate“) anpasst, um den Verlust zu minimieren, die Genauigkeit zu maximieren und die benötigte Trainingszeit exponentiell zu reduzieren [4]. Beispiele für solche Optimierungsverfahren sind der „gradient descent“, „RMS prop (Root Mean Square)“ und „adam's optimizer“ [4].

Die AF entscheidet, wie der kalkulierte Ausgangswert eines Knotens zu interpretieren ist. Dies fügt dem Netz eine Nicht-Linearität hinzu [4]. Somit können die verborgenen Schichten jeweils für andere Bereiche der Interpretation der Daten stehen. Linearität würde auf der anderen Seite bedeuten, dass das Netz in eine Funktion zusammengefasst werden könnte [4]. Beispiele sind die binäre „step function“, die Sigmoid-Funktion und die „RELU“-Funktion [4].

Eine Verlustsfunktion (VF) misst, wie genau das KNN den Datensatz modelliert, indem es den vorliegenden Verlust (engl.: „loss“) berechnet [4]. Gängige Funktionen sind die „Mean Squared Error“-Funktion und die „Binary Cross Entropy“-Funktion [4].

1.2 Aktuelle Forschungslage

1.2.1 Banerjee et al.

Eine Vielzahl an Klassifizierern (engl.: „classifiers“) sind mit KNN-Modellen bezüglich Vorhersage der Mortalität in RTAs verglichen worden. Darunter gehören Random Forest, Support Vector Machine, K-Nearest Neighbor Classifier, AdaBoost Classifier, XGBoost Classifier. Hierbei hat das KNN mit 7 „hidden layer“, 1 „input layer“, Adams Optimizer und einer „dropout class“ in der Eingangsschicht die beste Genauigkeit von 84,36% erzielt [5].

1.2.2 Maurya et al.

Vorhersehen von Kraftfahrzeuggeschwindigkeiten dienen als Basis für ein fortgeschrittenes Verkehrsmanagementsystem [6]. Hierbei wurden Lineare Regression, „random forest“-Regression, „Decision Tree“ und ein KNN miteinander Vergleichen uns eine „Performance Analyse“ durchgeführt. KNN und die „random forest“-Regression haben mit einem „R squared“-Wert [1] von jeweils 0,9301 und 0,9642 das beste Fitting erzielt [6].

[1] Bestimmtheitsmaß. Stellt die Anpassungsgüte einer Regression dar [7]

1.2.3 Zohra et al.

Anhand den Datensätzen aus „US Accidents (2016-2023)“ wurden ein KNN, ein Random Forest Klassifizierer und eine logistische Regression trainiert und getestet. Hier wurden jeweils Genauigkeitswerte von 81,1%, 90,7% und 87,03% erreicht [8].

2 Methodik

Im Folgenden wird die Vorgehensweise bei der Literaturrecherche beschrieben. Die Vorgehensweise orientert sich an der PRISMA-Leitlinie (Preferred Reporting Items for Systematic Reviews and Meta-Analyses). PRISMA hat sich ursprünglich als effektiver Vorreiter bzgl. SLR im medizinischen Sektor etabliert, mit dem Fachkräfte auf dem aktuellen Stand der Wissenschaft bleiben und bestehende Vorschriften aktualisiert werden [9]. Entnehmend der zahlreichen SLR-Veröffentlichungen in Bereichen der Informatik (siehe Datenbanken ACM/IEEE), hat sich PRISMA als hilfreiches Mittel zur SLR auch in anderen Wissenschaftsbereichen bewiesen und wird daher als Instrument der SLR in dieser Arbeit verwendet.

2.1 Die PRISMA-Leitlinie

PRISMA beinhaltet eine aus 27 Stichpunkten bestehende „Checkliste“ und ein 4-phasiges Fluss-Diagramm [9]. Da diese allerdings ursprünglich für die Medizin entwickelt wurden, werden jene auf die Informatik angepasst als Basis der SLR verwendet. Mithilfe dieser Elemente kann sowohl Literatur systematisch für die Aufnahme in einer SLR auf Eignung geprüft als auch der Prozess erleichtert und standardisiert werden.

Zunächst werden Forschungsfragen aufgestellt, die im Laufe der Recherche beantwortet werden sollen. Im Anschluss wird dann die Literatur anhand der ausgewählten Such-Strategie ausgewählt. Die gesammelte Literatur wird dann durch Ein- und Ausschlusskriterien gefiltert und auf Eignung für die SLR durch die Checkliste geprüft. Schließlich werden aus der eingeschlossenen Literatur qualitative und quantitative Daten extrahiert, mit anderer Literatur verglichen und in der eigentlichen Literaturanalyse zusammengetragen.

2.2 Forschungsfragen

Um die Literaturrecherche zu systematisieren, werden laut PRISMA [9] Forschungsfragen aufgestellt, damit ein systematisches Full-Text-Screening für die Eignung möglich ist. In Rahmen dieser Arbeit wurden folgende Fragen aufgestellt:

1. Welche Netztopologie sind verwendet worden?
2. Welche Genauigkeit konnte erzielt werden?
3. Wie viele Knoten sind in den verborgenen Schichten verwendet worden?
4. Wie lauten die Parameter des KNN („optimizer“, „loss function“ usw.) ?
5. Welche Eingabevariablen sind gewählt worden?

2.3 In- und Exklusionskriterien

Um gezielt die Potenz künstlicher neuronaler Netze für die Verkehrsanalyse zu untersuchen, sind spezifische Ein- und Ausschlusskriterien gewählt worden. Somit wird die Anzahl der aus der Suche ergebenen Artikel reduziert und verschafft einen besseren Überblick.
Die für diese Arbeit relevanten Artikel sind auf solche beschränkt worden, die

1. eine Verkehrsdatenanalyse untersuchen und die Daten mit einer KNN auswerten,
2. ggf. einen quantitativen Vergleich verschiedener ML-Modelle durchführen oder
3. quantitative Merkmale der KNN aufzählen.

Zu letzterem Punkt gehören Merkmale wie gewählte Netztopologie, Anzahl der Schichten, Anzahl der Knoten pro Schicht, gewähltes Optimierungsverfahren, die gewählte AF Genauigkeit und Eingangsvariablen.

2.4 Prozess der Recherche

Für die Literaturrecherche sind die Datenbanken ACM Digital Library und die IEEE Explore verwendet worden, da diese überwiegend Artikel in der Informatik veröffentlicht haben. Hierbei wurden die folgenden Suchstrings verwendet:

ACM Digital Library

[All: ”neural network?”] AND [[All: ”traffic flow”] OR [All: ”traffic control”] OR [All: ”accident”]] AND [E-Publication Date: (01/01/2023 TO 12/31/2024)]

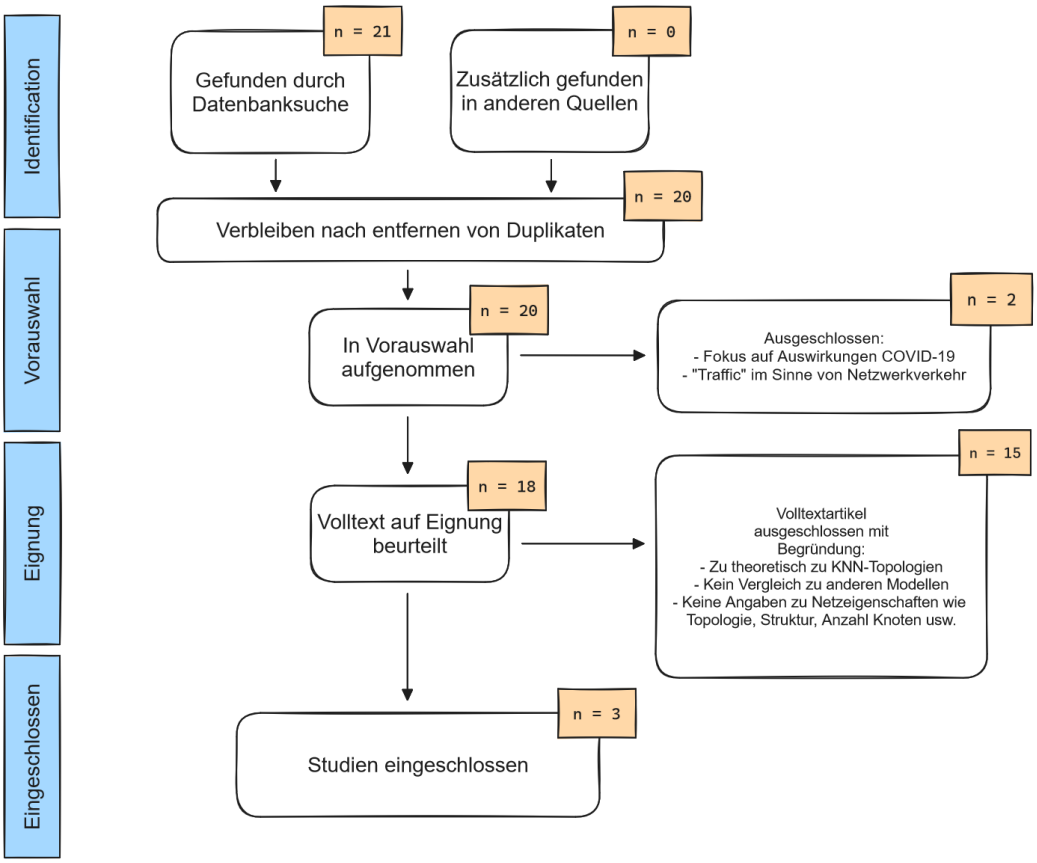
IEEE Explore

(”All Metadata”: ”artificial neural network?”) AND (”All Metadata:”traffic”) AND (”All Metadata:”control” OR ”All Metadata:”accident”)

Bei der „IEEE-Explore“-Datenbank wurde als Zeitraum-Filter der 01.01.2023 bis zum 31.12.2024 gewählt, da dieser nicht in den Such-String mit eingebunden werden konnte. Zusätzlich ist die Suche auf den Inhaltstyp „Research Article“ und die Verfügbarkeit „Open Access“ begrenzt worden, soweit von den Einstellungen beider Datenbanken möglich.

Die zusammengetragenen Artikel wurden im Anschluss mithilfe des PRISMA Fluss-Diagramms [9] auf Eignung überprüft (Abb. **??**) und in einer Tabelle zusammengefasst (Tab. **??**).

Abbildung 1: SLR nach PRISMA



3 Ergebnisse

Die Ergebnisse der Studien fallen sehr unterschiedlich aus. Sowohl Qian et al. als auch Das et al. befürworten die Nutzung von KNN in der Durchführung und Auswertung von

Verkehrsanalysen ([10] S. 2744, [11] S. 5), allerdings empfielt Bao et al., dass zukünftige Studien sich intensiver mit den einzelnen Faktoren der Verkehrsanalyse befassen sollten, da Verkehrsanalysen zu komplex sind, diese in einem Modell zusammenzufassen ([12] S. 784).

Diese Rückschlüsse werden auch von den erreichten Genauigkeiten der KNN-Modelle reflektiert: Absteigend sortiert erreichte Das et al. eine Genauigkeit von 94,62%, Qian et al. 78,56% und von Bao et al. einer Genauigkeit von 69,95% [11, 10, 12].

Entnehmend der Tabelle **??** gibt es dennoch einige Gemeinsamkeiten zwischen den Studien. Sowohl Qian et al. als auch Bao et al. verwendeten für ihre Netztopologie ein „back propagation neural network“ (BPNN) ([10] S. 2739, [12] S. 782), wohingegen Das et al. als besten Kandidat ein „feed forward neural network“ (FFNN) nominierte ([11] S. 2). Bezüglich den Aktivierungs- und Verlustsfunktionen verwenden Qian et al. und Das et al. jeweils die Softmax-AF und die Cross-Entropy-VF ([10] S. 2742, [11] S. 2). Bao et al. hat hier hingegen eine Sigmoid-Funktion als AF und den „Levenberg Marquardt“-Algorithmus als VF verwendet ([12] S. 783). Sowohl in der gewählten Lernrate, dem gewählten Optimierungsverfahren und den gewählten Eingabevariablen unterscheiden sich die drei Studien maßgeblich. Qian et al. nutzten 3 verborgene Knoten mit einer Lernrate von 0,01, während Bao et al. 388 verborgene Knoten und eine Lernrate von 0,005 für ihre Analysen verwendeten. Das et al. hat zu beiden Attributen keine Angeben gemacht.

Tabelle 1: Ergebnisse der Literaturrecherche

Autoren	Qian et al. [10]	Das et al. [11]	Bao et al. [12]
Topologie	BPNN	FFNN	BPNN
Aktivierungsfunktion	Softmax	Softmax	Sigmoid
Verlustsfunktion	Cross-Entropy	Cross-Entropy	Levenberg Marquardt
Verborgene Knoten	3	o.A.	388
Lernrate	0.01	o.A.	0.005
Optimierungsverfahren	Gradient Descent	Adam	o.A.
Genauigkeit	78,56%	94,62%	69,95%
Eingabevariablen			

4 Diskussion und Fazit

Im Rahmen der SLR werden im Folgenden die Ergebnisse kritisch untersucht, Limitationen sowohl der Studien als auch dieser SLR genannt und zusätzlich zum Fazit eine

Empfehlung für zuküftige Forschungen im Bereich der Verwendung von KNN für Verkehrsanalysen ausgesprochen.

4.1 Diskussion

Alle drei inkludierten Studien sind zu maßgeblich unterschiedlichen Ergebnissen auf unterschiedlichen Wegen gekommen. Das et al. und Bao et al. stechen mit Genauigkeiten von jeweils 94,62% und 69,95% aus den Rechercheergebnissen heraus. Im Rahmen einer Verkehrsanalyse sollten die verwendeten Modelle so zuverlässig wie nur möglich sein, allerdings besteht die Gefahr, dass das trainierte Modell einem sog. „over-fitting“ (OF) oder „under-fitting“ (UF) unterliegt. OF bezeichnet eine zu starke Anpassung des Modells an die Trainingsdaten, weswegen es neue Daten nicht korrekt auswertet [13]. Dies kann unter anderem passieren, wenn die Größe der Trainingsdaten zu klein, die Trainingsdaten für den gewählten Klassifizierungskontext irrelevante Datensätze beinhalten (auch genannt „verrauschte Daten“) oder die Parameter des Modells zu komplex gewählt wurden [13]. Das Pendant dazu ist ein vorliegen eines UF. Hierbei ist das Modell zu simpel, kann Muster dementsprechend nicht auffassen und lernen [14].

- ~~Interessante Differenz der Genauigkeiten~~
- Auf die schlechte genauigkeit von Bao et al. eingehen
- Allgemein aber darauf eingehen, dass KNNs in allen Artikeln die beste Genauigkeit erzielen konnten
- Eingabevariablen vergleichen
- Trainingsdaten vergleichen
- Das et al. Ausreißer?

4.2 Limitationen

- Beschränkt im betrachteten Zeitintervall
- Unklar, welche Konfigurationen die geeignetesten sind

4.3 Fazit

- KNN als bestes Modell zum Vorhersagen von RTAs
- Optimale Parameter sind schwer zu finden

4.4 Ausblick auf zukünftige Forschung

- Mehr Artikel mit genaueren Beschreibungen der Netze.
- Gerne auch andere Topologien erkunden.

Literatur

[1] World Health Organization (WHO). *Road Traffic Injuries*. 2023. URL: https://www.who.int/news-room/fact-sheets/detail/road-traffic-injuries.

[2] Systemanalyse Programmentwicklung (SAP). *Was ist Machine Learning?* 2024. URL: https://www.sap.com/germany/products/artificial-intelligence/what-is-machine-learning.html.

[3] International Business Machines Corporation (IBM). *Was ist ein neuronales Netz?* 2024. URL: https://www.ibm.com/de-de/topics/neural-networks.

[4] Deeksha Goplani. *Activation functions, loss functions & Optimizers*. 2023. URL: https://medium.com/@deeksha.goplani/activation-functions-loss-functions-optimizers-6bd0316898ae.

[5] Amitayas Banerjee et al. „Comparative Analysis of Machine Learning and ANN models for Mortality prediction in RTAs“. In: *2023 OITS International Conference on Information Technology (OCIT)*. 2023, S. 698–702. DOI: 10.1109/OCIT59427.2023.10431379.

[6] Ayush Maurya et al. „A Key Factor in Traffic Management - Vehicle Speed Prediction Using Machine Learning“. In: *2023 4th International Conference on Computation, Automation and Knowledge Management (ICCAKM)*. 2023, S. 1–6. DOI: 10.1109/ICCAKM58659.2023.10449542.

[7] Jim Frost. *How to interpret R-squared in Regression Analysis*. o. A. URL: https://statisticsbyjim.com/regression/interpret-r-squared-regression/.

[8] Ennaji Fatima Zohra et al. „Accident Severity Prediction using Machine Learning: A case study on the US Accidents Dataset“. In: *2023 17th International Conference on Signal-Image Technology & Internet-Based Systems (SITIS)*. 2023, S. 242–246. DOI: 10.1109/SITIS61268.2023.00044.

[9] David Moher et al. „Preferred Reporting Items for Systematic Reviews and Meta-Analyses: The PRISMA Statement“. In: (2009). DOI: 10.1371/journal.pmed.1000097.

[10] * Ruyi Qian und Xin Wang. „Prediction of accident severity based on BP neural networks“. In: *2023 35th Chinese Control and Decision Conference (CCDC)*. 2023, S. 2739–2744. DOI: 10.1109/CCDC58219.2023.10327407.

[11] * Surojit Das et al. „Machine Learning Based Approach for Predicting the Impact of Time of Day on Traffic Accidents“. In: *2023 26th International Conference on Computer and Information Technology (ICCIT)*. 2023, S. 1–5. DOI: 10.1109/ICCIT60459.2023.10441325.

[12] * Chun Bao et al. „Research on Classification and Prediction of General Traffic Accidents on National and Provincial Highways Based on BP Neural Networks“. In: *2024 4th International Conference on Neural Networks, Information and Communication Engineering (NNICE)*. 2024, S. 781–784. DOI: 10.1109/NNICE61279.2024.10498278.

[13] Amazon Web Services (AWS). *Was ist Overfitting?* 2023. URL: https://aws.amazon.com/de/what-is/overfitting/.

[14] International Business Machines Corporation (IBM). *What is underfitting?* 2024. URL: https://www.ibm.com/topics/underfitting.

Abbildungsverzeichnis

Tabellenverzeichnis