Schlussbericht

„KI-LOK: KI-Lokomotivesysteme – Prüfverfahren für KI-basierte Komponenten im Eisenbahnbetrieb“

Projektpartner

- ITPower Solutions GmbH
- neurocat GmbH
- Hitachi Rail (GTS Deutschland GmbH) - ehemals Thales Deutschland GmbH
- Fraunhofer Institut für Offene Kommunikationssysteme FOKUS
- Heinrich-Heine-Universität Düsseldorf

Projektlaufzeit: 01.04.2021 – 30.09.2024

Finanziert von der Europäischen Union
NextGenerationEU

Gefördert durch:

Bundesministerium für Wirtschaft und Klimaschutz

aufgrund eines Beschlusses des Deutschen Bundestages

Das diesem Bericht zugrunde liegende Vorhaben wurde mit Mitteln des Bundesministeriums für Wirtschaft und Klimaschutz unter den Förderkennzeichen 19I21007A gefördert. Die Verantwortung für den Inhalt dieser Veröffentlichung liegt bei den Projektpartnern.

INHALT

1. Durchgeführte Arbeiten und erzielte Ergebnisse

1.1. Hauptarbeitspaket 1 (HAP1): Fallstudien

1.1.1. Normung und Richtlinien für die KI-Absicherung

Gemeinsam mit den Projektpartnern erfolgte eine Recherche zu existierenden Normen und Richtlinien für die KI-Absicherung in den Bereichen Bahntechnik, Automotive und Medizintechnik. Die relevanten Dokumente wurden festgestellt und in Detail analysiert. Im Ergebnis der Analyse wurden Anforderungen an die KI-Absicherungsmethoden im Kontext des Projekts formuliert. Der Fokus der Arbeit von ITPower Solutions in diesem Arbeitspaket waren die Analyse relevanter Dokumente aus dem Bereich Medizintechnik und die Formulierung der daraus resultierenden Anforderungen.

1.1.2. KI-LOK Demonstrator

Die Testwerkzeugkette wurde prototypisch in einem Demonstrator anhand des Testszenarios einer Bahnstrecke in ländlicher Umgebung mit an einem Bahnhof wartenden Passagieren und zu erkennenden Signalen entlang der Strecke umgesetzt.

In Abbildung 1 rechts ist das Testszenario mit Hilfe der konzeptuellen Simulation dargestellt. Grüne Punkte stellen Bäume dar, das Gleis ist als schwarze Linie dargestellt. Die Rot, Orange und Cyan gefärbten Symbole stellen verschieden Signaltypen dar, die blauen Kreise zeigen die Position der Passagiere am, durch graue Rechtecke dargestellten, Bahnhof an.

Das Testszenario wurde automatisiert durch ein Sampling aus der Methode PEON (siehe 1.2) in verschiedenen Wetterbedingungen generiert, die verschiedene Einflüsse auf das KI-Perzeptionssystem vom Projektpartner Hitachi prüfen:

1. Leichter Schnee zur Mittagszeit: keine Schlagschatten, helle Umgebung. Voraussichtlich gute Bedingungen für das Perzeptionssystem.
2. Regen am Abend: dunkle Umgebung, Überlagerung des Bildes durch Regentropfen. Voraussichtlich herausfordernde Bedingungen für das Perzeptionssystem.
3. Starker Regen am späten Abend: sehr dunkle Umgebung, starke Überlagerung des Bildes durch Regentropfen. Voraussichtlich sehr herausfordernde Bedingungen für das Perzeptionssystem.
4. Klare Bedingungen in der Nacht: sehr dunkle Umgebung aber klare Sicht. Voraussichtlich herausfordernde Bedingungen für das Perzeptionssystem.
5. Leichter Nebel am Vormittag: gute Sicht, z.T. starke Schlagschatten. Voraussichtlich mittelmäßige Bedingungen für das Perzeptionssystem.

Eine qualitative Auswertung des Verhaltens der Perzeptionssystems ergab, dass die dunklen Umgebungen in 2., 3. einen starken Einfluss auf die Leistung der Objekterkennung und Segmentierung hatten. Der starke Regen in 3. hat das Perzeptionssystem so stark beeinflusst, dass es nicht mehr zu gebrauchen war. Starke Schlagschatten in 5. haben die Gleiserkennung und -segmentierung sehr stark beeinflusst und z.T. vollständig ausgehebelt. Des Weiteren kann allgemein gesagt werden, dass Überlagerungen von Teilen der zu erkennenden Objekte die Leistung des Perzeptionssystems stark negativ beeinflusst haben.

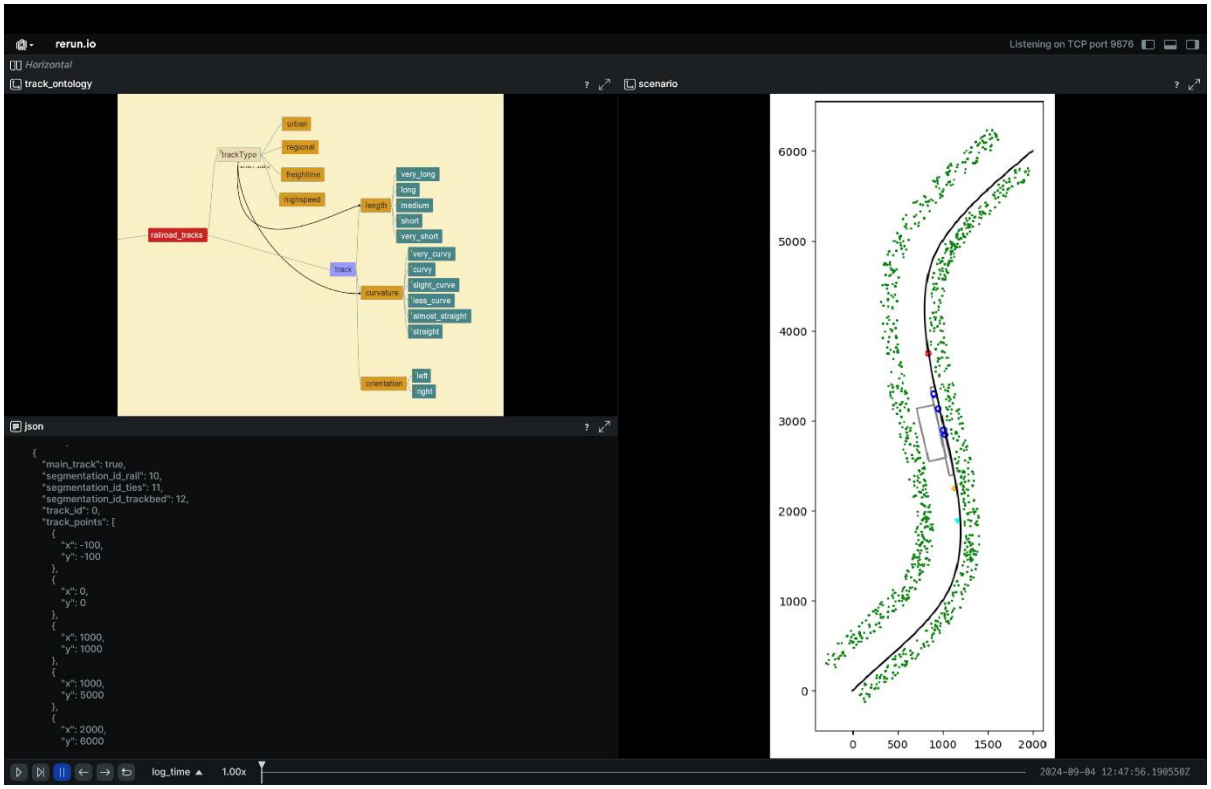

Abbildung 1: Visualisierung des Testszenarios in der PEON (oben links), dessen abstrakte Beschreibung im JSON-Format (unten links) und Konzeptuelle Simulation (rechts).

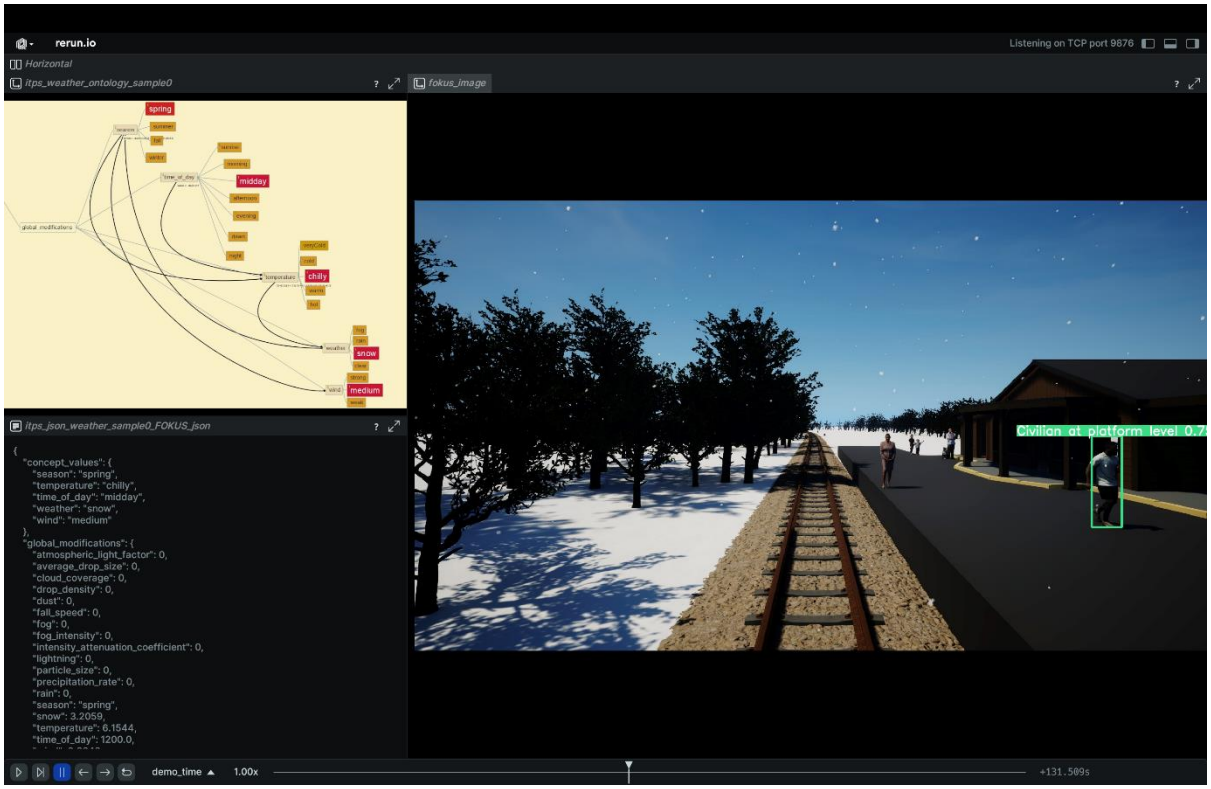

Abbildung 2: Die Wetterbedingungen des Testszenarios wurden automatisiert aus der PEON (oben links) gesamplet und über die JSON-Schnittstelle (unten links) an die 3D-Simulation (Beitrag des Projektpartners Fraunhofer FOKUS) zur Erstellung der konkreten Testdaten/ -bilder (rechts) übergeben.

Abbildung 3: Visualisierung der Bilderkennung mittels KI-Perzeptionssystems (Beitrag des Projektpartners Hitachi, links) und Certifying Control (Beitrag des Projektpartners HHU, rechts).

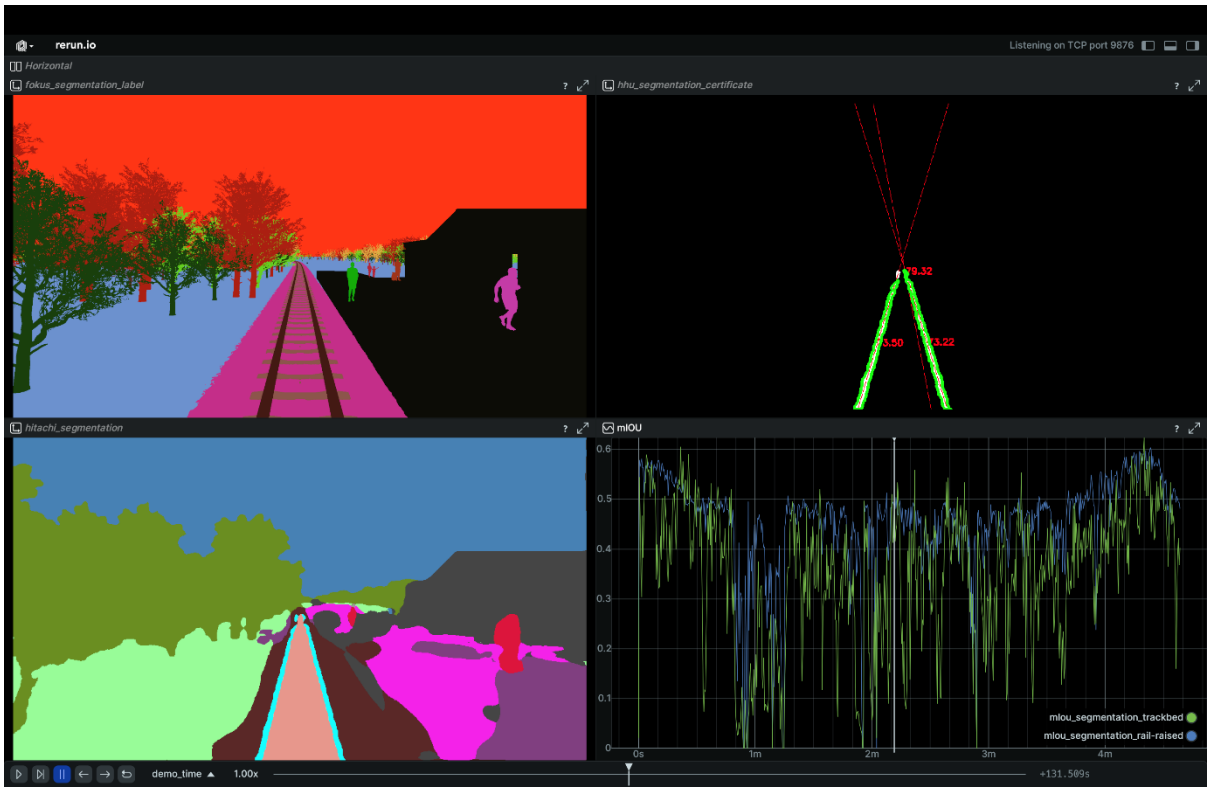

Abbildung 4: Validierung der Bildsegmentierung des verwendeten KI-Perzeptionssystems (Beitrag des Projektpartners Hitachi). Ground Truth der Objektpositionen (Beitrag des Projektpartners Fraunhofer FOKUS, oben links), Ergebnis der Segmentierung (unten links), Vergleich der Position der Schienen (Beitrag des Projektpartners HHU, oben rechts), Metrik für die Genauigkeit der Segmentierung der Schienen und des Gleisbetts (Beitrag des Projektpartners HHU, unten rechts).

Abbildung 5: Realisierung des Testszenarios mit starkem Nebel, umgesetzt Augmentation durch den Projektpartner neurocat (links). Fehlklassifikation einige Blätter als Auto ("Car", rechts).

1.1.3. Fallstudien

Der oben beschriebene Demonstrator wurde genutzt, um die Fallstudien 1 (Objekterkennung) und Fallstudie 2 (Lokalisierung) prototypisch umzusetzen und damit die Machbarkeit der entwickelten Methodik zu zeigen.

Dazu wurden die im Lichtraum zu erkennenden Objekte identifiziert (Personen, Autos, Signale, etc.), in der PEON eingefügt und Statistiken für die Auftrittswahrscheinlichkeit der Objekte modelliert oder recherchiert. Die identifizierten Objekte wurden dabei vereinfacht in der Ontologie modelliert und nicht alle möglichen Ausprägungen wurden übernommen.

Neben den Identifizierten Objekten wird die Erstellung der Bahnstrecke ebenfalls aus der PEON generiert (siehe auch Abschnitt Konzeptionelle Simulation unter 1.2), um vielfältige Streckenverläufe zu ermöglichen und diese Strecken mit verschiedenen Störungen in Form von Hindernissen (z.B. quer über dem Gleis liegende Baumstämme oder das Gleis passierende Autos) oder Umwelteinflüssen (z.B. starker Regen oder Nebel) anzureichern.

1.2. Hauptarbeitspaket 2 (HAP2): Validierungs- und Verifikationstechniken für ML

1.2.1. Systematischer Test von KI-Komponenten im Bahnbereich

Scenic Description

KI-Systeme sind nicht entlang bekannter Regeln programmiert, die sie zu befolgen haben, sondern über aufwendiges Lernen mit Trainingsdaten. Dabei ist die ausgewogene Auswahl der Trainingsdaten entscheidend, für die Güte des angelernten Systems, ihre Fähigkeit, auch in neuen Situationen adäquat zu reagieren (Generalisierung), sowie ihre Robustheit gegenüber Störungen. Wie aber ist ein

so komplexes Umfeld wie die möglichen Sichten aus einer Zugfahrerkabine während der Fahrt für solche Betrachtungen zugänglich?

Scenic Description ist ein aus der Automobilbranche stammender Ansatz, verschiedene Situationen beim Autonomen Fahren simulieren zu können und Trainings- oder Testdaten abzuleiten.

Die möglichen Situationen für das autonome Bahnfahren sind in vielerlei Hinsicht verschieden, einfacher, als im Automotivbereich. Überholverkehr und Kreisverkehre treten nicht auf und so reicht der Blick nach vorn. Um einen Einsatz dieses Ansatzes in der Bahntechnik zu ermöglichen, wurde in Zusammenarbeit mit dem Projektpartner Fraunhofer Fokus eine Ontologie zur Beschreibung von Szenen erstellt (Geometrie der Bahngleise, Topologie der Umgebung, Umwelteinflüsse, technische Systeme, etc.), sowie ihre Beziehungen probabilistisch modelliert, was die mögliche Komplexität weiter reduzieren hilft.

Um mittels der abstrakten Elemente der Ontologie konkrete Szenen zu beschreiben und zu generieren, sind konkrete Instanzen zu ihnen zu definieren. In einem Klassifikationsbaum wurden deshalb verschiedene Kompositionen und Klassen möglicher Fahrstrecken, Wetterverhältnissen, Akteure usw. definiert, deren Zusammenspiel dann eine Szene bildet. Im späteren Verlauf des Projektes wurde ausgehend von diesen Vorarbeiten ein stochastisches Modell möglicher Umgebungen erstellt werden.

Metamorphe Bildtransformationen zur Erweiterung von Trainings- und Testdaten Neuronaler Netzwerke

Es gibt vielfältige Techniken, stochastisch Trainings- und Testdaten zu generieren (GANs, Markov Automaten, ...). Doch hat man in diesen Fällen kein Labelling und folglich kein Orakel für den Test. Hier sind Techniken des Metamorphen Testens vorzuziehen. Dabei nutzt man Transformationen von Ein- und Ausgangspaaren, unter denen beide Größen sich kovariant transformieren. Auf diese Weise lassen sich Trainings- und Testdatenmengen mit gegebenen Erwartungen erweitern.

Natürliche Transformationen sind im Kontext der optischen Objekterkennung z.B. verwackelte Bilder, wie sie durch fehlende Arretierung der Kamera entstehen, oder durch Linsenverzerrungen. Aber auch geänderte Lichtverhältnisse und Tageszeiten modifizieren die Bilder auf realistische Weise.

Im Rahmen dieses Arbeitspakets wurden verschiedene metamorphe Bildtransformationen, die im Bahnbetrieb auftreten können, identifiziert sowie implementiert und auf Testbilder angewandt. Ziel war, ein erweitertes Testen von KI-Systemen bei realistischen Störungen. Anhand der Wahrscheinlichkeit ihres Auftretens, ihres Ausmaßes, und Anwendung in verschiedenen gefährlichen Situationen kann dann zu einem KI-System eine Risiko-Matrix definiert werden, welche ihre Eignung in kritischen Situationen fasst.

Der Einfluss verschiedener metamorpher Bildtransformationen wurde am Beispiel eines neuronalen Netzes zur Objekterkennung untersucht.

Unter Verwendung der mean Average Precision (mAP) als Maß für die Genauigkeit (Accuracy) der Vorhersage des Netzwerkes wurde mit 26 metamorphen Bildtransformationen (u.a. Verwackeln, Änderung von Lichtverhältnissen, Hoch- und Tiefpass Filter, Linsenfehler) die Abweichung in der mAP des gestörten neuronalen Netzes vom ungestörten Netz untersucht.

Es wurde beobachtet, dass Änderungen der Lichtverhältnisse und Farbveränderungen, oder allgemeine Transformationen im Farbraum, einen kleinen Einfluss auf die mAP haben, während Transformationen im Pixelraum (Rotation, Verzerrungen oder verschiedene Filter wie Hoch- und Tiefpass) genauso wie verschiedene Arten von Rauschen (Regen, Schnee oder Hitzeflimmern) eine große Verschlechterung der mAP herbeiführten.

Aus den Ergebnissen kann geschlossen werden, dass Störungen, die die Kanten von Objekten verändern, den größten Einfluss auf die mAP und damit auf die Accuracy haben.

Die oben aufgeführten Ergebnisse können unter Berücksichtigung von Rahmenbedingungen wie einer Operational Design Domain (ODD) genutzt werden, um eine Risikoanalyse durchzuführen. Eine exemplarische Risikoanalyse hatte zum Ergebnis, dass die bei Bewegung (z.B. der sich bewegende Zug mit On-Board Objekterkennung) zu erwartende radiale Bewegungsunschärfe in Verbindung mit großen Objekten wie z.B. einem Auto, Boot oder LKW ein großes Risiko darstellen.

Ein weiteres Ergebnis ist, dass unter Verwendung durch metamorphe Bildtransformationen erweiterter Datensätze und einer Risikoanalyse neuronale Netze prinzipiell auf Robustheit geprüft werden können.

Die Anwendung dieser Techniken auf Trainingsdaten kann zu besserer Robustheit des KI-Systems bei der Objekterkennung führen. Diese Fragen wurden im weiteren Verlauf der Arbeit untersucht.

Genauere Details finden sich in der Masterarbeit von Roman Krajewski „Untersuchung von metamorphen Bildtransformationen zur Erweiterung von Trainings und Testdaten Neuronaler Netzwerke“

Ontologie

Ziel ist, anhand einer Ontologie, erweitert um probabilistische Modelle, Szenarien für das Training und den Test eines Objekterkennungssystems zu generieren. In Zusammenarbeit mit Fraunhofer Fokus wurde zuerst eine einfache Ontologie (Geometrie der Bahngleise, Topologie der Umgebung, Umwelteinflüsse, technische Systeme, etc.) zur Beschreibung von Szenen erstellt und im Projektverlauf fortlaufend verfeinert. Diese Ontologie bildet schließlich die Grundlage für die prototypisch entwickelte Testtoolkette und ihrer Implementierung in einem Demonstrator. Die Genese dieser Ontologie wird in diesem Abschnitt erläutert

Für die Ontologie wird von einer formalen Beschreibung der verschiedenen möglichen Objekte in einem Szenario ausgegangen, in denen sich Züge, Fahrzeuge und Menschen in verschiedenen Arten bewegen können.

Im Projekt wurden bahnrelevante Szenarien als Grundlage für die Ontologie betrachtet. So sind die betrachteten Klassen der Ontologie

- die Streckengeometrie (Track Geometry), die die Streckenführung z.B. gerade und gekrümmte Streckenführung oder Steigungen enthält.
- die Streckenumgebung (Specific Environments), die z.B. die angrenzende Umgebung wie urban, ländlich oder alpin enthält (dies entspricht der Scenery im Kontext der Scenic Description).
- die Umgebungsbedingungen (Environmental Conditions), die z.B. Wetterbedingungen, Tages oder Jahreszeiten enthält.
- die Akteure (Actors), die eine Abhängigkeit zu den Specific Environments und den Environmental Conditions haben. So sind z.B. in einer alpinen Umgebung im Winter Skifahrer, oder in ländlichen Gegenden Kühe und Schafe zu erwarten. Im Sommer in einer urbanen Umgebung hingegen werden weder Skifahrer noch Kühe und Schafe zu erwarten sein.

Die Ontologie wurden die Klassen Environmental Conditions um kurzfristige Anpassungen (z.B. Bauarbeiten) und Actors um statische (z.B. Signale) sowie nicht stationäre Akteure (z.B. Fahrzeuge), denen dynamische Statusänderungen (z.B. Signalschaltung oder sich bewegende Fahrzeuge) von einer Szene zur nächsten zugewiesen werden können, erweitert.

Anstelle sämtlich möglicher Entitäten in Szenarien zu betrachten, wurde ein kleineres ontologisches Modell freigeschnitten, welches nur die möglichen Personen betrachtet, siehe Abbildung 6.

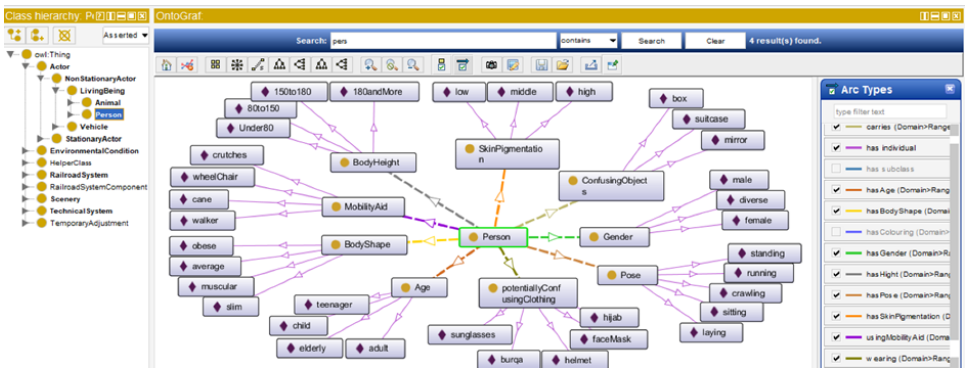

Abbildung 6: Eingeschränkte Ontologie möglicher Personen

Zur Analyse möglicher Abhängigkeiten von Attributen untereinander innerhalb einer Klasse wurde die Klasse „Person“ ausgewählt, da sie einen intuitiven Zugang auf der Grundlage von Alltagserfahrungen ermöglicht.

Bei der Klasse „Person“ werden einer generischen „Person“ Attribute wie; „age“, „ethnicity“, „gender“, „authorized“, und u.v.m., zugordnet. Im untersuchten Beispiel wird sich auf die Attribute; „speed“, „pose“, „riding“, „leading“, „mobilityaid“ und „carrying“ beschränkt, um die Kombinatorik zu begrenzen.

Zu dieser Ontologie wurde ein Wahrscheinlichkeitsraum definiert, der eine realistische Verteilung von Personen am Streckenrand beschreibt. Bestimmte Kombinationen von Eigenschaften wie ‚Ein Mensch mit Gehstock, der rennt‘ werden bereits in der Ontologie verboten. Das allgemeine Problem liegt in den verwickelten Abhängigkeiten verschiedener, zufälliger Eigenschaften von Personen: Eine Person kann jung oder alt sein, schnell laufen oder stehen, bei sich einen Koffer tragen oder nicht und ein Fahrrad schieben. Dabei ist es wahrscheinlicher, dass ein alter Mensch nicht so schnell läuft, und wenn doch, keinen Koffer bei sich hat und äußerst selten dazu noch ein Fahrrad schiebt. Hingegen wird seltener ein junger Mensch im Rollstuhl sitzen.

Der Wert eines Attributs kann einen Einfluss auf die Werte anderer Attribute haben, z.B. verbietet die „pose:laying“ jede Realisierung des Attributs „riding“ außer „riding:nothing“, da die betrachteten Fahrzeuge: „bike“, „car“, „horse“ nicht im Liegen geführt werden können. Eine Beispielrealisierung einer Person ist in Abbildung 7 zu sehen.

```
✓ sampled_person ···

{'person': 'person',
 'speed': '5 to 10',
 'pose': 'running',
 'leading': 'leading',
 'leadingwhat': 'dog',
 'carrying': 'carrying_nothing',
 'mobilityaid': 'no_mobilityaid',
 'ridingwhat': 'riding_nothing'}
```

Abbildung 7: Stochastisch generierte Instanz der Klasse "Person"

Des Weiteren kann auch die Festlegung eines Attributes vor einem anderen Einfluss auf die möglichen Realisierungen einer Klasse haben, z.B. könnte die Wahl von „riding:car" vor oder nach der Wahl des Attributs „pose" geschehen, was zu verschiedenen Einschränkungen führt. So sind, wie oben angedeutet, „pose:laying" und „riding:car" einander ausschließend.

Dazu wurde ein graphisches Modell der probabilistischen Abhängigkeiten entwickelt.

Die Erkenntnisse bei der Erstellung der Ontologie und von Abhängigkeiten zwischen verschieden Klassen der Ontologie waren Grundlage für die Implementierung der Toolkette.

Konzeptionelle Simulation

Die konzeptionelle Simulation dient dazu, die aus der oben beschriebenen Ontologie erstellten, abstrakten Testfälle auf einfache Art zu veranschaulichen. Die abstrakten Testfälle stellen die Grundlage für künstlich generierte konkrete Testdaten und Testdatensätze, die zur Validierung von KI genutzt werden. Prinzipiell können die erstellten Datensätze ebenfalls zum Training von KI genutzt werden, dieser Use Case wurde im Projekt allerdings weder verfolgt noch dessen Auswirkungen untersucht.

Die Auswahl der Szenarien zur Generierung der Datensätze geschieht zufällig aus den von der Ontologie zur Verfügung gestellten Klassen und Bedingungen (probabilistische Abhängigkeiten, siehe auch unten, Abschnitt Test-Workflow. Ein erster Entwurf einer Ontologie, in Form eines Klassifikationsbaums, ist in Abbildung 8 abgebildet.

Ist eine Szenerie, zufällig ausgewählt, z.B. Bahnstrecke in ländlicher Umgebung mit Feld auf einer Seite der Strecke und einem Wald auf der anderen Seite der Strecke, wird in einem weiteren Schritt die Szenerie um Objekte und Akteure, wie eine die Schienen kreuzende Straße, einem Fahrzeug auf der Straße, Verkehrszeichen und Signalen entlang der Strecke erweitert. Spezielle Akteure wie Fahrzeuge oder Signale können entsprechend einer zuvor definierten Dynamik im zeitlichen Verlauf (Abfolge von Szenarien aufeinander) ihren Zustand ändern, ein Fahrzeug bewegt sich entlang einer Straße und ein Signal wechselt sein Zeichen.

Die so generierten dynamischen Szenarien können in einer simplen zweidimensionalen Darstellung ausgegeben oder mittels einer Schnittstelle an eine weitere Verarbeitung weitergegeben werden. Im Rahmen des Projekts dienten die generierten konzeptionellen Simulationen, bestehend aus Ortsdaten, Akteuren und Dynamiken, als Grundlage für die Erzeugung (fotorealistischer) Validierungsdatensätze durch den Projektpartner Fraunhofer Fokus.

Abhängig von der Reichhaltigkeit der Ontologie, der Vollständigkeit der verschiedenen Objektklassen und ihren potentiellen Dynamiken lässt sich dieser Ansatz kontinuierlich erweitern.

Auf diese Weise werden Mit Hilfe einer konzeptuellen Simulation grobe Szenarien stochastisch generiert, aus denen in weiteren Verfeinerungsschritten realistische Szenarien abgeleitet werden. Dazu werden im ersten Schritt zufällige Gleisstrecken generiert. Ausgehend von einer zufälligen Anzahl von Kurven mit Längen und Krümmungsradien oder von der zufälligen Anzahl paralleler Gleise und Länge werden dann entsprechende Gleisstrecken generiert.

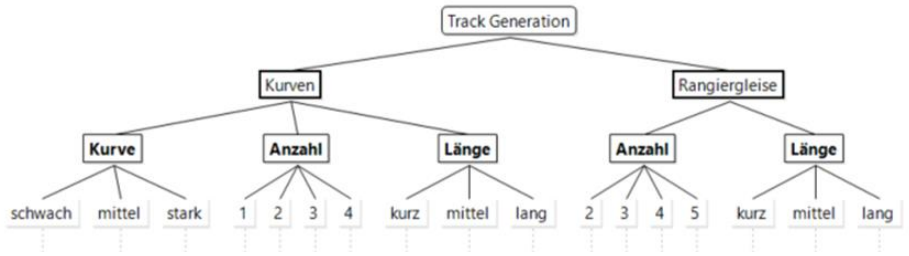

Abbildung 8: Klassifikationsbaum für die Gleisstrecken Generierung

In der ersten Version des Simulationstools konnten noch keine Randbedingungen über die generierten Weichenkrümmungen von Rangiergleisen berücksichtigt werden. Dies wurde in einem zweiten Schritt erfolgreich umgesetzt, so dass realistischere Strecken generiert werden, siehe Abbildung 9.

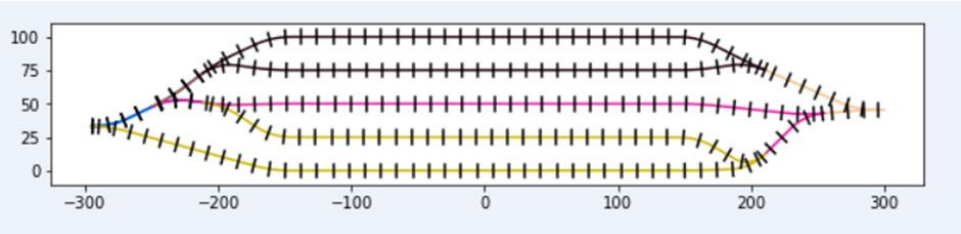

Abbildung 9: Stochastisch generierte Rangiergleise

In einem weiteren Schritt können nun Bäume, Häuser, etc. stochastisch den Strecken hinzugefügt werden.

Teststrategie

Testen sucht nach möglichem Fehlverhalten, verursacht durch Fehler. Aber was sind Fehler eines KI-Systems, geeignete Fehlermodelle für sie? Aufgrund ihrer statistischen Natur ist das Aufzeigen einer Fehleinschätzung, z.B. ein Schaf wird als Kuh erkannt, noch kein Fehler. Es wird immer Fehleinschätzungen geben, es fragt sich nur, wie häufig!

In einem gemeinsamen Dokument (Test ML basierter Systeme) der Projektpartner versuchten wir uns der Frage zu nähern: Was kann alles falsch gehen und wie kann man diese Fehler aufdecken?

Herausforderungen entstehen unter anderem durch den offenen Kontext der Anwendung, der stochastischen Natur sowie dem Zusammenspiel mit klassischen Software-Systemen. In dem Dokument konzentrierten wir uns auf die besonderen Anforderungen an den Test von KI-basierten System und welche Fehler und zugehörige Testmethoden sich ggfs. aus dem Test klassischer Software übertragen oder auf die neuen Herausforderungen adaptieren lassen. Von ITPower Solutions ist insbesondere das Kapitel 5 (*Test methods for testing ML-based systems*) verantwortlich bearbeitet worden.

Test-Workflow

Die Methoden der Modellierung einer Operational Design Domain (ODD) mit einer Ontologie und stochastischen Generierung von Szenarien über eine konzeptionelle Simulation wurden genutzt, um mit diesen Methoden einen systematischen Testansatz zu konzipieren und zu implementieren, der die statistische Natur von KI berücksichtigt und im „Safety Evaluation Process“ [1] Anwendung findet.

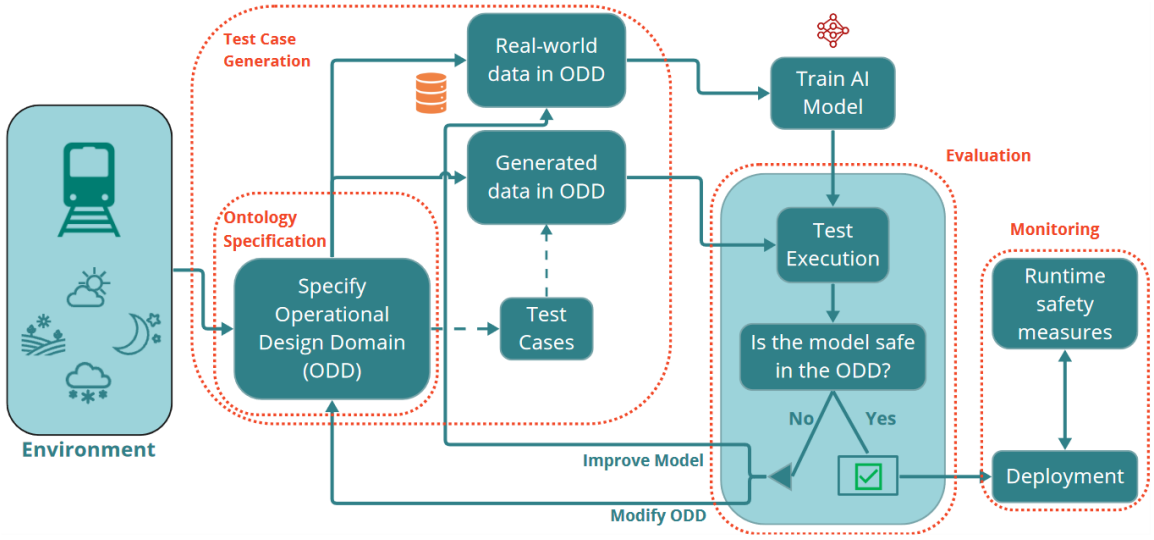

Abbildung 10: Abbildung des Safety Evaluation Process [1]

Eine weitverbreitete und etablierte Methode für systematische Softwaretests ist die Pair-Wise (oder allgemeiner n-Wise, für Pair-Wise ist n=2) Testmethode. Der Name beruht auf der Auswahlstrategie der Testfälle: hat eine Software eine gewisse Anzahl an Eingabeparametern, die wiederum gewisse Werte annehmen können z.B. A aus der Menge {a1, a2}, B aus {b1, b2, b3} und C aus {c1, c2}, so ergibt sich eine hohe Anzahl an Testfällen, um jede Kombination abzutesten. Im Beispiel würde es sich um 2x3x2=12 Testfälle handeln. In der Pair-Wise Methode werden nun alle Parameterwerte paarweise mindestens einmal abgetestet, was zu einer Reduktion der Testfälle führt. Im Beispiel kann die Anzahl an Testfällen von 12 auf 6 reduziert werden; T1: (a1, b1, c1), T2: (a1, b2, c2), T3: (a2, b3, c1), T4: (a2, b1, c2), T5: (a2, b2, c1), T6: (a1, b3, c2). Der Methode liegt die Annahme zu Grunde, dass das Verhalten der Parameter untereinander unabhängig ist. Diese Annahme kann jedoch im Allgemeinen bei KI-Komponenten nicht getroffen werden. Deshalb wurde die Pair-Wise Testmethode analytisch und experimentell auf ihre Tauglichkeit beim Test von KI geprüft und ihr ein neuer Ansatz, der Auswahl von Testszenarien auf Grundlage einer probabilistisch erweiterten Ontologie (PEON), entgegengestellt.

Im konkreten Fall des KI-LOK Projekts wird als Testobjekt ein Objekterkennungssystem in der Bahndomäne getestet. Dieses Objekterkennungssystem wurde auf Trainingsdaten einer Bahnspezifischen ODD angelernt wurde, mit der Aufgabe, die auf den Eingabebildern befindlichen Objekte und deren Ort zu klassifizieren. Die Objekte und deren Verortung werden dabei im Objekterkennungssystem über eine die ODD repräsentierende, Wahrscheinlichkeitsverteilung dieser Objekte klassifiziert. Um repräsentative Daten von Szenarien und darin enthaltenen Objekten zu erhalten, wird die betreffende ODD mit einer Ontologie modelliert. Die Ontologie stellt dabei Eingabeparameter für Testfälle durch Klassen, (z.B. Person) und die Werte der Eingabeparameter durch Attributwerte der Klassen (z.B. Alter: zwischen 21 und 35 Jahren) dar. Werden die Abhängigkeiten und Einflüsse der Werte verschiedener Attribute der ODD entsprechend untereinander berücksichtigt, so approximiert die Ontologie die Wahrscheinlichkeitsverteilung der in der ODD möglichen und nicht-möglichen Szenarien und den darin enthaltenen Objekten.

Wird nun die Testfallauswahl der zu prüfenden Szenarien auf Grundlage der Pair-Wise Testmethode gewählt, so entspricht dies der gleichverteilten Auswahl von Szenarien der ODD. Es kann davon ausgegangen werden, dass im Allgemeinen die Szenarien einer ODD nicht gleichverteilt sind und somit auch die Menge der durch Testfälle gefundenen Fehler, nicht repräsentativ für die im Betrieb zu

erwartende Menge an Fehlern ist. Werden die Szenarien gleichverteilt ausgewählt, wird im Vergleich zur wahren Verteilung der ODD systematisch eine größere Menge an Fehlern erwartet, als wenn die wahre Verteilung für die Testauswahl genutzt würde.

Weiter ermöglicht die Erstellung einer repräsentativen Verteilung für die Auftrittswahrscheinlichkeiten der ODD mithilfe der PEON erst die Bewertung der Statistik der Testergebnisse einer KI-Komponente innerhalb ihrer ODD. Erst eine repräsentative Vergleichsverteilung der ODD ermöglicht zu bestimmen, mit welcher Häufigkeit ein Fehlerfall innerhalb der ODD auftritt und ob ein Fehlerfall potenziell katastrophale Auswirkungen hat oder nicht.

Konstruktion einer probabilistisch erweiterten Ontologie (PEON)

Der Ansatz der probabilistisch erweiterten Ontologie verbindet stärken aus dem klassischen Software-Test und der Wahrscheinlichkeitstheorie. Im Folgenden wird skizziert, wie eine PEON konstruiert werden kann. Weitere Details finden sich in [2].

Nach dem bekannte Ansatz des Partition Testing, häufig umgesetzt über die Klassifikationsbaummethode, wird die Testparametermenge in eine Menge von Äquivalenzklassen partitioniert, so dass innerhalb einer Äquivalenzklasse die gleichen Testergebnisse erwartet werden (Uniformitätshypothese). Im Kontext einer Objekterkennung ist bekannt, dass verschiedene Lichtverhältnisse einen Einfluss auf die Klassifikationsleistung eines Objekterkennungsmodells hat. Deshalb ist es sinnvoll das Jahr in Jahreszeiten und den Tag in verschiedene Tageszeiten einzuteilen. Es sollte jedoch zu einer gegebenen Jahreszeit keinen Einfluss haben, ob ein Bild um 14 oder 15 Uhr aufgenommen wurde. Ein Unterschied wird hingegen sehr wohl erwartet, wenn ein Bild, um 14 Uhr nachmittags oder 22 Uhr in der Nacht aufgenommen wurde. Nach der Uniformitätshypothese wären demnach 14 und 15 Uhr in der gleichen Äquivalenzklasse (nachmittags), 22 Uhr wäre hingegen in einer anderen Äquivalenzklasse (nachts) zu erwarten.

Partition Testing gibt in diesem Kontext die Systematik der Testmethode vor, wohingegen die Wahrscheinlichkeitstheorie die Erstellung eines stochastischen Referenzmodells aus den Klassen der Ontologie ermöglicht. Jeder Klasse der Ontologie wird Wahrscheinlichkeitsverteilung der Auftrittswahrscheinlichkeit in der ODD zugewiesen. Die Wahrscheinlichkeitsverteilungen können, wenn vorhanden, aus statistischen Daten abgeleitet werden oder mit gesundem Menschenverstand modelliert werden (die so getroffenen Annahmen sollten empirisch validiert werden). Die so erhaltenen Verteilungen berücksichtigen im Allgemeinen nicht die Abhängigkeiten zu anderen Klassen der Ontologie, weshalb im Projektverlauf, auf Funktionen basierende, Algorithmen entwickelt wurden, mit denen die Abhängigkeiten zwischen verschiedenen Wahrscheinlichkeitsverteilungen modelliert werden können. Für eine Ontologie von Personen lassen sich zum Beispiel die statistischen Daten für die Größe, das Alter und die Größe in Abhängigkeit des Alters innerhalb der deutschen Bevölkerung leicht finden [3]. Die Haarlänge in Abhängigkeit des Alters hingegen wurde statistisch nicht erhoben, so dass diese mit einer funktionalen Abhängigkeit nach gesundem Menschenverstand modelliert wird, in der die Haarlänge im mittleren Alter häufiger vorkommt als in jungen Jahren oder bei Senioren (z.B. einer Gaußverteilung mit entsprechendem Mittelwert und Varianz).

Mit der so erstellten PEON lässt sich eine Größe bestimmen, die ohne diesen Ansatz nicht zugänglich war: das Testendekriterium.

Testendekriterium

Das Testendekriterium gibt an, unter welchen Bedingungen ein Test als abgeschlossen angesehen werden kann. Da es sich bei KI um stochastische Systeme handelt, reicht es nicht nur das Bestehen verschiedener Testfälle zu fordern, denn statistisch wird eine KI immer wieder Fehlerfälle erzeugen, sondern es muss ein statistisches Gütekriterium erfüllt werden, z.B. 10^{-3} Fehlerfälle pro Stunde für eine Klasse in der Ontologie. Dies bedeutet allerdings nicht, dass es ausreicht 1000 Testfälle

durchzuführen und nur einen negativen Testfall zu erlauben, da auch wenn 1000 Testfälle positiv verlaufen sind, die 10 folgenden fehlschlagen könnten. Das würde bedeuten, dass der Test zwar nach 1000 Testfällen positiv verlaufen würde, nach 1002 allerdings ein negatives Ergebnis zur Folge hätte. Um diesem Umstand Sorge zu tragen, müssen im Test zusätzlich zu einem Gütekriterium (im Beispiel 10^{-3} Fehlerfälle pro Stunde), Signifikanzparameter (Signifikanzniveau und ein Konfidenzintervall) gefordert werden, um die Zuverlässigkeit des Tests zu spezifizieren.

Mit Hilfe des Gütekriteriums, der Signifikanzparameter, des Erwartungswerts, der Varianz einer Klasse in der Ontologie und dem Zentralen Grenzwertsatz der Statistik lässt sich so die benötigte Anzahl an Tests zum Bestätigen oder Widerlegen des Gütekriteriums bestimmen. Dies entspricht dem Testendekriterium.

Auftrittswahrscheinlichkeit von Fehlern und Abschätzung der Datenqualität

Um die Auftrittswahrscheinlichkeit eines Fehlers entsprechend der gewählten ODD schätzen zu können, wird wie oben beschrieben eine Ontologie erstellt und dessen Klassen und Attribute um eine diskrete Wahrscheinlichkeitsverteilung, z.B. als Resultat einer statistischen Erhebung, ergänzt. Diese Konstruktion erzeugt aus der Ontologie ein Bayes'sches Netzwerk, dass die ODD modelliert und statistische Vorhersagen über die Auftrittswahrscheinlichkeit verschiedener Objekte und Szenarien zulässt.

In Abbildung 11 sind als Ausschnitt aus der Ontologie die Abhängigkeiten der Klasse „Person" und dessen Attribute „Sex", „Age" und „Height" durch ein grafisches Modell dargestellt. Für diese Attribute existieren statistische Erhebungen ihrer einzelnen Ausprägungen in der deutschen Bevölkerung, so dass eine Wahrscheinlichkeitsverteilung für das Auftreten der einzelnen Attributwerte definiert und die Auftrittswahrscheinlichkeit einer Ausprägung an jedem Entscheidungspunkt angegeben werden kann. Wird die gesamte Ontologie so mit Wahrscheinlichkeitsverteilungen aus statistischen Daten erweitert und dort wo keine Statistik vorliegt eine Schätzung vorgenommen, kann eine, von den Trainingsdaten unabhängige, repräsentative Beschreibung der ODD approximiert werden. So soll ermöglicht werden, auftretende Fehler entsprechend ihrer Auftrittswahrscheinlichkeit in der ODD zu bewerten und z.B. für Risikobewertungen heranzuziehen.

Die Unabhängigkeit der Ontologie von den Trainingsdaten der KI-Komponente ermöglicht darüber hinaus ebenfalls eine Beurteilung der Repräsentativität der Trainingsdaten. Je ähnlicher die Anzahl an Szenarien und Objekten im Trainingsdatensatz der Vorhersage der Ontologie für diese Szenarien und Objekte, desto besser bilden sie die ODD ab.

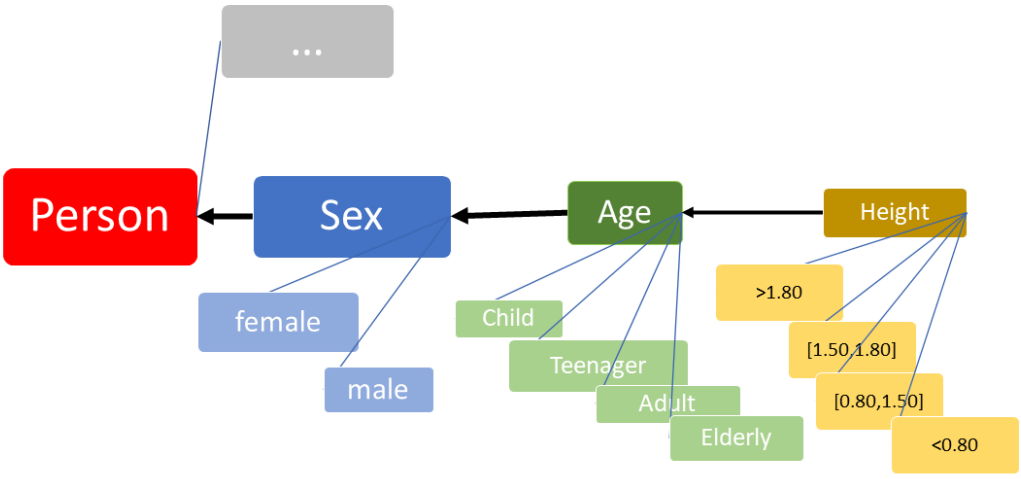

Abbildung 11: probabilistisch, grafisches Modell. Pfeile stellen Abhängigkeiten dar.

Dieser Ansatz wurde genutzt, um ein systematischen Testvorgehen zu konzipieren und zu implementieren, der die statistische Natur von KI berücksichtigt.

Des Weiteren wurde diese Methode in einer Testwerkzeugkette zur Erzeugung und Ausführung von Testfällen integriert. Die probabilistisch erweiterte Ontologie soll dabei aus einer abstrakten Testfallbeschreibung als Eingabe Szenarien und Objekte, entsprechend einer gewählten Wahrscheinlichkeit (repräsentativ für die ODD, eingeschränkt auf Randfälle, etc.) auswählen und diese Eingaben dem 3D-Simulationdwerkzeug zur fotorealistischen Simulation bereitstellen (siehe Abbildung 12).

Die probabilistischen Abhängigkeiten zwischen verschiedenen Klassen einer Ontologie können, soweit vorhanden, durch statistische Daten oder durch geeignete Annahmen angegeben werden. Diese Herangehensweise skaliert jedoch schlecht, da im Allgemeinen jede Klasse von mehreren weiteren Klassen abhängig sein kann und damit mit einem hohen Aufwand verbunden ist, um die passenden Statistiken zu erheben bzw. aufzufinden oder Annahmen an die Abhängigkeiten zu erstellen. Um dieses Problem zu lösen, werden Klassen mit unabhängigen Wahrscheinlichkeitsverteilungen genähert und diese Näherung, über eine Lösung des optimalen Transportproblems, stochastisch gekoppelt. So können abhängigen Klassen, funktional abhängige Wahrscheinlichkeitsverteilungen zugewiesen werden, ohne auf umfangreiche Statistiken oder Annahmen angewiesen zu sein.

Es lassen sich verschiedene Ontologien aufstellen, die verschiedene Anforderungen wie z.B. Repräsentativität der ODD, Robustheit gegenüber technischen Störungen und Umwelteinflüssen oder ethnische und geschlechtliche Gleichbehandlung genügen. Werden so aufgestellte Ontologien probabilistisch erweitert, können mit ihnen Performanzkriterien für die ODD und Randfälle geprüft werden. Des Weiteren weisen analytische Betrachtungen und Vergleiche zwischen den N-Wise Testmethoden (zur kombinatorischen Auswahl von Eingabedaten) und der Probabilistisch Erweiterten ONtologie (PEON) darauf hin, dass die N-Wise-Methode die Performanz einer Bilderkennungs-KI nicht eindeutig bewerten kann, während die PEON dies gewährleisten kann. Die N-Wise Testmethode berücksichtigt die probabilistische Natur von KI nicht ausreichend und enthält für N nicht maximal (N < Anzahl zur Auswahl stehender Klassen) Freiheitsgrade, die in der Testfallerstellung das Ergebnis der Tests nicht vorhersagbar verzerren. Sei als Beispiel N = 2 und die Anzahl der zur Auswahl stehenden Klassen 5, dann sind N-Wise Paare, die abgetestet werden müssen, zu wählen. Das heißt, in jedem Testfall werden die Werte von 2 Klassen eingeschränkt, während die übrigen 3 Klassen uneingeschränkt zufällige Werte annehmen können – dies sind die erwähnten Freiheitsgrade. Innerhalb der Freiheitsgrade hängt es von der spezifischen Performanz der KI innerhalb der Klasse ab, ob sie über- oder unterdurchschnittlich performt ist. Die N-Wise Testmethode kann daher im Allgemeinen keine Schranke für die Performanz der untersuchten KI angeben. Für N maximal (N = Anzahl zur Auswahl stehender Klassen) werden die Testfälle entsprechend der Gleichverteilung über die Testdaten erzeugt, da hier jede Testfallkombination genau einmal auftritt, was ebenfalls eine unangemessene Gewichtung der Fehlerfälle zum Ergebnis hat. Dies führt zu einer Unterbewertung der Performanz. Die PEON berücksichtigt die probabilistische Natur von KI hingegen in angemessener Weise und liefert daher zuverlässige Ergebnisse und ermöglicht so einen neuen, systematischen Testzugang.

Abbildung 12: Simulations- und Testautomatisierungskette

Experimentelle Validierung des Konzepts

Das oben beschriebene Testkonzept wurde mit einem überschaubaren Aufwand einer ersten experimentellen Validierung am Beispiel des CenterNet hourglass Modells [4] als System Under Test (SUT) und dem COCO-Datensatz [5] unterzogen.

Dazu wurde eine Ontologie aus den Superklassen „person“, „vehicle“, „outdoor“, „accessory“ als Parameter und deren Subklassen als Werte erstellt. Diese ist auf Abbildung 13 dargestellt. Ein Testfallsatz wird mittels eines abgewandelten AETG-Pair-Wise Algorithmus erstellt. Dieser erstellt zum COCO-Datensatz einen möglichst n-Wise abgedeckten Testfallsatz. Ein Testfall besteht dabei aus einem Bild, auf dem Objekte verschiedener Subklassen vorhanden sind. Jedes Bild deckt also die Kombinationen an Subklassen ab, die auf dem Bild vorhanden sind. Sind zu einer bestimmten Kombination an Subklassen keine Bilder im Testdatensatz vorhanden, können diese nicht abgedeckt werden. Je größer n, desto geringer ist also die Wahrscheinlichkeit, dass das n-Wise Kriterium erfüllt werden kann. Außerdem gibt es weniger Kombinationen an Bildern aus dem Datensatz, die das n-Wise Kriterium erfüllen, je größer n gewählt wird.

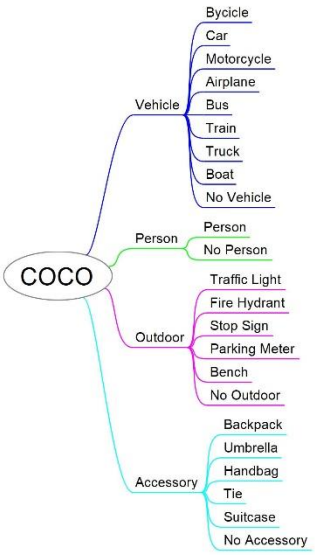

Abbildung 13: Ontologie zur Auswahl von Testfällen auf Grundlage des Coco Datensatzes

Um die Anwendbarkeit des Pairwise-Testing Verfahrens auf das Modell zu untersuchen, wurde die Performanz des Netzwerks bei zufälliger Auswahl nach dem n-Wise Verfahren verglichen. Da die Testfallsätze kleiner sind, je größer n gewählt wird, wurden bei kleinerem n mehrere Proben generiert und die durchschnittliche Performanz aller Durchläufe ermittelt, sodass für alle n mindestens 648 Bilder getestet wurden, was der vollen Kombination an Subklassen oder 4-Wise Testing entspricht.

Die Ergebnisse des Experiments sind auf Abbildung 14 zu erkennen. Es zeigt sich, dass das Neuronale Netzwerk mit zunehmendem n schlechter abschneidet. Daraus lässt sich ableiten, dass die Performanz des Netzwerks schlechter wird, je eher sich die Verteilung der Testdaten der Gleichverteilung annähert. Da bei kleinerem n die Testdatensätze kleiner sind und deshalb mehrfach gesampelt werden, ähnelt die gesamte Verteilung aller Durchläufe eher der Originalverteilung des Datensatzes. Dies zeigt die Wichtigkeit einer präzisen Ontologie, die die Wahrscheinlichkeitsverteilungen der Testfälle berücksichtigt.

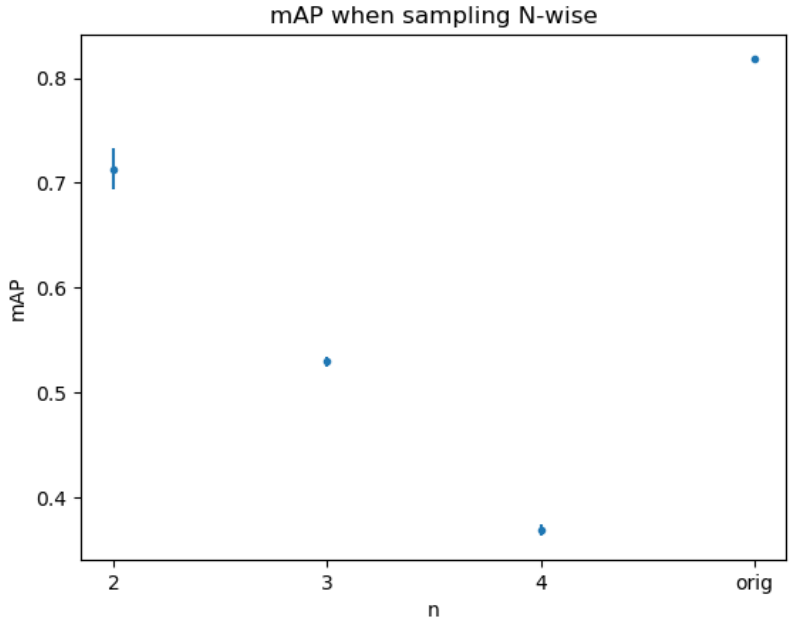

Abbildung 14 Ergebnisse des n-Wise Sampling Experiments

Eine tiefergehende Analyse bestätigte die ersten Experimente. In [6] können alle Details der tiefergehenden Analyse gefunden werden, im Folgenden werden deren Kernpunkte zusammengefasst.

Um das Konzept zu validieren und nachzuweisen, dass das Testergebnis einer KI-Komponente von der Auswahl der Testdaten abhängt, wurden die Datenqualität eines Datensatzes und die Genauigkeit (Accuracy) eines Klassifikationsmodels in Abhängigkeit der Testauswahlmethode untersucht.

Datenqualität

Zur Untersuchung der Datenqualität eines Datensatzes, wurde der PETA Datensatz [7] im Vergleich zur deutschen Bevölkerung analysiert und die Attribute „age“, „sex“, „hair length“ und „hair color“ mit einer PEON modelliert. Dabei wurden vier Attribute ausgewählt, um den Aufwand in Grenzen zu halten. Für die Attribute „age“ und „sex“ wurden leicht zugängliche Statistiken des statistischen Bundesamtes [3] benutzt. Für die Attribute „hair length“ und „hair color“ wurde die Verteilung in der Deutschen Bevölkerung von drei Forschern abgeschätzt, diese Abschätzungen gingen ebenfalls in die Testauswahl mit PEON ein. Für den produktiven Einsatz der Methode empfehlen wir die Schätzungen von Auftrittswahrscheinlichkeiten durch empirische Erhebungen abzusichern. Aus Ressourcengründen wurde in diesem Projekt darauf verzichtet.

Vergleicht man nun die marginalen Statistiken des statistischen Bundesamts/ der PEON und des PETA Datensatzes, wird in Bezug auf die Verteilung des Geschlechts im Datensatz eine Verzerrung deutlich (siehe Abbildung 15 und Abbildung 16). Außerdem ist klar ersichtlich, dass Personen über 60 Jahren (personalLarger60) im PETA Datensatz gegenüber der deutschen Bevölkerung stark unterrepräsentiert sind (siehe ebenfalls Abbildung 15 und Abbildung 16).

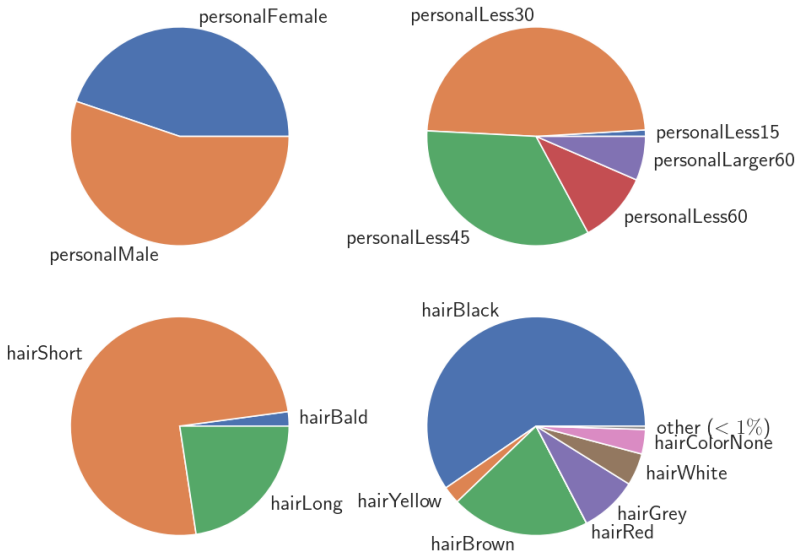

Abbildung 15: Marginale Verteilung der Attribute "sex", "age", "hair length", "hair color" (v.l.o.n.r.u.) entsprechend des PETA Datensatzes.

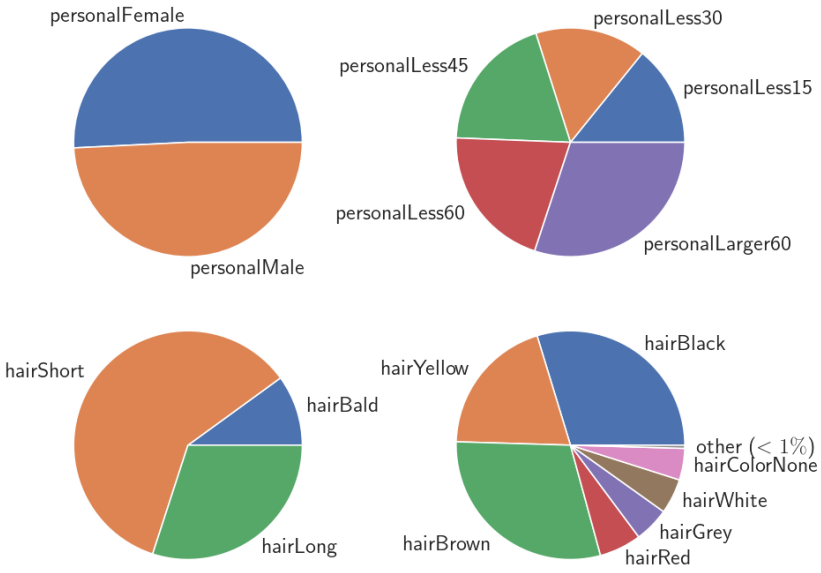

Abbildung 16: Marginale Verteilung der Attribute "sex", "age", "hair length", "hair color" (v.l.o.n.r.u.) entsprechend der probabilistisch erweiterten Ontologie (PEON).

Vergleich der Testauswahlmethoden

Das benutztes Klassifikationsmodel wurde auf der LeNet5 [8] Architektur, mit leichten Anpassungen, erstellt und auf dem PETA Datensatz [7] trainiert. Das Klassifikationsmodell erreichte auf dem Evaluationsdatensatz (20% des PETA Datensatzes, die nicht für das Training des Klassifikationsmodells benutzt wurden), während des Trainings, eine Genauigkeit (Accuracy) von über 0.95 für die Klassen

„age“, „sex“, „hair length“ und „hair color“ des PETA Datensatzes. Dieses Klassifikationsmodell wurde benutzt um folgende Testauswahlmethoden miteinander zu vergleichen:

1. ***Gleichverteilt (Uniform)****:* Die Wahrscheinlichkeit der Auswahl eines jeden Attributs eines Datensatz ist gleich.
2. ***Marginal Distribution data set (Dataset-based (marginal))***: Die Wahrscheinlichkeit der Auswahl eines Attributs im Datensatz entspricht der Häufigkeit des Auftretens des Attributes im Datensatz.
3. ***Conditional Distribution data set (Dataset-based (conditional))***: Die Wahrscheinlichkeit der Auswahl eines Attributs im Datensatz entspricht der Häufigkeit des Auftretens des Attributes im Datensatz unter Berücksichtigung der vorher gezogenen Attribute.
4. ***Marginal Distribution Estimation (Ontology-based (marginal))***: Die Wahrscheinlichkeit der Auswahl eines Attributs im Datensatz entspricht der nach PEON modellierten Auftrittswahrscheinlichkeit des Attributs.
5. ***Conditional Distribution Estimation (Ontology-based (conditional))***: Die Wahrscheinlichkeit der Auswahl eines Attributs im Datensatz entspricht der nach PEON modellierten Auftrittswahrscheinlichkeit des Attributs unter Berücksichtigung der nach PEON modellierten Auftrittswahrscheinlichkeit anderer Attribute.

An Abbildung 17 lässt sich erkennen, dass die Testdatenauswahlmethode einen Einfluss auf die erhaltenen Testergebnisse hat.

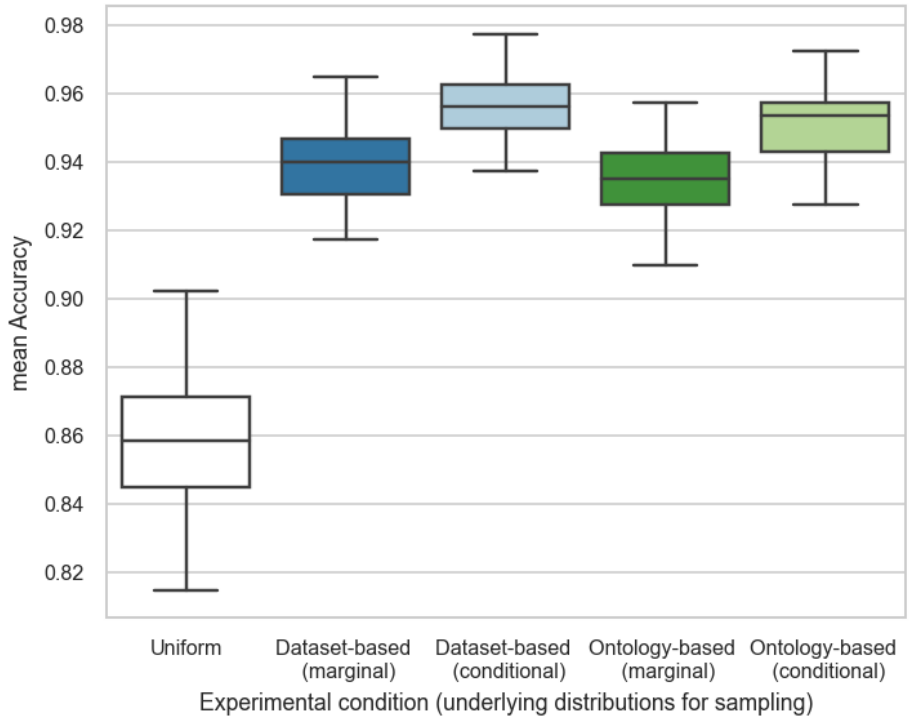

Abbildung 17: Ergebnisse für die im Test gemessene Genauigkeit (mean Accuracy) mit dem gleichen zugrunde liegenden Testdatensatz, aber unterschiedlicher Testdatenauswahl (s.o. Auflistung 1-5).

Dass unterschiedliche Testauswahlverfahren einen Einfluss auf die Testergebnisse haben, kann an Hand des Beispiels N-Wise erklärt werden:

Ein Klassifikationsmodell weißt jeder Äquivalenzklasse, innerhalb einer Partition der Attribute einer Testdatenmenge, im Mittel eine feste Klassifikationsgenauigkeit zu. Im Beispiel in Abbildung 18 weist das Klassifikationsmodell der Äquivalenzklasse "sex": "personalFemale", "hair length": "hairShort", „age“: „personalLess45“ (Index 2 an der Achse „Age“), „hair color“: „black“ (Index 0 der Achse „Hair

Color“) eine Klassifkationsgenauigkeit von etwa 0.5 zu, während die Äquivalenzklasse "sex": "personalFemale", "hair length": "hairShort", „age“: „personalLess15“ (Index 0 an der Achse „Age“), „hair color“: „black“ (Index 0 der Achse „Hair Color“) eine Klassifikationsgenauigkeit von etwa 0.9 zuweist.

Je nachdem in welchem Verhältnis nun Äquivalenzklassen mit besserer oder schlechterer Klassifikationsgenauigkeit in den Testdatensatz eingehen, ergibt sich eine unterschiedliche Genauigkeit des Klassifikationsmodells im Test.

Die Testdatenauswahlmethoden 1-5 wählen, basierend auf ihren Wahrscheinlichkeitsverteilungen, jeweils unterschiedlich viele Testdaten aus einer gegebenen Äquivalenzklasse aus, so produzieren sie die unterschiedlichen Genauigkeitswerte in Abbildung 17.

Ein Beispiel für die Gleichverteilung (1), wäre in Annäherung die N-Wise Testauswahlmethode. Je nach der Wahl von N und Anzahl der vorhandenen Attribute (M) ergeben sich M-N Freiheitsgrade, die je nach implementiertem Algorithmus unterschiedlich aus den vorhandenen Äquivalenzklassen auswählen. Für N=M entspricht N-Wise exakt der Gleichverteilung.

Conditional Distribution data set (3) entspricht der gleichverteilten Auswahl aus dem Trainingsdatensatz, so wie sie im Test der Modellperformanz in der Trainingsphase üblich ist. Hier ist nicht per se sichergestellt, dass die Verteilung der Trainingsdaten der tatsächlichen repräsentativen Verteilung der ODD genügen. Daraus folgt, dass ohne eine Bewertung des Testdatensatzes (und in diesem Fall automatisch des Trainingsdatensatzes) keine Aussage über die Repräsentativität und Eignung des Modells für seine Einsatzumgebung (ODD) gemacht werden kann. Denn es ist nicht transparent in welchem Maß welche Äquivalenzklasse in den Testdatensatz eingeht.

PEON nutzt Conditional Distribution Estimation (5). Wird diese Methode zur Modellierung der ODD basierend auf statistischen Daten, empirischen Erhebungen oder bekannten Wahrscheinlichkeitsverteilungen benutzt, um Abhängigkeiten zwischen Attributen zu erstellen, lässt sich eine, ausgewogene und repräsentative stochastische Referenzverteilung der ODD zu erstellen. In diesem Fall treten die Testfälle innerhalb einer Äquivalenzklasse entsprechend der Betriebsumgebung (ODD) auf.

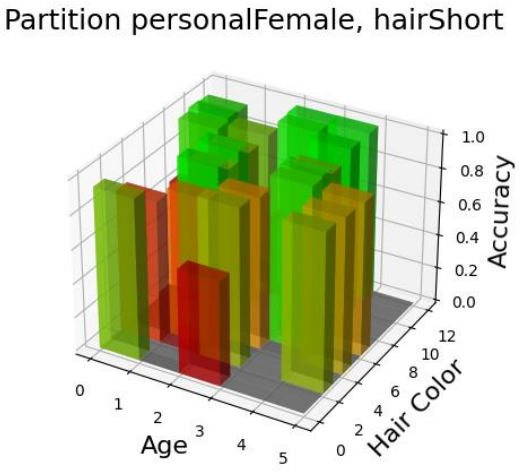

Abbildung 18: Visualisierung der mittleren Genauigkeit (Accuracy) der Äquivalenzklassen innerhalb der Partition zur Testdatenteilmenge "sex": "personalFemale", "hair length": "hairShort" in Abhängigkeit der Attribute „age“ und „hair color“.

Wie oben bereits erwähnt, kann bei der Zulassung sicherheitskritischer KI-Komponenten und -Systeme eine Varianz der Testergebnisse auf Grundlage verschiedener Testauswahlmethoden nicht toleriert werden. Daher bietet die PEON eine Methode, um KI-Komponenten und -Systemtests auf eine einheitliche, statistisch repräsentative und ausgeglichene Grundlage zu stellen, und Varianzen durch Testauswahl entgegenzuwirken. Denn eine repräsentative Vergleichsverteilung ermöglicht Testergebnisse von KI-Komponenten und -Systemen zu bewerten.

Der Technologiereifegrad der PEON-Methode, die über die Auswahl von Testeingaben hinaus durch entwicklungsabhängige Testauswertung auch ein Orakel für den Test von KI-Komponenten liefert, als Teil der KI-Testwerkzeugkette liegt bei TRL 6, was durch die Verwendung im Demonstrator gezeigt wurde.

1.2.2. Evaluation von Explainable AI Methoden

Ein Problem bei der Nutzung von Machine Learning Modellen (ML-Modellen) in sicherheitskritischen Bereichen ist ihre Anfälligkeit gegenüber Manipulation durch Adversarial Examples. Adversarial Examples stellen Datenobjekte dar, die von einem ML-Modell fehlklassifiziert werden und der größeren Klasse der Out-Of-Distribution (OOD) Daten zugeordnet werden können. Der Begriff OOD leitet sich dabei aus der Sichtweise ab, dass ML-Modelle Wahrscheinlichkeitsverteilungen lernen, die ihren Trainingsdaten reproduzieren.

Im Rahmen von Hauptarbeitspaket 2 wurde eine Recherche zu aktuellen Tools und Methoden zu Erklärbarkeit (Explainability) von ML-Modellen durchgeführt. Neben bekannten Tools wie LIME (Local Interpretable Model-agnostic Explanations) oder SHAP (SHapley Additive exPlanations) und Blackbox Methoden, die über Änderungen an den Inputdaten den (approximierten) Einfluss dieser auf den Output darstellen, wurden Methoden zur Erkennung von OOD Daten betrachtet.

Auf Grundlage der Maximum Mean Discrepancy Methode [9], ein Spezialfall des Mahalanobis Abstands, der erfolgreich als Metrik im Training von generativen Modellen und zur Outlier/ Novelty Detection eingesetzt wird [10][11], wurde ein Erweiterter Ansatz [12], der durch ein ML Modell erlernte, semantische Informationen in die Bewertung der Genauigkeit der Vorhersage (Accuracy) mit einschließt, evaluiert und als Proof-of-Concept implementiert.

Ein State-of-the-Art Ansatz zur Erkennung von OOD Daten [13], der auf den oben genannten Methoden aufbaut und sehr hohe AUROC-Scores (96,4 bis 100,0) bei der Erkennung von OOD Daten erzielt, wurde evaluiert und in eine Proof-of-Concept Implementierung für ML-Modelle als Greybox Ansatz (Zugriff auf Layer der ML-Modelle und Trainingsdaten werden benötigt) umgesetzt.

Prototyp einer Out-of-Distribution (OOD) Analysemethode

Aufbauend auf der Proof-of-Concept Implementierung der OOD-Methoden aus [12] und [13], wurden diese Methoden in einem Prototyp auf ein Analyseverfahren auf Grundlage einer mehrdimensionalen Gaußverteilung erweitert, um die interne Struktur von ML-Modellen auf wiederkehrende Zusammenhänge zu untersuchen. Dabei wird ein modularer Ansatz verfolgt, um bei Bedarf die Analysemethode zu erweitern oder auf leistungsfähigere Methoden zu wechseln.

Im Prototyp soll mit Zugriff auf die internen Schichten des Neuronalen Netzes und deren Ausgaben neben der Entscheidung folgendes untersucht werden:

- Stammt ein Datenpunkt eines Datensatzes aus der vom ML-Modell gelernten Verteilung oder nicht?
- Ist ein globaler (nicht Datenpunkt abhängiger) Zusammenhang zwischen Fehlklassifikationen des untersuchten ML-Modells und dessen internen Aktivierungen ableitbar?

Bei stichpunktartigen Untersuchungen der Aktivierungen interner Layer eines Bildklassifizierungsmodels wurde festgestellt, dass durch adversariale Attacken fehlklassifizierte Bilder sich im ML-Modell durch ein stark von der Norm abweichendes Bild der Aktivierungen zeigen (siehe Abbildung 19 und Abbildung 20). Dieser Umstand soll zugänglich und prüfbar gemacht werden, in dem die angesprochene Analysemethode durch ein Training auf den Aktivierungen des zu untersuchenden ML-Modells während der Inferenz trainiert werden und über Varianzen und Kovarianzen die Abweichungen zu den natürlichen Eingaben (Bilder aus der von dem ML-Modell gelernten Verteilung/ ODD) messbar gemacht werden.

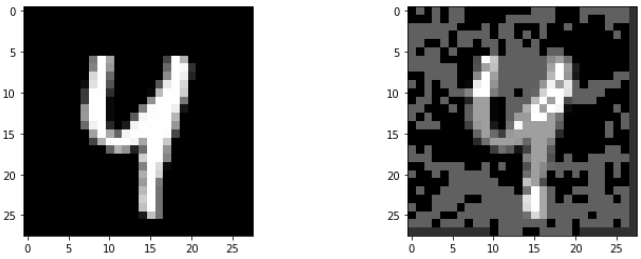

Abbildung 19: Links: ungestörte Eingabedaten der Klasse 4, klassifiziert als Klasse 4. Rechts: gestörte Eingabedaten der Klasse 4, klassifiziert als Klasse 9.

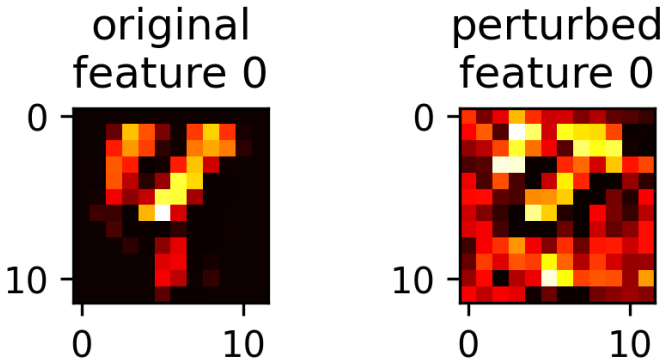

Abbildung 20: Repräsentation eines Features in einem Hidden Layer. Links: ungestörte Eingabedaten der Klasse 4, klassifiziert als Klasse 4. Rechts: gestörte Eingabedaten der Klasse 4, klassifiziert als Klasse 9.

Bei der Implementierung der Out-of_Distribution (OOD) Erkennung mittels mehrdimensionaler Gaußverteilung (multi variate Normal - MVN) hat sich folgendes Fehlerbild ergeben:

Für einfache Beispiele mit niedrig dimensionalen Eingabedaten, die wenig Rechenzeit benötigen, werden zuverlässig Mittelwert und Kovarianzmatrix der MVN gefittet, mit denen in einem zweiten Schritt die Entscheidung getroffen wird, ob ein Datenpunkt innerhalb der von einer KI gelernten Wahrscheinlichkeitsverteilung (in distribution) oder außerhalb (out of distribution) liegt. Werden hingegen höher dimensionale Eingabedaten eines Test-ML-Modells als Eingabe genutzt, versagt das Fitting der MVN und es werden scheinbar zufällige Mittelwerte und Kovarianzmatrix Elemente erzeugt, die keine Aussage darüber zulassen, ob ein Datenunkt in distribution oder out of distribution liegt.

Bei der Fehleranalyse konnten zwei Fehler identifiziert werden. Neben einer inkorrekt angewendeten Metrik zum Fitting der mehrdimensionaler Gaußverteilung (Multi Variate Normal - MVN) wurde auch der Vergleich verschiedener Features innerhalb eines Layers und Layer übergreifend nicht korrekt in der OOD Methode berücksichtigt.

Der Fehler in der Metrik wurde angepasst und behoben, so dass das Fitting funktioniert.

Der Vergleich verschiedener Features innerhalb eines Layers und Layer übergreifend hingegen konnte jedoch nicht korrigiert werden.

Darüber hinaus besteht das Fehlerbild (scheinbar zufällige Mittelwerte und Kovarianzmatrix Elemente) weiter und es wird davon ausgegangen, dass noch weitere unentdeckte Fehler vorhanden sind. Da der abgeschätzte Aufwand zur Erreichung eines fund8ierten Ergebnisses den Projektrahmen sprengte, wurde die Arbeit an der Methode nicht abgeschlossen.

1.3. Hauptarbeitspaket 3 (HAP3): Modellbasiertes Testen von KI-basierten Bahntechnikkomponenten

1.3.1. Ziele und Strategien zum automatisierten Testen von KI-basierten Objekterkennungssystemen in der Bahntechnik

Was muss alles zu einer Zertifizierung eines KI-Systems nachgewiesen werden? Welche Ziele sind zu erreichen und wie kommen wir dahin?

In einem gemeinsamen Dokument *Objectives and strategies for automated testing of AI-based perception systems in railroad engineering* wurden von den verschiedenen Projektpartnern dazu Ziele und Strategien aus ihren Kompetenzbereichen abgeleitet und hier festgehalten. Von ITPower Solutions wurde zu den folgenden Themen Beiträge erstellt:

- *safeguarding robustness (Absicherung der Robustheit)*
- *Safeguarding the quality of the data (Absicherung der Datenqualität)*
- *Testing for dependence on spurious correlations (Test auf Abhängigkeit von Scheinkorrelationen)*
- *Automated identification of test oracles (ground truth) (Automatisierte Identifikation von Testorakeln)*

Diese Arbeit wurde vom Projektpartner Fraunhofer FOKUS bei der ETSI eingereicht.

1.3.2. Konzeption einer Testwerkzeugkette für KI

Gemeinsam mit Fraunhofer FOKUS wurde an einem Konzept für eine Testwerkzeugkette für KI gearbeitet. Dabei standen die Fragen des Testorakels und der Erstellung von Testdaten mit ausreichend Varianz in Szenarien, Umweltbedingungen und potenziellen Störungen im Vordergrund.

Mit der probabilistisch erweiterten Ontologie (PEON) konnte ein systematischer Ansatz für die Erstellung von abstrakten Testfällen als Startpunkt in der Form eines (statistischen) Referenzmodells für die Testwerkzeugkette (Abbildung 12) identifiziert werden. Da die abstrakten Testfälle sowohl eine Beschreibung der Szenen und den darin enthaltenen Objekten, sowie deren Auftrittswahrscheinlichkeit enthalten, liefert die PEON ebenfalls ein Testorakel für die Auswertung der Tests. Die abstrakten Testfälle dienen dann in der Testwerkzeugkette als Eingabedaten für die Erstellung (fotorealistischer) Testdaten durch die 3D-Simulation von Fraunhofer FOKUS.

Im Rahmen von Hauptarbeitspaket 3 wurde PEON für die Nutzung in der Testwerkzeugkette angepasst. Mehr Informationen dazu in Hauptarbeitspaket 4 (HAP4): Absicherungsmethodik und Werkzeugkette.

1.4. Hauptarbeitspaket 4 (HAP4): Absicherungsmethodik und Werkzeugkette

1.4.1. Implementierung der Testwerkzeugkette für KI

Die durch die Projektpartner im Projekt erarbeitete Methodik wurde als Teil des Safety Evaluation Process (Abbildung 10) in einer Testwerkzeugkette zusammengefasst und in der Form eines Demonstrators implementiert.

Einen wichtigen Aspekt der Umsetzung des Demonstrators stellen die Schnittstellen zwischen den Softwarekomponenten der Projektpartner dar. Um die Schnittstellen zu identifizieren, wurde aus der bereits beschriebenen Testwerkzeugkette (Abbildung 12) ein schematisches Ablaufdiagramm erstellt (Abbildung 21), das die Schnittstellen klar darstellt.

Für ITPower Solutions ergaben sich Schnittstellen zwischen der Probabilistisch Erweiterten ONtologie (PEON) und der 3D Simulation von Fraunhofer FOKUS (1 in Abbildung 21), aber auch zur Augmentierung von Störungen von neurocat (2 in Abbildung 21). An der Implementierung der anderen Schnittstellen (3-5 in Abbildung 21) war ITPS als Koordinator beteiligt.

Schnittstelle 1 stellt den Austausch der erzeugten abstrakten Testfallbeschreibungen aus der Probabilistisch Erweiterten ONtologie (PEON), die dann in der Testautomatisierungskette durch die 3D Simulation von Fraunhofer FOKUS in Szenarien umgewandelt werden, um daraus konkrete Testfälle (Bilder) zu erstellen, dar. In der abstrakten Testfallbeschreibung werden Störungen (z.B. Regen oder Nebel) definiert, die entweder in 3D simuliert oder nachträglich als Augmentierung auf vorhandene Bilder realisiert werden können. Da beide Varianten möglich sind, enthält diese Schnittstelle einen Parameter zur Auswahl, ob eine Störung in der 3D Simulation oder als Augmentierung umgesetzt werden soll.

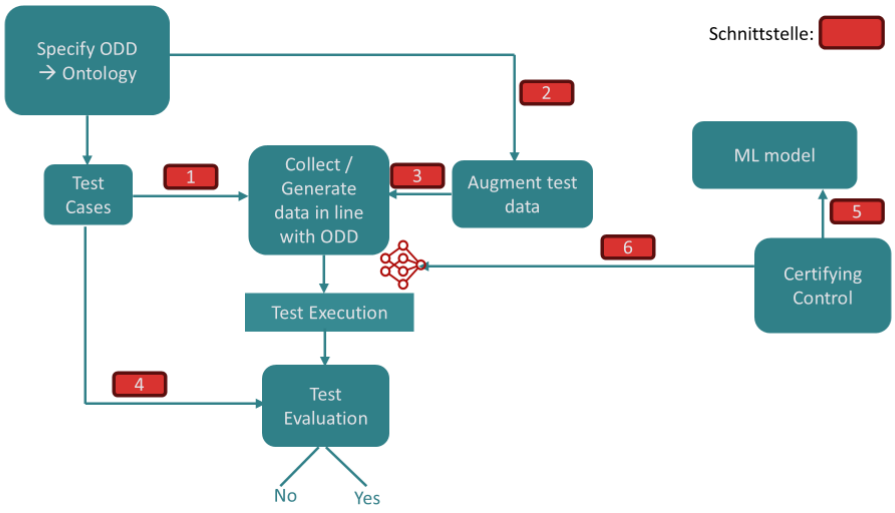

Abbildung 21: Schematischer Testablauf der Test-Toolchain

Tool zur Erstellung von Ontologien

Zur Erstellung von Ontologien wurde die freie Software Freeplane genutzt. Bei Freeplane handelt es sich um ein Tool zur Erstellung von Mindmaps in Form von Knoten und Verbindungen dieser durch gerichtete und ungerichtete Kanten. Sowohl den Knoten, als auch den Kanten können benutzerdefinierte Attribute hinzugefügt werden. Diese Attribute wurden bei der Erstellung von probabilistisch erweiterten Ontologien (PEON) genutzt, um stochastische Daten, wie diskretisierte Wahrscheinlichkeitsverteilungen und Abhängigkeiten (siehe Attribut „Marginal_Distribution“, „Dependancies“ auf der rechten Seite in Abbildung 22) an die Knoten anzufügen, sowie um

Abbildungsvorschriften zwischen den Werten verschiedener Wahrscheinlichkeitsverteilungen zu definieren (siehe „Text am Anfang“ auf der rechten Seite von Abbildung 23).

Freeplane liegt eine XML-Datenstruktur zugrunde, in der die Knoten, Kanten und Attribute gespeichert werden. Um abstrakte Testfälle aus der Ontologie zu erhalten, wurde ein Tool entwickelt, dass die XML-Datei einer Ontologie nutzt, die relevanten Werte und Größen parst und auf Grundlage der Wahrscheinlichkeitsverteilungen, Abhängigkeiten und Abbildungen innerhalb der Ontologie, automatisiert abstrakte Testfälle generiert. Die Abstrakten Testfälle werden in einem mit den Projektpartnern abgestimmten JSON-Format (siehe Abbildung 24) gespeichert und können so in der Testwerkzeugkette weiterverwendet werden.

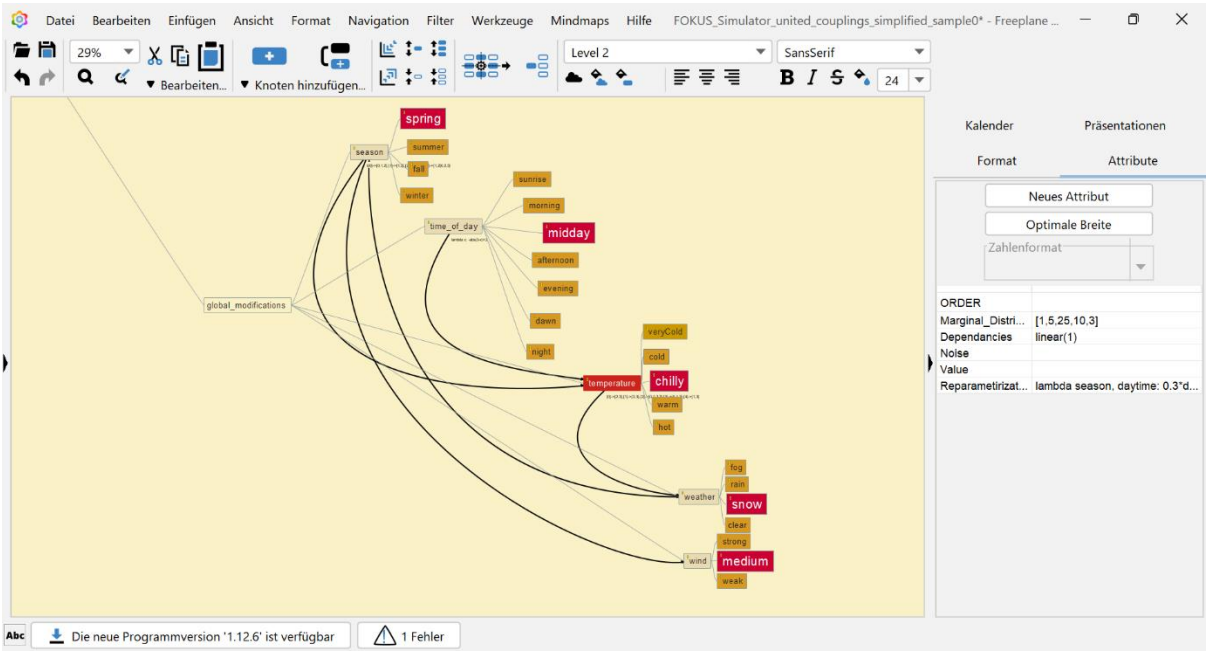

Abbildung 22: Oberfläche der Open Source Software Freeplane, in der der Ausschnitt „global_modifications“ (Einflüsse auf Szenarien durch Wetter, Jahres- und Tageszeiten) aus einer größeren Ontologie des Bahnbereichs zu sehen ist.

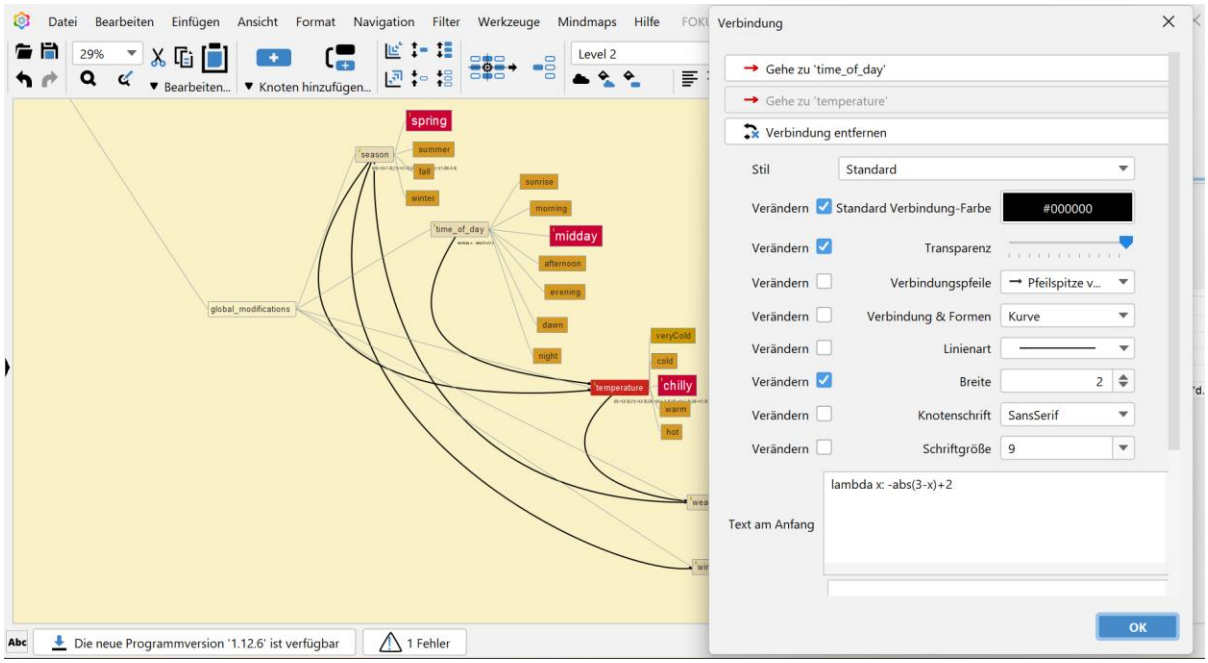

Abbildung 23: gerichtete Kante (Verbindung) zwischen den Knoten "time_of_day" und "temperature", in der im Feld "Text am Anfang" eine funktionale Abbildungsvorschrift der Werte von "time_of_day" auf "temperature" definiert ist.


```
{
  "samples": [
    {
      "global_modifications_source": "Fokus",
      "concept_values": {
        "temperature": "cold",
        "cloud_coverage": "medium",
        "fog": "weak",
        "rain": "NoRain",
        "snow": "medium",
        "dust": "None",
        "wind": "weak",
        "season": "fall",
        "time_of_day": "midday"
      },
      "global_modifications": {
        "time_of_day": 12.0,
        "cloud_coverage": 4.2652,
        "rain": 0.0,
        "snow": 4.0571,
        "lightning": 0,
        "wind": 2.4053,
        "fog": 2.0834,
        "dust": 0.0,
        "temperature": 0.8958,
        "season": "fall",
        "average_drop_size": 0.216,
        "drop_density": 2115.9155,
        "precipitation_rate": 55.3269,
        "particle_size": 0.0021,
        "fall_speed": 0.0094,
        "intensity_attenuation_coefficient": 0.0059,
        "atmospheric_light_factor": 0.0274,
        "fog_intensity": 1.2096
      },
```


Abbildung 24: Spezifikation der "global_modifications" eines abstrakten Testfalls, der automatisch aus obiger PEON generierten wurde.

Implementierung der Testwerkzeugkette als Demonstrator in Rerun

Rerun ist eine freie Software zur interaktiven Visualisierung verschiedener multimodaler Datenobjekte (z.B. Bilder, Videos, geometrische Objekte in 2D und 3D, Text, mathematische Funktionen, …), die es erlaubt das dynamische Verhalten der Objekte aufzuzeichnen und als interaktive Aufzeichnung wieder

abzuspielen. Rerun kann sowohl über die Kommandozeile, als auch mit einer GUI genutzt werden. Für den Demonstrator wurde Rerun mit der GUI als Viewer verwendet (siehe Abbildung 1 - Abbildung 4).

Rerun wurde über die vorhandene Python-Schnittstelle zur Aufzeichnung von Testszenarien konfiguriert. Die Konfiguration wurde vor der Aufzeichnung der Testszenarien zwischen den Projektpartnern abgestimmt, um ein Ineinandergreifen der einzelnen Komponenten, d.h.

- PEON-Methode von ITPower Solutions,
- 3D Simulation des Projektpartners Fraunhofer FOKUS,
- KI-Perzeptionssystem des Projektpartners Hitachi,
- Augmentierung von Bildern des Projektpartners neurocat und
- Certifying Control und Segmentierungsauswertung des Projektpartners HHU,

sicherzustellen.

Als Ergebnis wurden verschiedene Wetterbedingungen in einem Testszenario umgesetzt und als Rerun Aufzeichnung für Präsentationen vorbereitet (siehe KI-LOK Demonstrator im Abschnitt 1.1.2).

1.5. Hauptarbeitspaket 5 (HAP5): Übergreifende Aktivitäten

Projektbegleitend wurden mehrere übergreifende Aktivitäten durchgeführt.

1.5.1. Koordination und Projektleitung

ITPower hat die Koordination der Arbeitspakete übernommen, den Fortschritt des Projektverlaufs geleitet und dafür Sorge getragen, dass das Projekt trotz Herausforderungen durch zeitweise aufgetretenen Kapazitätsengpässen, Umfirmierung und Insolvenz eines Projektpartners am Ende des Projekts, erfolgreich zu Ende gebracht werden konnte.

Weiter fanden regelmäßig alle zwei Wochen Projekttreffen aller Projektpartner unter Leitung und Koordination von ITPower statt. Der Schwerpunkt wurde dabei wechselnd auf die verschiedenen Hauptarbeitspakete gelegt, deren Inhalt von den Hauptarbeitspaketleitern bestimmt wurde.

In den Wochen in denen keine Projekttreffen stattfanden, wurden zuerst regelmäßige, zweiwöchige Kolloquien zu Themen des Projekts (z.B. detaillierte Vorstellung aktueller Arbeiten oder Vorstellung von relevanter Literatur) veranstaltet, zu denen auch Teilnehmer außerhalb des Projektes (z.B. Studenten von Universitäten und Instituten, Forscher mit verwandten Themengebieten) eingeladen wurden.

Zum Ende des Projekts wurden die Projekttreffen und Kolloquien in regelmäßige wöchentliche Treffen zur Abstimmung und Koordination der Implementierung des Demonstrators unter Leitung von ITPower umgewandelt.

1.5.2. Projekttreffen im Rahmen des Förderprogramms „Neue Fahrzeug- und Systemtechnologien“

Im Laufe des Projekts wurde aktiv der Kontakt mit anderen Projekten des Förderprogramms „Neue Fahrzeug- und Systemtechnologien“ gesucht. Als Ergebnis nahm KI-LOK am Workshop „Automated Train Operation - Aktueller Stand und Ausblick in einem industriellen Schlüsselbereicht“ am 22.02.2024 teil, stellte seine aktuellen Ergebnisse dar und beteiligte sich aktiv an der Identifizierung aktueller Herausforderungen bei der Automation des Bahnbetriebs und der Konzeption von Lösungsansätzen.

Ferner wurde ein regelmäßiger Austausch mit dem Deutschen Zentrum für Schienenverkehrsforschung (DZSF), das sich bereits in der Antragsphase des Projekts als

Projektbegleiter bereit erklärt hatte, gepflegt. Der Austausch zwischen ITPower Solutions und DZSF wird voraussichtlich auch über die Projektlaufzeit hinaus aufrechterhalten wird.

1.5.3. Transfer in Kundenprojekte

Die Ergebnisse des KI-LOK-Projekts, vor allem die PEON-Methode zur Ermittlung repräsentativer und ausgeglichener Testeingaben und automatisierten Testauswertung (Orakel), hat Einzug in das Dienstleistungsportfolio von ITPower genommen. Zurzeit wird einerseits die Umsetzung des Prototyps in Produkt konzipiert und andererseits intensive Akquise von Dienstleistungsprojekten betrieben.

1.5.4. Veröffentlichungen

Im Rahmen des KI-LOK Projekts wurden die Arbeiten und Ergebnisse von ITPower durch verschiedene Publikationen, die meisten davon in Zusammenarbeit mit den Projektpartnern, veröffentlicht. Die Liste der Veröffentlichungen findet sich in Abschnitt 6 (Erfolgte oder geplante Veröffentlichungen der Ergebnisse).

2. Wichtigste Positionen des zahlenmäßigen Nachweises

Die wichtigsten Positionen des zahlungsmäßigen Nachweises sind die Personalkosten, welche durch die Zahlung von Gehältern an die Projektmitarbeiter während der Projektlaufzeit entstanden.

3. Notwendigkeit und Angemessenheit der geleisteten Arbeiten

Die Qualitätssicherung von KI-basierten Komponenten im Eisenbahnbetrieb durch analytische Methoden ist technisch und wissenschaftlich innovativ und hochkomplex. Alle geleisteten Arbeiten zur Durchführung der Arbeitspakete und zum Erzielen der entsprechenden Ergebnisse, die zur Entwicklung von innovativen Methoden und ihre Umsetzung in Werkzeuge dienten, wie in diesem Bericht dargestellt, waren notwendig und angemessen.

4. Voraussichtlicher Nutzen und Verwertbarkeit des Ergebnisses

Die durch ITPower Solutions erarbeiteten Projektergebnisse füllen eine methodische Lücke im Bereich des Tests von KI-Komponenten auf und eignen sich deshalb hervorragend für die Verwertung im Bahnbereich, aber auch in anderen Industriebranchen, die klassifizierende KI-Komponenten entwickeln und einsetzen. Ferner verbirgt die Kombination der von allen Projektpartnern aufeinander abgestimmten entwickelten Methoden und Werkzeuge ein noch größeres Potential an Verwertbarkeit.

ITPower Solutions startete bereits in der Projektlaufzeit mit der Verwertung der Projektergebnisse durch Gründung und Etablierung eines neuen Geschäftsfelds zur KI. Hier steht die im KI-LOK-Projekt entwickelte systematische Testmethode auf der Basis der probabilistisch erweiterten Ontologie (PEON) im Zentrum. Auch die im Rahmen des Projekts entwickelte Methode der metamorphen Bildtransformationen zur Erweiterung von Trainings- und Testdaten sowie der konzeptionellen Simulation liegen im Dienstleistungsportfolio des KI-Geschäftsfeld. Es ist geplant, durch zukünftige F&E-Aktivitäten die vorhandene Umsetzung der Methoden in Werkzeuge weiterzuentwickeln, so dass auch die Werkzeuge als eigenständige Produkte auf dem Markt angeboten werden können.

Nach dem Projetabschluss fanden Gespräche mit dem Projektpartner Fraunhofer FOKUS zum gemeinsamen Angebot der von den beiden Partnern entwickelten Methoden und Tools statt. Es wurde vereinbart, die bereits aufeinander abgestimmte Generierung von Testfällen gemäß der PEON-Methode (von ITPower) und Generierung von Fotorealistischen Simulationen aus diesen Testfällen (von FOKUS) gemeinsam auf dem Markt anzubieten, den Bedürfnissen der Kunden anzupassen und weiterzuentwickeln. Die Vorbereitungen dieser Vermarktungsstrategie laufen aktuell auf Hochtouren.

5. Während der Durchführung des Vorhabens bekannt gewordener Fortschritt bei anderen Stellen

Während der Projektlaufzeit und nach dem Projektabschluss wurden keine relevanten neuen Forschungsergebnisse im Bereich KI-Systeme für die Bahn von dritter Seite bekannt, welche die Forschungsarbeit im KI-LOK Projekt vorwegnehmen oder überflüssig machen.

Inhaltlich verwandte Forschungsarbeiten zu denen des KI-LOK-Projektes wurden insbesondere in den Projekten des Förderprogramms „Neue Fahrzeug- und Systemtechnologien“, in dem auch KI-LOK gefördert wurde, durchgeführt. Während sich die meisten Projekte dieses Programms, die einen Bezug zur Bahntechnik haben, mit den konstruktiven Aspekten des automatisierten Fahrens und der darin eingesetzten KI beschäftigen und sich dadurch von KI-LOK abgrenzten, widmet sich das Projekt safe.trAIn mit dem Prozess des Safety-Engineerings von Perzeptionssystemen im Bahnbereich und berücksichtigt dabei auch den Test dieser Systeme. Allerdings wurde der Test in safe.trAIn auf einer prozessualen Ebene betrachtet, während KI-LOK den Test von KI-Komponenten auf einer methodischen Ebene erforschte. Im Rahmen des von TÜV Rheinland organisierten ATO-Workshops, der am 21.02.2024 in Köln stattfand, fand ein reger Austausch zwischen den beiden Projekten statt. Es wurde unter anderem festgestellt, dass die jeweils prozessualen und methodischen Ansätze der Projekte KI-LOK und safe.trAIn im Bereich des Tests von KI-Komponenten sich gegenseitig gut ergänzen.

Ferner markiert der KI-Standard DIN SPEC 91516 „Menschliche Leistungsfähigkeit bezüglich der dynamischen Fahraufgabe zur Spezifikation von KI für ATO“, der in naher Zukunft veröffentlicht wird, einen Meilenstein im Bereich der Qualitätssicherung und Zertifizierung des automatisierten Fahrens im Bahnbereich. Durch diesen Standard können Referenzwerte für die Qualität von KI-Komponenten definiert werden. Diese tragen zur Bestimmung von Testendekriterien bei, welche in den im KI-LOK-Projekt entwickelten Testmethode PEON verwendet werden können. Es bestand ein kontinuierlicher Austausch zwischen dem KI-LOK-Projekt und dem Deutschen Zentrum für Schienenverkehrsforschung (DZSF), das an der Entwicklung des o.g. Standards maßgeblich beteiligt war und gelichzeitig das KI-LOK-Projekt begleitete. Dieser Austausch wird nach dem Projektabschluss fortgeführt.

6. Erfolgte oder geplante Veröffentlichungen der Ergebnisse

H.-W. Wiesbrock, J. Grossmann
Probabilistically Extended Ontologies: A Basis for Systematic Testing of ML-Based Systems
SAE Technical Paper 2024-01-3002, 2024, https://doi.org/10.4271/2024-01-3002

H.-W. Wiesbrock, J. Grossmann
Outline of an Independent Systematic Blackbox Test for ML Systems
The 6th IEEE International Conference on Artificial Intelligence Testing (IEEE AITEST 2024), 07/2024, Shanghai, China
https://doi.org/10.1109/AITest62860.2024.00009

N. Grube, M. Massah, M. Tebbe, P. Wancura, H.-W. Wiesbrock, J. Grossmann, S. Kharma,
On a systematic test of ML-based systems: Experiments on test statistics
The 6th IEEE International Conference on Artificial Intelligence Testing (IEEE AITEST 2024), 07/2024, Shanghai, China
https://doi.org/10.1109/AITest62860.2024.00010

H.-W. Wiesbrock, S. Sadeghipour
Wie kann KI systematisch getestet werden? Ein methodischer Blackbox Test für KI-Systeme
6. ASQF Net Week, 06/2024

R. Krajewski, S. Stritz
Künstliche Bilder für künstliche Intelligenz – Testdatengenerierung mittels metamorpher Bildtransformationen
Elektronik Automotive, 02/2024; Abrufbar als E-Paper oder unter https://www.elektroniknet.de

G. Hemzal, T. Strobel, J. Großmann, M. Leuschel, D. Knoblauch, M. Kucheiko, N. Grube, R. Krajewski: **KI-LOK – Ein Verbundprojekt über Prüfverfahren für KI-basierte Komponenten im Eisenbahnbetrieb**
Signal + Draht 04/ 2023

Grossmann, J. et al.
Test and Training Data Generation for Object Recognition in the Railway Domain.
In: Masci, P., Bernardeschi, C., Graziani, P., Koddenbrock, M., Palmieri, M. (eds) Software Engineering and Formal Methods.
SEFM 2022 Collocated Workshops. SEFM 2022. Lecture Notes in Computer Science, vol 13765. Springer, Cham. https://doi.org/10.1007/978-3-031-26236-4_1

J. Großmann, N. Grube, S. Kharma, D. Knoblauch, R. Krajewski, M. Kucheiko, H.-W. Wiesbrock
Test- und Trainingsdatengenerierung für die Objekterkennung im Bahnbereich
AI4EA 2022

G. Hemzal, T. Strobel, J. Großmann, B.-H. Schlingloff, M. Leuschel, S. Sadeghipour, J. Firnkorn
KI-LOK – Ein Verbundprojekt über Prüfverfahren für KI-basierte Komponenten im Eisenbahnbetrieb
Signal + Draht 10/ 2021

H.-W. Wiesbrock, N. Grube,
Testing ML-based Systems Using Probabilistically Extended Ontologies
Vortrag bei AI4EA 2024 Workshop + Veröffentlichung in Springer LNCS

H.-W. Wiesbrock, S. Sadeghipour,
Testverfahren für ML-basierte Systeme - Konsistentes und vollständiges Testen von Feedforward-Netzen
Vortrag bei Embedded Software Engineering Kongress im Dezember 2024

Referenzen

[1] J. Roßbach, O. De Candido, A. Hammam, und M. Leuschel, „Evaluating AI-Based Components in Autonomous Railway Systems: A Methodology“, in *KI 2024: Advances in Artificial Intelligence*, Bd. 14992, A. Hotho und S. Rudolph, Hrsg., in Lecture Notes in Computer Science, vol. 14992. , Cham: Springer Nature Switzerland, 2024, S. 190–203. doi: 10.1007/978-3-031-70893-0_14.

[2] H.-W. Wiesbrock und J. Grossmann, „Outline of an Independent Systematic Blackbox Test for ML Systems“, November 2023. [Online]. Verfügbar unter: https://arxiv.org/abs/2401.17062

[3] Statistisches\ Bundesamt\ Deutschland, „Bevölkerung in Deutschland (Internetausgabe)“. Wiesbaden, 2023. [Online]. Verfügbar unter: https://service.destatis.de/bevoelkerungspyramide/index.html

[4] K. Duan, S. Bai, L. Xie, H. Qi, Q. Huang, und Q. Tian, „CenterNet: Keypoint Triplets for Object Detection“, in *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, 2019, S. 6568–6577. doi: 10.1109/ICCV.2019.00667.

[5] T.-Y. Lin *u. a.*, „Microsoft COCO: Common Objects in Context“, in *Computer Vision – ECCV 2014*, D. Fleet, T. Pajdla, B. Schiele, und T. Tuytelaars, Hrsg., Cham: Springer International Publishing, 2014, S. 740–755.

[6] N. Grube *u. a.*, „On a Systematic Test of ML-Based Systems: Experiments on Test Statistics“, in *2024 IEEE International Conference on Artificial Intelligence Testing (AITest)*, Shanghai, China: IEEE, Juli 2024, S. 11–20. doi: 10.1109/AITest62860.2024.00010.

[7] Y. DENG, P. Luo, C. C. Loy, und X. Tang, „Pedestrian Attribute Recognition At Far Distance“, in *Proceedings of the 22nd ACM International Conference on Multimedia*, in MM ’14. New York, NY, USA: Association for Computing Machinery, 2014, S. 789–792. doi: 10.1145/2647868.2654966.

[8] Y. Lecun, L. Bottou, Y. Bengio, und P. Haffner, „Gradient-based learning applied to document recognition“, *Proceedings of the IEEE*, Bd. 86, Nr. 11, S. 2278–2324, 1998, doi: 10.1109/5.726791.

[9] A. Gretton, K. M. Borgwardt, M. J. Rasch, B. Schölkopf, und A. Smola, „A kernel two-sample test“, *J. Mach. Learn. Res.*, Bd. 13, Nr. null, S. 723–773, März 2012.

[10] T. Denouden, R. Salay, K. Czarnecki, V. Abdelzad, B. Phan, und S. Vernekar, „Improving Reconstruction Autoencoder Out-of-distribution Detection with Mahalanobis Distance“, 6. Dezember 2018, *arXiv*: arXiv:1812.02765. doi: 10.48550/arXiv.1812.02765.

[11] Y. Hou, Z. Chen, M. Wu, C.-S. Foo, X. Li, und R. M. Shubair, „Mahalanobis Distance Based Adversarial Network for Anomaly Detection“, in *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Mai 2020, S. 3192–3196. doi: 10.1109/ICASSP40776.2020.9053206.

[12] R. Gao *u. a.*, „Maximum Mean Discrepancy Test is Aware of Adversarial Attacks“, in *Proceedings of the 38th International Conference on Machine Learning*, PMLR, Juli 2021, S. 3564–3575. Zugegriffen: 15. Februar 2022. [Online]. Verfügbar unter: https://proceedings.mlr.press/v139/gao21b.html

[13] H. Okamoto, M. Suzuki, und Y. Matsuo, „Out-of-Distribution Detection Using Layerwise Uncertainty in Deep Neural Networks“, Dez. 2019, Zugegriffen: 1. November 2022. [Online]. Verfügbar unter: https://openreview.net/forum?id=rklVOnNtwH