Finanziert von der Europäischen Union
NextGenerationEU

Gefördert durch:
Bundesministerium für Wirtschaft und Klimaschutz
aufgrund eines Beschlusses des Deutschen Bundestages

Schlussbericht

Projekt: „safe.trAIn
– Sichere KI am Beispiel fahrerloser Regionalzug“

Teilvorhaben: „Nachweis der Vertrauenswürdigkeit von KI-Methoden für den autonomen Schienenverkehr“

Förderkennzeichen: 19I21039L

Zuwendungsempfänger: *Fraunhofer-Gesellschaft zur Förderung der angewandten Forschung e.V. (im Folgenden „Fraunhofer“)* mit den beteiligten Instituten: *Fraunhofer-Institut für Kognitive Systeme IKS* und *Fraunhofer-Institut für Intelligente Analyse- und Informationssysteme IAIS*

Berichtszeitraum: 01.01.2022-31.03.2025

Autoren: Gereon Weiß, Christian Drabek, Sujan Sai Gannamaneni, Iwo Kurzidem, Andreas Kreutz, Michael Mock, Poulami Sinhamahapatra, Hoai My Van

Inhalt

I. Kurzdarstellung

I.1 Aufgabenstellung

Das Projekt safe.trAIn verfolgte das Ziel, die sichere Anwendung Künstlicher Intelligenz (KI) im fahrerlosen Schienenverkehr zu erforschen. Im Fokus stand die Entwicklung eines GoA4-Systems (vollautomatisierter Zugbetrieb ohne Personal an Bord) für Regionalzüge, das auf KI-gestützten Perzeptionsfunktionen basiert. Klassische Automatisierungstechnologien reichen für diesen Anwendungsfall nicht aus – insbesondere die sichere Hinderniserkennung in offenen Umgebungen stellt eine zentrale Herausforderung dar. Das Projekt hatte zum Ziel, diese Herausforderung hinsichtlich verschiedener Aspekte zu verfolgen. Eine Herausforderung war die Entwicklung von Prüfmethoden und Werkzeugen zur Bewertung der Vertrauenswürdigkeit von KI anhand des Anwendungsfalls. Hierfür wurde auch die Spezifikation einer passenden Sicherheitsarchitektur für KI-basierte Funktionen im Zugbetrieb verfolgt. Für die Absicherung von KI-Funktionen wurde die Konzeption und Validierung eines virtuellen Testfelds beabsichtigt. Darüber hinaus sollten die Ergebnisse sowohl hinsichtlich Standardisierung als auch Übertragung auf verwandte Anwendungsdomänen berücksichtigt werden. Die Projektlaufzeit erstreckt sich von Januar 2022 bis März 2025. Beteiligt waren Partner aus Industrie, Forschung, Normung und Prüfung.

Die Fraunhofer-, vertreten in safe.trAIn durch die Institute Fraunhofer IKS und Fraunhofer IAIS, übernahm im Projekt eine zentrale Rolle bei der methodischen Absicherung von KI-Funktionen. Im Teilvorhaben „Nachweis der Vertrauenswürdigkeit von KI-Methoden für den autonomen Schienenverkehr" verfolgte Fraunhofer verschiedene Aufgabenschwerpunkte. Hierzu gehört die Entwicklung eines systematischen Nachweiskonzepts (Safety Case) für KI-Funktionen und quantifizierbarer Methoden / Metriken zur Bewertung deren Verlässlichkeit. Hierbei leitete Fraunhofer das Arbeitspaket zu Methoden und Werkzeugen zur Herstellung und zum Nachweis der Vertrauenswürdigkeit von KI-Funktionen. Des Weiteren wirkte Fraunhofer federführend an der Entwicklung und Nutzung der Operational Design Domain (ODD) mit und unterstützte die Erstellung der Sicherheitsargumentation, u.a. durch Einbeziehung der Argumentation zur Sicherheit der KI-Funktionen mit Hilfe der Goals Structuring Notation (GSN). Zur Standardisierung wirkte Fraunhofer bei der Erstellung der DIN DKE SPEC 99002 „Terminologie – KI in Bahnanwendungen" mit. Diese definiert grundlegende Begriffe für das neuartige Thema KI im Bahnumfeld. Die DIN DKE SPEC 99004 „Spezifikation von ODD im Schienenverkehr" wurde von Fraunhofer initiiert und geleitet. Sie beinhaltet unter anderem eine Beschreibung zum Vorgehen bzgl. der Integration der ODD im Entwicklungsprozess von Bahnanwendungen und eine ODD-Taxonomie für den Schienenverkehr. Somit wurde eine solide Grundlage für die Spezifikation und Verwendung von ODDs für zukünftige KI-Anwendungen im Bahnbereich geschaffen. An der Verbreitung der Projektergebnisse beteiligte sich Fraunhofer vielseitig, z. B. durch zahlreiche Veröffentlichung wissenschaftlicher Beiträge, Vorträge oder Beteiligung an Anwenderkreisen. Somit konnte Fraunhofer maßgeblich zur Zulassungskonformität und Vertrauenswürdigkeit von KI im Bahnbereich beitragen und übertragbare Grundlagen für zukünftige Anwendungen sicherer KI in sicherheitskritischen Domänen schaffen.

I.2 Voraussetzungen

Das Vorhaben safe.trAIn wurde als Verbundprojekt realisiert, gefördert durch das Bundesministerium für Wirtschaft und Klimaschutz sowie der Europäischen Union. Durch die Förderung wurde die Möglichkeit geschaffen, eine intensive Zusammenarbeit und kontinuierlichen fachlichen Austausch in diesem hoch potenzialen und risikobehafteten Themenbereich zu ermöglichen. Die Arbeiten wurden in enger Abstimmung mit den Projektpartnern, dem Verbundkoordinator und Projektträger durchgeführt.

Das Verbundvorhaben safe.trAIn konnte auf einer Reihe günstiger technischer, organisatorischer und marktseitiger Voraussetzungen aufbauen, die eine erfolgreiche Durchführung des Projekts ermöglichten. Auf technischer Ebene war insbesondere die bereits vorhandene Expertise der beteiligten Partner im Bereich der KI-gestützten Perzeption, sicherheitskritischer Systeme und Bahntechnologien von zentraler Bedeutung. Diese Vorarbeiten schufen eine gute Ausgangslage für die Erforschung von Lösungen für einen fahrerlosen Regionalzug und ermöglichten eine praxisnahe Umsetzung der Projektziele. Ein weiterer Erfolgsfaktor war die Bereitstellung einer gemeinsamen technischen Infrastruktur, die eine enge Zusammenarbeit im Konsortium unterstützte. Dazu zählten unter anderem eine Umgebung zur Verwaltung von KI-Modellen und Daten sowie eine gemeinsame GitLab-Umgebung. Diese Infrastruktur ermöglichte eine koordinierte Entwicklung und Integration der verschiedenen Komponenten und Methoden in verschiedenen Iterationen als sogenannte Minimal Viable Products (MVPs). Zudem waren auf methodischer Ebene weitere wichtige Voraussetzungen gegeben. Die Definition von Anforderungen an die technische Lösung sowie Akzeptanzkriterien für KI-Systeme im Bahnumfeld wurde durch die enge Verzahnung mit bestehenden Normen und Standardisierungsaktivitäten unterstützt. Die aktive Beteiligung von Partnern aus der Standardisierung begünstigte, regulatorische Anforderungen frühzeitig zu berücksichtigen und die Übertragbarkeit der Ergebnisse auf zukünftige Zulassungsverfahren einzubeziehen.

Nicht zuletzt war das Projekt durch ein hohes Maß an Interdisziplinarität und Kooperationsbereitschaft im Konsortium geprägt. Die Partner brachten komplementäre Kompetenzen aus den Bereichen KI, Bahntechnik, Normung, Sicherheit und Softwareentwicklung ein. Dies schuf die Voraussetzung für eine qualitativ hochwertige, termin- und budgetgerechte Umsetzung der Projektziele sowie für eine hohe Relevanz der Ergebnisse für Industrie und Forschung.

I.3 Planung und Ablauf des Vorhabens

Das Verbundvorhaben safe.trAIn wurde ursprünglich mit einer Laufzeit von drei Jahren von Januar 2022 bis Dezember 2024 geplant. Im Projektverlauf wurde die Laufzeit jedoch um drei Monate verlängert, sodass das Projekt Ende März 2025 erfolgreich abgeschlossen wurde. Ziel des Projekts war es, die sichere Anwendung von KI im fahrerlosen Schienenverkehr zu ermöglichen –

insbesondere durch die Entwicklung von Prüfmethoden, Sicherheitsarchitekturen und eines virtuellen Testfelds für KI-basierte Perzeptionssysteme.

I.3.1 Struktur und Arbeitspakete

Die Projektstruktur gliederte sich in sechs zentrale Arbeitspakete (AP):

- AP1: Anforderungen an die Sicherheitsnachweisführung
- AP2: Prüfmethoden und -werkzeuge zur Vertrauenswürdigkeitsbewertung von KI
- AP3: Sicherheitsarchitektur für KI-basierte Funktionen im GoA4-Betrieb
- AP4: Virtuelles Testfeld und Sicherheitsbewertung
- AP5: Standardisierung und Verbreitung der Ergebnisse
- AP6: Projektmanagement und Koordination

Diese Arbeitspakete waren inhaltlich und zeitlich eng miteinander verzahnt. Die Umsetzung des Projekts erfolgte in iterativen Sprints von mehreren Wochen. Die Arbeitspakete AP1 bis AP5 wurden hierbei jeweils durchlaufen, wobei die Ergebnisse iterativ weiterentwickelt und verfeinert wurden. Als Zwischenziele wurden sogenannte MVPs definiert, die inkrementelle Erweiterungen der verfolgten Gesamtlösung darstellten. Dieses Vorgehen ermöglichte es, neue technologische Entwicklungen flexibel zu integrieren und die entwickelten Konzepte mehrfach zu evaluieren und zu optimieren.

I.3.2 Zeitlicher Ablauf und Meilensteine

Die Projektplanung sah folgende zentrale Meilensteine vor:

- M6 (06-2022): Abschluss der initialen Analyse des Stands von Technik und Wissenschaft
- M12 (12-2022): Abschluss des ersten Innovationszyklus mit initialen Ergebnissen
- M24 (12-2023): Abschluss des zweiten Innovationszyklus mit überarbeiteten Ergebnissen
- M30 (12-2024): Verfügbarkeit finaler Methoden und Konzepte
- M39 (03-2025): Offizieller Projektabschluss mit finalen Ergebnissen

I.3.3 Projektsteuerung

Die Koordination des Projekts erfolgte durch den Verbundkoordinator. Zur Steuerung und Qualitätssicherung wurde ein organisatorischer Steuerkreis, bestehend aus mindestens einem Vertreter pro Projektpartner, und technischer Steuerkreis eingerichtet, bestehend aus dem Projektkoordinator, dem technischen Koordinator, den Arbeitspaketleitern sowie je einem Vertreter jedes Projektpartners. Der Steuerkreis traf sich regelmäßig und bei Bedarf, um den Projektfortschritt zu überwachen, Risiken zu identifizieren und Maßnahmen zur Zielerreichung abzustimmen. Zusätzlich zu den halbjährlichen Zwischenberichten wurden im Sinne eines agilen Vorgehens ca. alle acht Wochen informelle Kurzberichte an den Projektträger durch den Verbundkoordinator übermittelt, welche die Hauptergebnisse der jeweiligen Sprints dokumentierten. Fraunhofer

koordinierte die eigenen Arbeiten und Zusammenarbeit mit den Partnern über einen Projektkoordinator.

I.4 Wissenschaftliche und technische Ausgangslage

Das Projekt safe.trAIn knüpfte an den aktuellen Stand der Wissenschaft und Technik im Bereich der KI-gestützten Automatisierung sicherheitskritischer Systeme an – insbesondere im Kontext des fahrerlosen Schienenverkehrs. Während KI-basierte Perzeptionssysteme im Straßenverkehr bereits intensiv erforscht und teilweise erprobt wurden, fehlten bislang übertragbare Konzepte für den Bahnbereich, insbesondere im Hinblick auf die Sicherheitsnachweisführung und Zulassungsfähigkeit.

Ein zentrales Problem bestand darin, dass klassische Automatisierungstechnologien in offenen, komplexen Umgebungen – wie sie im Regionalverkehr vorherrschen – nicht ausreichen, um einen vollautomatisierten Betrieb (GoA4) zu realisieren. Gleichzeitig existierten im Bereich des hochautomatisierten Fahrens auf der Straße bereits fortgeschrittene Verfahren, insbesondere auf Basis von Deep Learning, die jedoch nicht ohne Weiteres auf den Bahnbereich übertragbar sind. Gründe hierfür liegen unter anderem in den strengeren regulatorischen Anforderungen, der Notwendigkeit größerer Sensorreichweiten sowie der fehlenden Erklärbarkeit und Nachvollziehbarkeit neuronaler Netze. Zur systematischen Erfassung und Bewertung des Stands der Technik wurden im Projektverlauf gezielte Literaturrecherchen, Normenanalysen und Gap-Analysen durchgeführt. Dabei kamen unter anderem wissenschaftliche Publikationen, Whitepapers, Standardisierungsdokumente sowie diverse Informations- und Dokumentationsdienste zum Einsatz.

Die Projektpartner konnten auf umfangreiche Vorarbeiten in verwandten Projekten zurückgreifen, etwa aus dem Automotive-Bereich (z. B. VDA KI-Absicherung), aus der Bahntechnik (z. B. ATO-Sense, BerDiBa, TAURO) sowie aus der industriellen KI. Zur Erarbeitung der eigenen Projektergebnisse wurde in safe.trAIn auf eine Vielzahl etablierter Verfahren und Konzepte zurückgegriffen. Dies beinhaltete beispielsweise KI-Methoden zur Unsicherheitsquantifizierung, zur Erhöhung der Robustheit oder auch Verfahren zur Erklärbarkeit von KI. Insbesondere wurden bestehende und in der Bearbeitung befindende Standards und Normen berücksichtigt. Einerseits betraf dies existierende Normen für den Schienenverkehr als auch für KI und Sicherheit von KI-Funktionen. Als Datengrundlage wurden öffentlich verfügbare Datensätze, wie RailSem19, verwendet, um Methoden zu entwickeln und validieren. Darüber hinaus wurden reale Betriebsdaten von Projektpartnern, wie vom Testgelände der Havelländischen Eisenbahn (HVLE) genutzt, um die entwickelten Methoden praxisnah zu validieren.

Diese fundierte Ausgangslage ermöglichte es dem Projektkonsortium, auf einem soliden wissenschaftlich-technischen Fundament aufzubauen und gezielt neue Methoden zur Absicherung von KI im Bahnbereich zu entwickeln.

I.5 Zusammenarbeit mit anderen Stellen

Im Rahmen des Projekts safe.trAIn wurde die Zusammenarbeit mit externen Stellen über das eigentliche Konsortium hinaus gezielt gefördert, um Synergien zu nutzen, den Wissenstransfer zu stärken und die Anschlussfähigkeit der Projektergebnisse sicherzustellen. Zahlreiche

Projektpartner waren bereits vor safe.trAIn in verwandte Projekte eingebunden, wodurch ein direkter Ergebnisaustausch möglich war. Dies ermöglichte es, auf bestehenden Erkenntnissen aufzubauen und gleichzeitig neue Impulse in laufende Aktivitäten zurückzuspielen. Darüber hinaus fand ein kontinuierlicher Austausch mit weiteren Partnern aus Wissenschaft und Forschung sowie relevanter Industrien zum Thema KI in sicherheitskritischen Systemen statt.

Gezielter Austausch wurde über die safe.trAIn Anwenderkreise forciert, in denen Projektzwischenergebnisse mit interessierten Expert:Innen diskutiert und reflektiert werden konnten. Hierdurch entstand ein intensiver Austausch über das Projekt hinaus, welcher auch zur Initiierung von DIN SPEC Standardisierungsaktivitäten und Einbeziehung zahlreicher Nicht-Konsortialpartner führte. Als weiteres Beispiel kann der Austausch im Rahmen des vom Projektträger organisierten Workshops zu „Automatic Train Operation (ATO) – Aktueller Stand und Ausblick in einem industriellen Schlüsselbereich" genannt werden, bei dem zahlreiche Partner aus verschiedenen aktuellen Projekten zur Automatisierung im Bahnverkehr aus dem Förderprogramm vertreten waren.

Ein zentrales Ziel von safe.trAIn war die Überführung der entwickelten Methoden in zukünftige Normen und Standards. Daher wurde die Zusammenarbeit mit relevanten Gremien und Initiativen aktiv gesucht und gepflegt. Hierzu zählten unter anderem DIN und VDE/DKE mit Beteiligung an der Erstellung von DIN SPECs (z. B. 99002, 99004) und Einbindung in nationale und internationale Normungsaktivitäten oder auch Zertifizierte KI bzgl. Entwicklung von Prüfgrundlagen für vertrauenswürdige KI-Systeme. Durch diese Vernetzung konnten Anforderungen aus der Praxis frühzeitig in die Standardisierungsarbeit eingebracht und die Anschlussfähigkeit der Projektergebnisse an regulatorische Entwicklungen unterstützt werden.

II. Eingehende Darstellung

II.1 Verwendung der Zuwendung und erzieltes Ergebnis

Ziel des Teilvorhabens war es, Methoden und Lösungen für den Nachweis der Vertrauenswürdigkeit von KI-Methoden für den autonomen Schienenverkehr zu erforschen. Dies erfolgte im Verbundprojekt in sechs Arbeitspaketen (APs), welche folgende Schwerpunkte gesetzt haben: Anforderungen an die Sicherheitsnachweisführung, Methoden und -Werkzeuge zur Herstellung und zum Nachweis der Vertrauenswürdigkeit von KI-Funktionen, Fahrzeugarchitektur im GoA4-Betrieb mit dem Fokus auf sichere KI-basierte Funktionen, Virtuelles Testfeld und Sicherheitsbewertung, Standardisierung und Verwertung sowie Projektmanagement. Die Verwendung der Zuwendung und erzielte Ergebnisse werden im Folgenden anhand dieser Themenstränge dargestellt.

II.1.1 AP1 Anforderungen an die Sicherheitsnachweisführung

Das Ziel dieses Arbeitspakets konzentrierte sich auf die Anwendbarkeitsanalyse bestehender Normen und Standards im Bereich KI, insbesondere in Bezug auf funktionale Sicherheit. In diesem Rahmen wurden Anforderungen an Prüfmethoden und -verfahren, sowie-messbaren Metriken abgeleitet, um die Qualität und Sicherheit von KI-Systemen zu gewährleisten. Das Ergebnis des Arbeitspakets sind Anforderungen und Richtlinien für die Zulassung und Produktsicherheit von Systemen mit KI-basierten Funktionalitäten im Bahnbereich. Der Fokus von Fraunhofer lag darin, quantifizierbare Metriken zu identifizieren, definieren und analysieren. Die Arbeiten zu quantifizierbaren Metriken zur Bewertung der Vertrauenswürdigkeit, Erklärbarkeit und Safety-Performance von KI wurde von Fraunhofer angeleitet. In diesem Zuge wurde untersucht, inwiefern in safe.trAIn objektive Bewertungen für Vertrauenswürdigkeit, Erklärbarkeit und Safety-Performance von KI-Funktionen mithilfe von geeigneten Metriken gemessen werden kann. Daraus wurden folgende Einzelziele abgeleitet:

- Entwicklung eines allgemeinen Vorgehens mit dem KI-spezifische Metriken identifiziert und kategorisiert werden können. Dabei werden die spezifischen, bekannten Unzulänglichkeiten inkludiert (d.h., Vertrauenswürdigkeit, Erklärbarkeit und Safety-Performance)
- Analyse und Definition von Anforderungen an quantifizierbare Metriken
- Anwendung des Vorgehens für den betrachteten Anwendungsfall in safe.trAIn

Das entwickelte Vorgehen zur Identifizierung und Einordnung von Metriken hat zum Ziel, den Status Quo zu ermitteln und darauf aufbauend den Bedarf an weiteren Metriken zu ermitteln. Hierbei werden projektspezifische Besonderheiten von safe.trAIn einbezogen, d.h., dass bereits eine gewisse Anzahl an Metriken durch die verschiedenen Projektpartner zum Start des Projekts eingebracht werden konnten. Diese Menge an Metriken diente als Grundlage dieser Analyse.

Damit ergeben sich die folgenden Schritte zur Identifikation und Kategorisierung von quantifizierbaren Metriken:

(1) Geeignete und sinnvolle Metriken im Sinne des definierten Anwendungsfalls sammeln

(2) Analyse des Abdeckungsgrades der gesammelten Metriken bezüglich relevanter Sicherheitsaspekte im Sinne der Anwendung

(3) Definition und Ableitung KI-spezifischer Metriken für identifizierte Lücken in der Abdeckung relevanter Sicherheitsaspekte

Dieses Vorgehen wurde durch Fraunhofer mit den beteiligten Partnern in AP1 besprochen und dokumentiert (siehe Abbildung 1). Für die Identifikation von Metriken wurde auf die Arbeiten und Ergebnisse aus AP2 zurückgegriffen. Insbesondere die bis zu diesem Zeitpunkt definierten und dokumentierten Metriken im Sinne der Anwendung waren die Basis für die Bewertung.

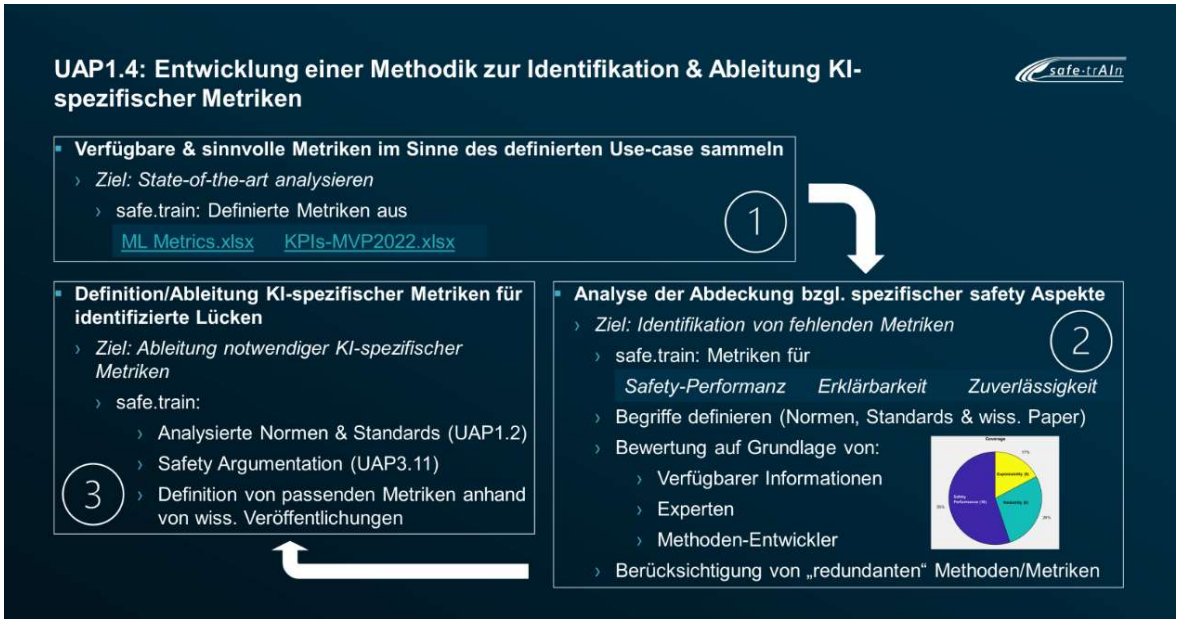
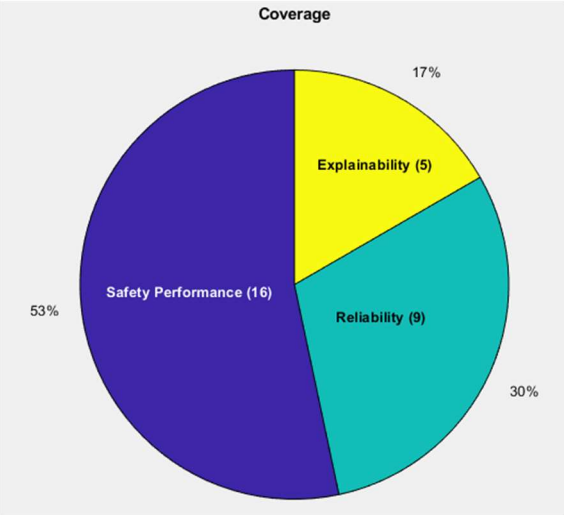


Abbildung 1: *Vorgehen zur Bestimmung von Metriken für Vertrauenswürdigkeit, Erklärbarkeit und Safety-Performance von KI. Diese Übersicht zeigt die drei allgemeinen Schritte und deren Ziel. Des Weiteren sind auch Details über die Realisierung innerhalb von safe.trAIn angegeben.*

Die Analyse zur Bestimmung relevanter Sicherheitsaspekte umfasste die bereits genannten Aspekte: Vertrauenswürdigkeit, Erklärbarkeit und Safety-Performance. Um eine Kategorisierung zu ermöglichen, mussten diese Aspekte konkret gefasst und definiert werden. Dazu wurde auf aktuelle Normen und Standards zurückgegriffen, welche innerhalb des Arbeitspakets gesammelt und analysiert worden sind. Dies geschah im Austausch zwischen Fraunhofer und den weiteren beteiligten Partnern aus AP1. Die Bewertung beinhaltete neben den genannten Sicherheitsaspekten auch eine Untersuchung hinsichtlich Redundanz. Somit ist nicht nur die absolute Anzahl an Metriken für einen bestimmten Aspekt relevant, sondern vor allem, inwiefern sich diese

Metriken ergänzen oder ähneln. Abbildung 2 zeigt die Ergebnisse dieser Analyse für die bereits bestehenden Metriken in safe.trAIn.

Abdeckung ohne Redundanz-Analyse:



Abdeckung mit Redundanz-Analyse:

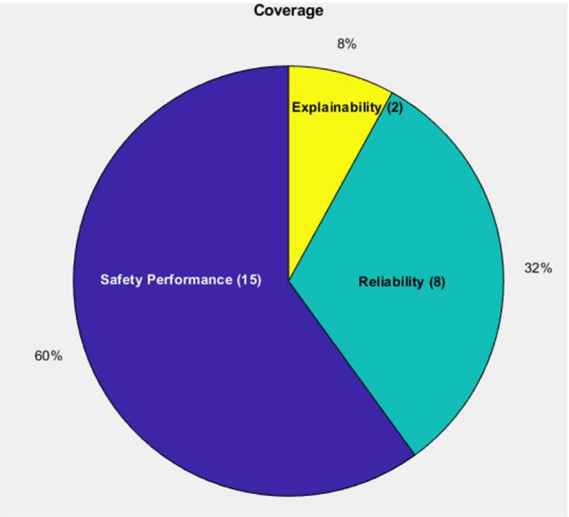


Abbildung 2: *Auswertung der Abdeckung definierter Metriken in safe.trAIn bezüglich Safety-Performance, Erklärbarkeit und Zuverlässigkeit* [1]

Die Analyse zur Ermittlung der Redundanz erfolgte anhand von Informationen durch die Dokumentation innerhalb safe.trAIn (AP2) sowie durch Experten und Metriken-Entwicklern (AP1, AP2 und AP3). Im letzten Schritt wurden Lösungen für die aufgedeckten Lücken erforscht. Die Analyse zeigt den aktuellen Status und lässt somit auch die Identifikation von unzureichenden Metriken zu. Um diese Lücken zu schließen, bedarf es zweier Phasen:

1. Die Definition von geeigneten Metriken für die kritischen Sicherheitsaspekte
2. Eine Bewertung dieser Metriken hinsichtlich deren Beitrag, um die konkreten kritischen Sicherheitsaspekte zu schließen oder zu minimieren (unter Berücksichtigung bereits definierter Metriken)

Wie Abbildung 2 zeigt, wurde zu Beginn von safe.trAIn identifiziert, dass insbesondere sehr wenige Metriken für die Erklärbarkeit verwendet wurden. Infolgedessen hat Fraunhofer zunächst anhand von Normen und Standards, dem zu erarbeiteten Sicherheitsnachweis aus AP3 sowie dem EU AI Act eine Vorauswahl und Voranalyse für geeignete Metriken angefertigt. Diese beinhaltet die folgenden Punkte: Funktionsweise, Vor- und Nachteile, Implementierungsauswand, Gegenüberstellungen zu bereits existierenden Methoden und Mehrwert für safe.trAIn aus Sicherheitsperspektive. Die finalen Metriken wurden dem Konsortium als mögliche Ergänzungen vorgestellt und sind in Tabelle 1 dargestellt.

Metrik	**☑ Vorteile, ☒ Nachteile & ↻ Nutzen für safe.trAIn**
CLRP	☑ Keine Änderungen am Basisnetzwerk erforderlich ☑ Instanz-spezifische (lokale Erklärbarkeit) für: ☑ Klassendiskriminierung

[1] *Die absolute Anzahl an Metriken pro Kategorie ist angegeben durch die Ziffern in Klammern. Die relative Verteilung der Metriken zueinander ist gegeben durch die prozentuale Angabe. Metriken, deren Messmethodik auf ähnlichen Prinzipien basiert, gelten als redundant und werden entsprechend nur einmal gezählt. Dadurch reduziert sich die effektive Anzahl an Metriken pro Kategorie.*

	☑ Erklärung für jede Klasse
	☒ Nur lokale Erklärbarkeit (d.h., jeder Input separat) ☒ Relativ rechenintensiv ☒ Nicht auf alle Arten von Netzwerken anwendbar
	↻ Erstellt Heatmaps pro Klasse anstelle einer einzelnen Heatmap für den kompletten Input. Ermöglicht das Zurückverfolgen der Neuronen, die pro Klasse aktiviert wurden.
Soft DT	☑ Keine Änderungen am Basisnetzwerk erforderlich ☑ Erklärbares Modell in zwei Varianten: ☑ DT ersetzt DNN ☑ Generiert Regeln für das initiale NN
	☒ Für *DT ersetzt DNN*: ☒ (deutlich) schlechtere Performance als das entsprechende NN ☒ Verwendet NN um mehr Trainingsdaten zu generieren: Problem, da NN selbst eine Black-Box ist sind gezielte Verbesserung schwierig ☒ Für *Generiert Regeln (für das initiale NN)*: ☒ Skaliert nicht gut für komplexe NN ☒ Durchwachsene Performance für reale, performante NNs ☒ Erfordert ausgewertete Testdaten
	↻ Könnte prinzipiell Regeln erstellen, die Einblicke in das Basis-NN geben ↻ Unterschiedliche Funktionalität im Vergleich zu allen anderen Methoden
Grad-CAM	☑ Keine Änderungen am Basisnetzwerk erforderlich ☑ Erklärung(en) für die Vorhersage des Netzwerks durch: ☑ Enthält spezifische Details (z.B. Form, Farbe) ☑ Ist exakt für eine einzelne Vorhersage ☑ Kann auch für die Datenanalysen verwendet werden
	☒ Nur lokale Erklärbarkeit (d.h., jeder Input separat)
	↻ Ähnlich zur bestehenden Methode "ODD-Konzeptabdeckungsmetrik", da beide gradientenbasierte Methoden sind ↻ Grad-CAM benötigt keine a-priori Konzepte (d.h., geringer Implementierungsaufwand), jedoch ist dies für Erklärbarkeit nicht unbedingt ein Vorteil.
SHAP	☑ Keine Änderungen am Basisnetzwerk erforderlich ☑ Ein approximatives " Modell (lineare Funktion auf binären Variablen) quantifiziert die Bedeutung aller Eingangsmerkmale für das Endergebnis ☑ Das approximative Modell verwendet nicht den vollständigen Eingangsraum, sondern nutzt eine Zuordnung von Eingangsmerkmalen zu "Superpixel" oder Konzepten
	☒ SHAP-Berechnung ist *sehr* rechenintensiv ☒ Die Methode erstellt ein Black-Box-Modell, dessen Erklärungen selber nicht erklärbar sind
	↻ Ermöglicht die Schätzung, wie viel jedes Eingangsmerkmal zu jedem Ausgang beiträgt ↻ Im Fall der Objekterkennung, wird gezeigt wie Superpixel-Cluster zu jedem Konzept beitragen (z.B. in Kombination mit TCAV)

Tabelle 1: *Übersicht aller analysierten Metriken die Erklärbarkeit liefern. Diese Auswahl enthält bereits nur Metriken welche ergänzend zu bereits vorhandenen Metriken fungieren (vgl.* Abbildung 2).

Um eine möglichst objektive Grundlage zur Priorisierung dieser Metriken (siehe Tabelle 1) zu schaffen, wurde eine anonymisierte Umfrage durch Fraunhofer erstellt und an alle relevanten safe.trAIn Projektpartner weitergeleitet. Das Ziel dieser Umfrage war es herauszufinden, welche der verfügbaren Metriken für Erklärbarkeit den größten Mehrwert für safe.trAIn erbringen kann und somit favorisiert verfolgt werden sollte.

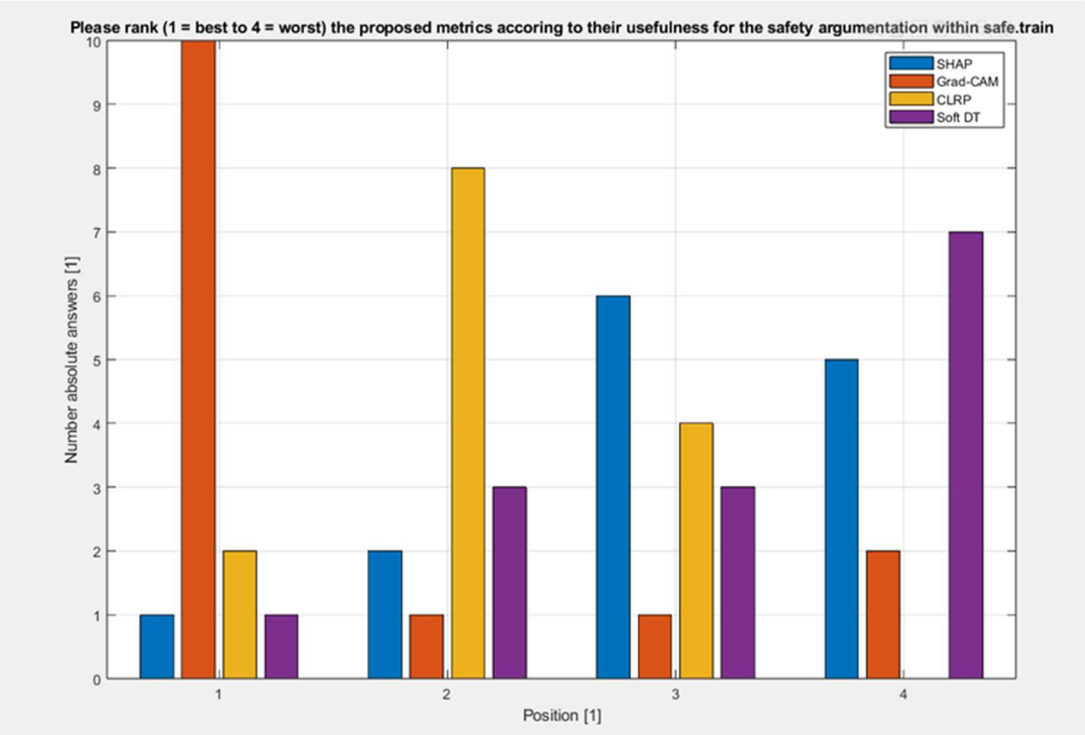


Abbildung 3: *Bewertung der Metriken zu Erklärbarkeit durch das safe.trAIn Konsortium, hinsichtlich deren Zweckmäßigkeit für die safe.trAIn Sicherheitsargumentation*

Die erzielten Umfrageergebnisse hat Fraunhofer an die verantwortlichen Metriken- und Methodenentwickler aus AP2 zurückgespielt.

Bei der Anwendung des Vorgehens bei safe.trAIn sind neben den bereits vorgestellten Resultaten weitere „Lessons Learned" entstanden. Diese gewonnenen Erkenntnisse, sind in den folgenden einzelnen Schritten aufgeführt:

Erkenntnisse aus Schritt (1):

- Eine verständliche, gemeinsame Dokumentation aller bereits definierter Metriken ist unabdinglich für die Identifikation, Kategorisierung sowie Analyse.

Erkenntnisse aus Schritt (2):

- Interviews mit Metrik-Entwicklern sind besonders hilfreich bzgl. einer korrekten Bewertung
- Bei der Bewertung der Metriken sind 1-n Beziehungen möglich, d.h. eine Metrik kann mehrere Sicherheitsaspekte gleichzeitig adressieren.
- Allerdings muss eine Metrik nicht zwangsläufig einen Sicherheitsaspekt abdecken.
- Die Bewertung von Redundanz ist für viele Metriken nicht trivial und der Grad der Korrelation kann in vielen Fällen nicht ohne Unsicherheit bestimmt werden.

Erkenntnisse aus Schritt (3):

- Aktuell gibt es kein etabliertes Verfahren zur Beurteilung, ob eine bestimmte Anzahl an Metriken als *ausreichend* für eine Eigenschaft betrachtet werden kann. In safe.trAIn wurden die folgenden Ansätze diskutiert:

- Analyse des GSN-Sicherheitsnachweises und Abschätzung ob die definierten Evidenzen durch bereits definierte Metriken eine ausreichende Abdeckung erfahren. Problem: Eine GSN-Argumentation muss vollständig und in ausreichender Güte vorliegen oder erarbeitet werden.
- Metriken aus Normen und Standards systematisch analysieren, um zu ermitteln ob alle Anforderungen erfüllt sind. Problem: Aktuelle Normen und Standards definieren nur sehr abstrakte Konzepte und Ideen. Es ist nicht möglich konkrete Abdeckungsgrade abzuleiten.

II.1.2 AP2 Methoden und Werkzeuge zur Herstellung und zum Nachweis der Vertrauenswürdigkeit von KI-Funktionen

In safe.trAIn übernahm Fraunhofer die Gesamtkoordination und -organisation der Arbeiten im Rahmen des AP2 zum Nachweis der Vertrauenswürdigkeit von KI-Funktionen. Fraunhofer leitete die State-of-the-Art (SotA) Analyse zur Absicherung von KI-Funktionen an. Hierfür wurde der Transfer von bestehenden Methoden zur Absicherung von KI-Systemen auf den Anwendungsfall eines fahrerlos fahrenden Regionalzugs untersucht. Es wurden insbesondere Methoden betrachtet, welche die Goal Structuring Notation (GSN) verwenden. Darüber hinaus wurden Einsatzmöglichkeiten von sogenannten Foundation-Modellen für die Bildverarbeitung untersucht. Die Ergebnisse dienten als Grundlage für weitere Arbeiten des Vorhabens. Unter Leitung von Fraunhofer wurde ein Konzept zur systematischen Erstellung eines Nachweises für die Vertrauenswürdigkeit von KI-Funktionen ausgearbeitet. Dabei wurden exemplarisch Safety Cases (im GSN-Format) entwickelt, welche die Beiträge des AP2 vereinen. Hierfür wurden für ausgewählte KI-Teilsysteme, d.h. Track Detektor und Mid-Level Fusionskomponente, ein Safety Case entwickelt, wobei der Fokus auf den Aspekten Robustheit, Performance, Datenqualität und Transparenz/Erklärbarkeit lag. Hierfür fand ein regelmäßiger Austausch mit den Arbeiten zur Sicherheitsnachweisführung in AP3 und zur entwickelten Landscape of AI Safety Concerns (LAISC) statt, insbesondere auch hinsichtlich notwendiger sicherheitsrelevanter Evidenzen unter der Berücksichtigung der GoA3/4 Architektur. Die Sicherheitsargumentation wurde iterativ, unter Abstimmung der involvierten KI-, Metriken-EntwicklerInnen und SicherheitsexpertInnen entwickelt. Für jeden Aspekt wurde ein übergeordnetes Sicherheitsziel definiert, das in weitere Teilziele heruntergebrochen wurden, bis messbare Evidenzen für Gegenmaßnahmen definiert werden konnten. Zusätzlich wurde eine Kategorisierung der in AP2 entwickelten KI-Metriken durchgeführt, welche einen Nachweis zur Wirksamkeit der im Safety Case definierten Maßnahmen liefern. Die entwickelten Metriken wurden des Weiteren auf ihre Vorteile und Grenzen hin analysiert, so dass sie als Evidenzen passenden Safety Goals im GSN-Baum zugeordnet werden konnten. Abschließend wurde die Konsistenz zu weiteren Projektergebnissen (z.B. LAISC, Validation Reports) überprüft und sichergestellt.

Für die Methodenentwicklung zur Nachweisführung wurden verschiedene Lösungen von Fraunhofer entwickelt, welche mit den erzielten Ergebnissen im Folgenden detailliert beschrieben werden.

II.1.2.1 Semantic Performance Discrepancy zur Erkennung systematischer Schwächen

Die Metrik *Semantic Performance Discrepancy (SPD)* unterstützt die Bewertung von Deep Neural Networks (DNNs) für sicherheitskritische Anwendungen (vgl. Abbildung 4). Das primäre Ziel dieser Metrik ist es, systematische Schwächen oder Verzerrungen zu identifizieren, die ein DNN während des Trainings erlangen kann. Sie überträgt einen Ansatz der klassischen KI zur Schwachstellensuche in strukturierten Daten (motiviert vom Algorithmus "Sliceline") auf die Anwendung für DNNs, die Bilddaten verarbeiten – hier zum Zwecke der Personenerkennung. Diese Schwächen entstehen oft aufgrund von Faktoren wie falschen Korrelationen in den Daten, verrauschten oder falschen Beschriftungen und unausgewogenen Datenverteilungen. Durch die Identifizierung von Leistungsdiskrepanzen in semantisch bedeutsamen Teilmengen von Daten ermöglicht die Metrik eine detailliertere Modellbewertung als herkömmliche mittelwertbasierte Metriken. Die erzielten Ergebnisse sollen als Grundlage für die Formulierung von Sicherheitsnachweisen für KI-gesteuerte Systeme dienen.

II.1.2.1.1 Überblick über die Metrik

Die Metrik umfasst zwei Phasen: die Erzeugung von Metadaten und die systematische Ermittlung von Schwachstellen. Im ersten Schritt werden semantische Metadaten aus Datensätzen extrahiert, indem das *CLIP (Contrastive Language Image Pre-training)*-Modell eine Zero-Shot-Klassifizierung vornimmt. Mit Hilfe von Ground-Truth Bounding Boxes werden Fußgängerbilder aus Datensätzen wie RailSem19[2] zugeschnitten. Diese zugeschnittenen Bilder werden dann von CLIP verarbeitet. Dabei werden textuelle Aufforderungen, die verschiedene Dimensionen der *Operational Design Domain (ODD)* darstellen, mit Bildeinbettungen verglichen, um jede Instanz über mehrere semantische Achsen zu klassifizieren. Das Ergebnis dieses Schritts ist eine strukturierte CSV-Datei, die in der jede Zeile Metadaten für eine einzelne Fußgängerinstanz enthält.

In der zweiten Phase werden diese Metadaten mit den Leistungsdaten des getesteten DNN zusammengeführt. Typischerweise sind das Metriken wie *Intersection over Union (IoU)* auf Objektebene. Die kombinierten Daten werden dann mit dem *SliceLine*-Algorithmus analysiert. SliceLine konstruiert ein Gitter potenzieller Datenteilmengen auf der Grundlage eines oder mehrerer semantischer Attribute und bewertet die Leistungsverschlechterung in jeder Teilmenge. Teilmengen oder „Slices", die eine deutlich geringere Leistung als der Gesamtdatensatz aufweisen, werden als „weak slices" gekennzeichnet. Die Metrik gibt keinen einzelnen skalaren Wert aus, sondern liefert eine detaillierte Bewertung der Modellschwächen in Verbindung mit bestimmten semantischen Dimensionen.

[2] RailSem19: A Dataset for Semantic Rail Scene Understanding, https://www.wilddash.cc/railsem19

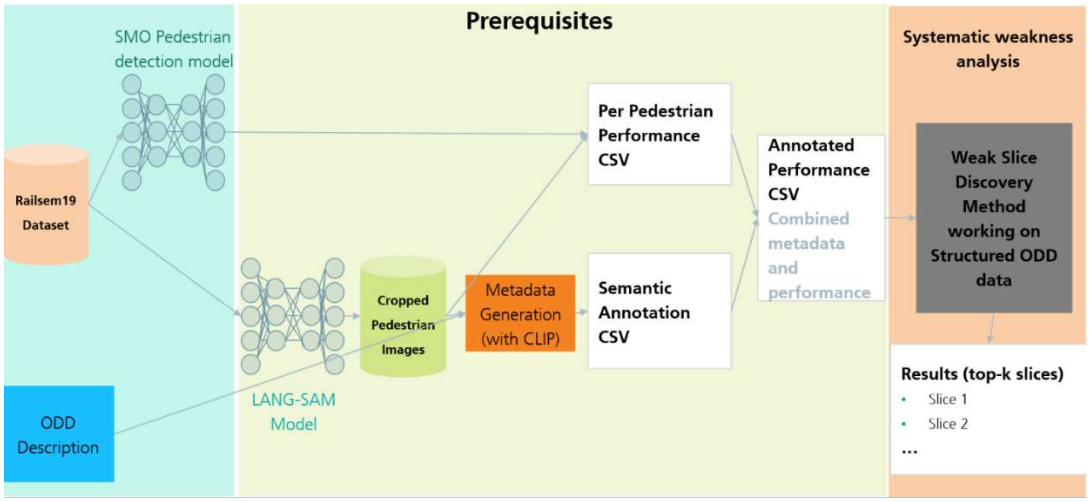


Abbildung 4: *Gesamtablauf unserer Methode*

Die Metrik wurde in Python implementiert. Sie erfordert mehrere Schlüsseleingaben: einen Datensatz mit Fußgängervielfalt (wie RailSem19), trainierte Objekterkennungs- oder semantische Segmentierungsmodelle (z. B. Panoptic FCN) und eine von Sicherheitsingenieuren erstellte ODD im JSON-Format. Darüber hinaus wird eine menschliche Bewertung in der Methode einbezogen, um die Genauigkeit der CLIP-generierten Metadaten zu quantifizieren und Korrekturen auf der Grundlage der beobachteten Abweichungen zu ermöglichen.

II.1.2.1.2 Erzeugung von Metadaten

Eine Kernkomponente der Metrik ist die Erzeugung semantischer Metadaten pro Objekt. Diese Metadaten ermöglichen es, Modelle nicht nur auf der Grundlage von Gesamtergebnissen zu bewerten, sondern anhand semantisch sinnvoller Partitionen, die gemeinhin als „Slices" bezeichnet werden. Diese Slices sind auf die Dimensionen der ODD abgestimmt, wie bei Personen z. B. Alter, Geschlecht, Kleidung und Verhaltensmerkmale. Eine zentrale Herausforderung bei der Umsetzung einer solchen Methodik in großem Maßstab ist die zuverlässige Generierung dieser Metadaten aus Rohbilddaten. Eine manuelle Beschriftung ist kostspielig und bei großen Datensätzen inkonsistent. In diesem Zusammenhang stellen die Ergebnisse von Fraunhofer in safe.trAIn einen wichtigen Beitrag zur Operationalisierung der Metrik dar. Diese wurden u.a. auf dem Workshop über sichere künstliche Intelligenz für automatisiertes Fahren (SAIAD) bei der internationalen Konferenz CVPR vorgestellt. Die Arbeit evaluiert die Durchführbarkeit, Leistung und Zuverlässigkeit von CLIP (einem Basismodell, das auf Bild-Text-Paaren trainiert wurde) für die semantische Kennzeichnung von für das Autonome Fahren (AF) relevanten Datensätzen.

Semantic dimension	Attributes			
Gender	Male	Female		
Skin color	White	Dark		
Age	Young	Adult		
Hair color	Black	Blond	Gray	Brown
Clothing color	Bright-color	Dark-color		
Blurry	True	False		
Construction-worker	True	False		

Abbildung 5: *ODD für die Bewertung von DNNs im Projekt*

Ziel dieser Arbeiten war es zu untersuchen, ob CLIP zuverlässig für die Generierung von Metadaten im Kontext des AF verwendet werden kann. Die Forschungshypothese war, dass CLIP (aufgrund seines Abgleichs von visuellen und textuellen Einbettungsräumen) als Zero-Shot-Klassifikator über mehrere semantische Dimensionen hinweg dienen kann, ohne jegliche Feinabstimmung. Die Evaluierung umfasst einen synthetischen AF-Datensatz mit kontrollierten Annotationen, die öffentlich verfügbaren RailSem19- und Cityscapes[3]-Datensätze sowie den CelebA[4]-Datensatz. Letzterer dient aufgrund seiner feinkörnigen Attributbeschriftungen für Gesichtsbilder als Ausgangsbasis.

Die Methode beginnt mit der Extraktion von ausgeschnittenen Objektbildern (z. B. Fußgänger) aus größeren Szenen unter Verwendung von Bounding Boxes, die der sogenannten „Ground Truth" entsprechen. Jedes ausgeschnittene Bild wird dann durch den CLIP-Vision Encoder bearbeitet, während eine Reihe von manuell erstellten Textaufforderungen (z. B. „eine Person, die eine rote Jacke trägt") durch den CLIP-Textkodierer verarbeitet wird. Die Klassifizierung basiert dann auf dem Nächster-Nachbar-Prinzip im Einbettungsraum. Um die Robustheit der Klassifizierung zu verbessern, wird eine sogenannte Ensemble-Prompt-Strategie verfolgt: Jedes semantische Konzept wird durch eine Sammlung von Prompts anstelle einzelner Phrasen repräsentiert. Das endgültige Label wird durch einen Auswahlmechanismus aufgrund der Prompt-Vorhersagen bestimmt. Zusätzlich wird eine Methode zur Rauschunterdrückung verwendet, die Vorhersagen mit hoher Entropie oder geringer Konfidenz entfernt, um Fehlklassifizierungen zu reduzieren.

Die Ergebnisse der Studie zeigen, dass die Verwendung von CLIP für die Generierung semantischer Metadaten sowohl vielversprechend als auch kritisch zu betrachten ist. Bei synthetischen Datensätzen, bei denen die Position, die Beleuchtung und das Erscheinungsbild von Objekten kontrolliert wurden, zeigt CLIP eine hohe Präzision und Wiedererkennung für eine Reihe von Attributen. Bei real aufgenommenen Datensätzen wie RailSem19 und Cityscapes fällt die Leistung jedoch deutlich ab. Bei auffälligen Attributen wie Alter und Geschlecht erreicht CLIP eine recht hohe Leistung. Bei subtileren Merkmalen wie der Beschaffenheit der Kleidung, Accessoires oder

[3] The Cityscapes Dataset, https://www.cityscapes-dataset.com/
[4] Large-scale CelebFaces Attributes (CelebA) Dataset, https://mmlab.ie.cuhk.edu.hk/projects/CelebA.html

der Körperhaltung gibt es jedoch Schwierigkeiten, insbesondere wenn diese Merkmale nicht auffällig sind oder in mehrdeutigen Kontexten auftreten.

Eine detaillierte Aufschlüsselung der Ergebnisse zeigt, dass die Ensemble-Prompt-Methode die Gesamtgenauigkeit im Vergleich zur Single-Prompt-Inferenz erhöht. Dennoch sind die Genauigkeitsgewinne attributabhängig und aufgrund von Datenvielfalt, Verdeckungen und unterschiedlichen Kamerastandpunkten weniger konsistent in realen Datensätzen. Die Technik der Rauschunterdrückung eliminiert effektiv unsichere Vorhersagen, wodurch die Genauigkeit verbessert wird, allerdings auf Kosten einer geringeren Trefferquote. Dieser Kompromiss wird bei sicherheitskritischen Bewertungen als akzeptabel angesehen, bei denen falsch-positive Ergebnisse eher toleriert werden können als falsch-negative.

Hervorzuheben ist, dass die Leistung von CLIP bzgl. der Metadaten anhand von Menschen beschrifteten Grunddaten quantifiziert wurde. Durch die Berechnung der Präzision, der Wiedererkennung und der F1-Werte für die untersuchten Teilmengen werden empirische Belege für die Grenzen der Verwendung von CLIP als eigenständiger Annotationsmechanismus geliefert. Dies ist von direkter Bedeutung für die Metrik, die in hohem Maße von der Zuverlässigkeit der Metadaten für eine aussagekräftige Slice-basierte Analyse abhängt.

II.1.2.1.3 Auswirkungen auf die Integration in der Metrik

Die vorangegangene Analyse bestätigt, dass CLIP eine skalierbare, erste Annotationsebene für Metadaten auf Objektebene bereitstellen kann, wodurch die Notwendigkeit einer umfassenden manuellen Beschriftung drastisch reduziert werden kann. Diese Möglichkeiten unterstützen direkt den Workflow der Metrik und ermöglichen eine effiziente Umwandlung von Bilddaten in eine strukturierte semantische Form. Die erforschten Ensemble- und Rauschunterdrückungstechniken, die beide mit dem Vorgehen kompatibel sind, bieten praktische Verbesserungen zur Sicherstellung der Annotationsqualität. Gleichzeitig verdeutlichen die in den Arbeiten identifizierten Einschränkungen, dass bei der Verwendung von CLIP-generierten Metadaten Vorsicht geboten ist, insbesondere für nachfolgende Sicherheitsanalysen. Die Metrik zielt darauf ab, schwache Leistungsabschnitte zu identifizieren, die auch semantisch interpretierbar und umsetzbar sind. Hierbei könnte eine schlechte Metadatentreue zur Identifizierung von falschen oder nicht vorhandenen Modellschwächen führen. Dieses Risiko ist besonders immanent, wenn Modelle für regulatorische Zwecke oder für die Sicherheitszertifizierung bewertet werden.

II.1.2.1.4 Ergebnisse der Bewertung

Die Metrik wurde sowohl im Vergleich zu modernen, existierenden systematischen Schwachstellenerkennungsmethoden als auch im Rahmen des Projekts bewertet. Auf dem CelebA-Datensatz wurde der SPD-Ansatz mit DOMINO, Spotlight und SVM FD unter Verwendung eines auf ImageNet21k trainierten ViT-Modells verglichen. Während einige State-of the-Art (SOTA)-Methoden Slices mit höherer Leistungsverschlechterung identifizierten, lieferte SPD besser interpretierbare und umsetzbare Beschreibungen dieser Slices. Dabei ist die Einbeziehung des semantischen Kontexts ein entscheidender Vorteil, der es Sicherheitsingenieuren ermöglicht, die Ergebnisse direkt mit den ODD-Dimensionen in Verbindung zu bringen. Darüber hinaus wurde die Metrik verwendet, um das entwickelte Fully Convolutional Networks for Panoptic Segmentation (Panoptic FCN)-Fußgängererkennungsmodell auf Basis des RailSem19-Datensatzes zu bewerten. Aufgrund

der Vielfalt der Fußgängervarianten und -szenen in diesem Datensatz konnte die Metrik die fünf schwächsten Slices identifizieren (vgl. Abbildung 6), die eine erhebliche Leistungsverschlechterung aufwiesen und mit sicherheitsrelevanten ODD-Dimensionen abgeglichen wurden. Quantitative Details wie die Größe der Slices, die False-Negative-Rate und die durchschnittliche Verschlechterung wurden hierfür erhoben.

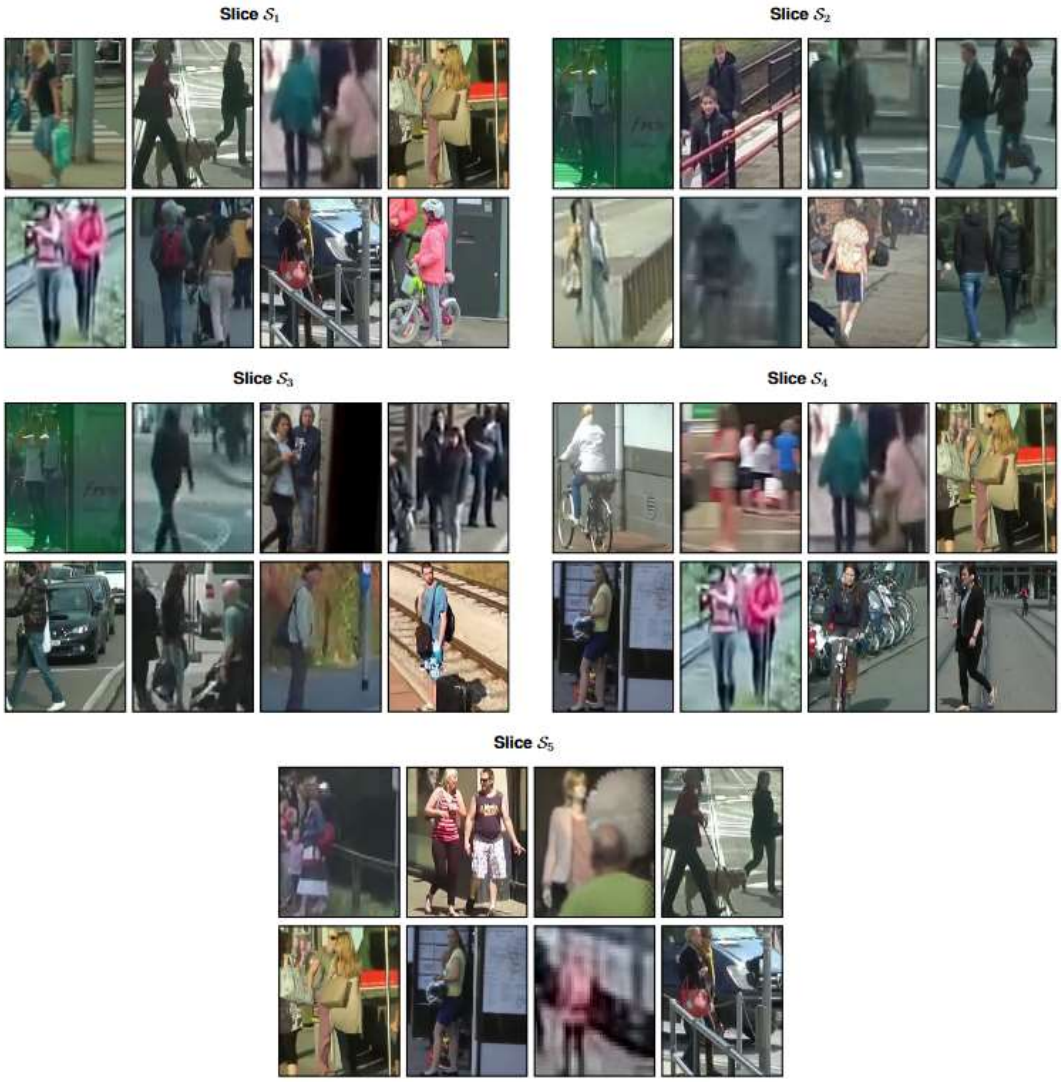


Abbildung 6: *Top-5 Slices, die vom erforschten Ansatz als schwache Slices identifiziert wurden*

II.1.2.1.5 Fazit

Die Metrik Semantic Performance Discrepancy bietet eine Methode zur Erkennung und Erklärung systematischer Schwächen in Deep-Learning-Modellen, insbesondere in sicherheitskritischen Kontexten wie der Fußgängererkennung. Durch die Ausrichtung der Modellbewertung an semantisch definierten operativen Dimensionen stellt die Metrik sicher, dass identifizierte Fehler nicht nur statistisch signifikant, sondern auch real relevant sind. Sie ergänzt herkömmliche Modellvalidierungsansätze um eine tiefere, besser interpretierbare Analyseschicht, die sowohl den Modellentwicklern als auch den Sicherheitsingenieuren bei ihren jeweiligen Aufgaben hilft.

Damit ist die Metrik ein robustes Werkzeug für das übergeordnete Ziel, vertrauenswürdige und sichere KI-Systeme zu entwickeln. Die Metrik und die Experimente werden in der Fachzeitschrift „Transactions on Machine Learning Research" (TMLR) veröffentlicht.

Allerdings gibt es auch Grenzen für den Einsatz dieser Metrik. Erstens sind die mit CLIP erzeugten Metadaten von Natur aus verrauscht. Obwohl die menschliche Bewertung dies abschwächt, lassen sich Kennzeichnungsfehler nicht vollständig ausschließen. Zweitens kann keine einzelne Methode alle ODD-Dimensionen erschöpfend abdecken. Eine Integration zusätzlicher Sensordaten könnte die semantische Abdeckung jedoch verbessern. Drittens muss der Datensatz eine ausreichend vielfältige Repräsentation der semantischen Dimensionen enthalten, damit die Metrik effektiv funktioniert. Der SliceLine-Algorithmus beschränkt sich auf vierdimensionale Kombinationen semantischer Attribute, um eine Fehlerfortpflanzung durch verrauschte Beschriftungen zu verhindern.

II.1.2.2 Visuelle Inspektionsabdeckung mit ScrutinAI

Die visuelle Inspektionsabdeckung ist eine Metrik, die sich auf die Zeit bezieht, die ein Experte oder Prüfer damit verbringen, die Leistung des untersuchten DNN mit einem visuellen Analysewerkzeug (*ScrutinAI*) auf einem Testdatensatz im Detail zu analysieren. Diese werkzeuggestützte Analyse erlaubt es, Erkenntnisse über die Schwächen oder Stärken des zu testenden DNN auf einem Testdatensatz unter Berücksichtigung der Metadaten von Objekten (Personen) zu gewinnen. Im „besten Fall" liefert die Metrik den Nachweis, dass auch mit dieser Toolunterstützung keine sicherheitsrelevanten Schwächen entdeckt wurden. Kann dieser Nachweis nicht erbracht werden, ermöglicht die Metrik eine detaillierte Rückmeldung an den DNN-Entwickler und an die Sicherheitsingenieure über bestehende sicherheitsrelevante Probleme.

Die Metrik wurde auf einen Entwicklungsstand des geplanten safe.trAIn-Personendetektor-Modells (Panoptic-Modell) angewendet, wobei Railsem19-Daten als Testdatensatz verwendet wurden. Der Railsem19-Testdatensatz wurde ursprünglich für die Segmentierung und nicht für die Objektdetektion erstellt und veröffentlicht. Die Metadaten für die Annotation von Personen im Datensatz wurden im Rahmen der von Fraunhofer entwickelten Semantic Performance Discrepancy-Metrik (s. vorigen Abschnitt II.1.2.1) generiert.

Der numerische Ergebniswert der Metrik hierbei ist 0 (s. auch unten „Ausgabe der Metrik"). Weitere Ergebnisse der Anwendung der Metrik und Einblicke in spezifische Probleme mit dem getesteten Modell sind weiter unten aufgeführt. Annotationen für Daten des in safe.trAIn betrachteten Testgeländes der Havelländischen Eisenbahn (HVLE) in Berlin waren zum Zeitpunkt der Auswertung noch nicht verfügbar. Da die im Projekt vorhandenen HVLE-Daten außerdem nur zwei Personen enthalten, bieten sie nicht genügend Varianz als Testdatensatz für die Personenerkennung.

Die Metrik für die visuelle Inspektion der Abdeckung basiert auf einem *Human-in-the-Loop-Ansatz*. Das Tool unterstützt einen menschlichen Entwickler oder Sicherheitsexperten (beide Rollen werden im Folgenden ohne weitere Unterscheidung als „Analyst" bezeichnet) bei der interaktiven Suche nach Leistungsschwächen, Ausreißern sowie Fehlern im DNN oder in den Daten. Die Methode befasst sich hauptsächlich mit dem Sicherheitsproblem der mangelnden Erklärbarkeit. Sie ermöglicht es dem Analysten, die Ursachen für fehlerhafte Vorhersagen zu ermitteln, indem er die DNN-Entscheidungen mit einem interaktiven visuellen Analysewerkzeug aus

verschiedenen Perspektiven untersucht. Ziel ist es, dem Analysten die Möglichkeit zu geben, die während des Analyseprozesses gewonnenen Erkenntnisse zu nutzen, um Hinweise für den Ursprung oder das Fehlen spezifischer Unzulänglichkeiten zu finden. Dies trägt schließlich zu einer Gesamtargumentation für ein vertrauenswürdiges DNN-Modell bei.

Es hat sich gezeigt, dass DNNs im Trainingsprozess Verzerrungen oder systematische Schwächen erlernen, die auf falsche Korrelationen, verrauschte Labels und unausgewogene Verteilungen zurückzuführen sind. Zusätzlich zur automatischen Erkennung semantischer Leistungsdiskrepanzen, um nach systematischen Schwächen zu suchen, kann ein Human-in-the-Loop-Ansatz die Möglichkeit bieten, DNNs auf systematische Schwächen zu evaluieren. Das verwendete Visual Analytics (VA)-Tool stellt Modelleingaben, -ausgaben und Metadaten durch verknüpfte, interaktive Elemente visuell dar (vgl. Abbildung 77). Beispielsweise werden Modelleingaben und -vorhersagen mit Bounding-Box-Kommentaren und Pop-up-Informationen dargestellt. Metadaten werden als Histogramme, Diagramme und Tabellen visualisiert, die eine textuelle und grafische Auswahl von Datenpunkten ermöglichen. Die dynamische Anpassung von Parametern oder die Auswahl interessanter Datenuntergruppen ermöglichen eine Untersuchung des Einflusses und der gegenseitigen Abhängigkeit von Attributen. Dies erlaubt spezifische Unzulänglichkeiten zu finden, was zur Gesamtargumentation eines vertrauenswürdigen DNN-Modells verwendet werden kann.

Abbildung 7: *Der Analyst kann mit Hilfe des Tools interaktiv Teilmengen von Testdaten definieren und auswählen. Das Tool liefert Metadaten und DNN-Leistungsstatistiken für die ausgewählte Teilmenge und unterstützt interaktive Drilldowns bis auf die Ebene einzelner Bilder.*

Die Metrik benötigt folgende Eingabedaten und kann damit die unten aufgeführten Aufgabedaten erzeugen:

Eingabe in die Metrik

- Datensätze
  - Testdatensatz Datensatz Bildordner
  - Ground-Truth-Anmerkungen für Objekte (Klasse, Bounding Box)
  - Metadaten-Annotationen der Bilder und für die Objekte innerhalb der Bilder entlang der Dimensionen der ODD. Diese Metadaten-Anmerkungen können automatisch generiert werden, indem CLIP und die ODD-Definition als JSON-Datei bereitgestellt werden
- DNN-Vorhersagen pro Objekt
  - Für jedes Objekt (Fußgänger) muss die Information bereitgestellt werden, ob das DNN das Objekt erkannt hat (True Positive, TP) oder nicht erkannt hat (False Negative, FN), vorzugsweise mit Bounding-Box-Informationen über die Erkennung. Außerdem sollten vorhergesagte Bounding-Boxen ohne entsprechende Objekte der Grundwahrheit (False Positives, FPs) bereitgestellt werden.
- Alternativ kann auch der Zugriff auf die Interferenzschnittstelle des zu testenden DNN oder die DNN-Gewichte als Eingabe bereitgestellt werden.

Ausgabe der Metrik

Hypothesen bezüglich der DNN-Leistung in verschiedenen Teilmengen der Testdaten der Sicherheitsingenieure oder ML-Entwickler können mit dem VA-Tool bestätigt oder verworfen werden. Der numerische Wert der Metrik ist eine Zahl zwischen 0 und 1, definiert als das Verhältnis der Personenstunden, die das Tool von Analytikern zur Bewertung des zu testenden DNN verwendet wurde, im Vergleich zur erwarteten erforderlichen Stundenzahl, unter der Voraussetzung, dass in dieser Inspektionszeit keine inakzeptablen Leistungseinschränkungen des zu testenden DNN gefunden wurden. Wurden unannehmbare Leistungseinschränkungen festgestellt, wird der Wert der Kennzahl auf 0 gesetzt.

Abdeckung der LAISC:

- Unzureichende Spezifikation der ODD: Das Tool unterstützt den Analysten bei der Identifizierung von semantischen Regionen im Testdatensatz, in denen das zu testende DNN nicht richtig funktioniert. Diese identifizierten Regionen können mit der ODD-Spezifikation verglichen werden, was möglicherweise zu Erkenntnissen über Unzulänglichkeiten in der ODD-Spezifikation führt.
- Mangelndes Verständnis der Daten: Das Tool ermöglicht die Auswahl von Teilmengen der Daten entlang der ODD sowie weiterer Metadaten, neben der Untersuchung der Leistung des zu testenden Modells auf der ausgewählten Teilmenge. Dies hilft auch dem Menschen, die zugrunde liegenden Daten besser zu verstehen und interpretieren.
- Unzureichende Datenrepräsentation: Es können semantische Regionen oder Szenarien erkannt werden, in denen das Modell unterdurchschnittlich abschneidet (z. B. schlechte Erkennung von Kindern), was oft mit einer unzureichenden Repräsentation der semantischen Region in den Trainingsdaten zusammenhängt.

- Falsche Datenlabels: Es können semantische Regionen oder Szenarien erkannt werden, in denen das Modell unterdurchschnittlich abschneidet (z. B. schlechte Erkennung von Kindern). In den durchgeführten Analysen wurde die Verwendung unterschiedlicher „Stile" der Bounding-Box-Notation in der Ground Truth und in den Vorhersagen als systematische Ursache für FPs erkannt.
- Mangelnde Erklärbarkeit: Der Hauptvorteil des Tools besteht darin, dem Analysten eine effiziente Möglichkeit zu geben, das Modellverhalten in verschiedenen semantischen Regionen und Szenarien besser zu verstehen. Das Tool hilft unter anderem dabei, Modellverhalten zu erkennen, das auf falschen Korrelationen beruht. Zum Beispiel betrifft dies False Positives (FPs), die auf eisenbahnspezifischen Infrastrukturelementen generiert werden.
- Mangel an Robustheit: Das Tool hilft bei der Identifizierung von Fällen inkonsistenter Leistung des Modells und trägt somit zur Robustheit als kontrafaktische Analyse bei. Die Anwendung im Benchmarking-Abschnitt zeigt Beispiel Bilder, in denen nebeneinanderstehende Personen auf demselben Bild inkonsistent erkannt werden (einige werden erkannt, andere nicht).

II.1.2.2.1 Bewertung der safe.trAIn-Modelle

Die Metrik wurde auf das in safe.trAin verwendete-Personendetektormodell auf der Grundlage des Panoptic FCN angewandt. Dieses Modell wurde an einer Teilmenge des Railsem19-Datensatzes analysiert, wobei der Schwerpunkt auf den Bildern des ursprünglichen Railsem19-Datensatzes lag, die Personen enthalten. Der Testdatensatz wurde um Bounding Boxes für Fußgänger und Metadaten-Annotationen für Fußgänger erweitert. Als ersten Einstiegspunkt liefert das Tool eine Gesamtstatistik für den gesamten Testdatensatz.

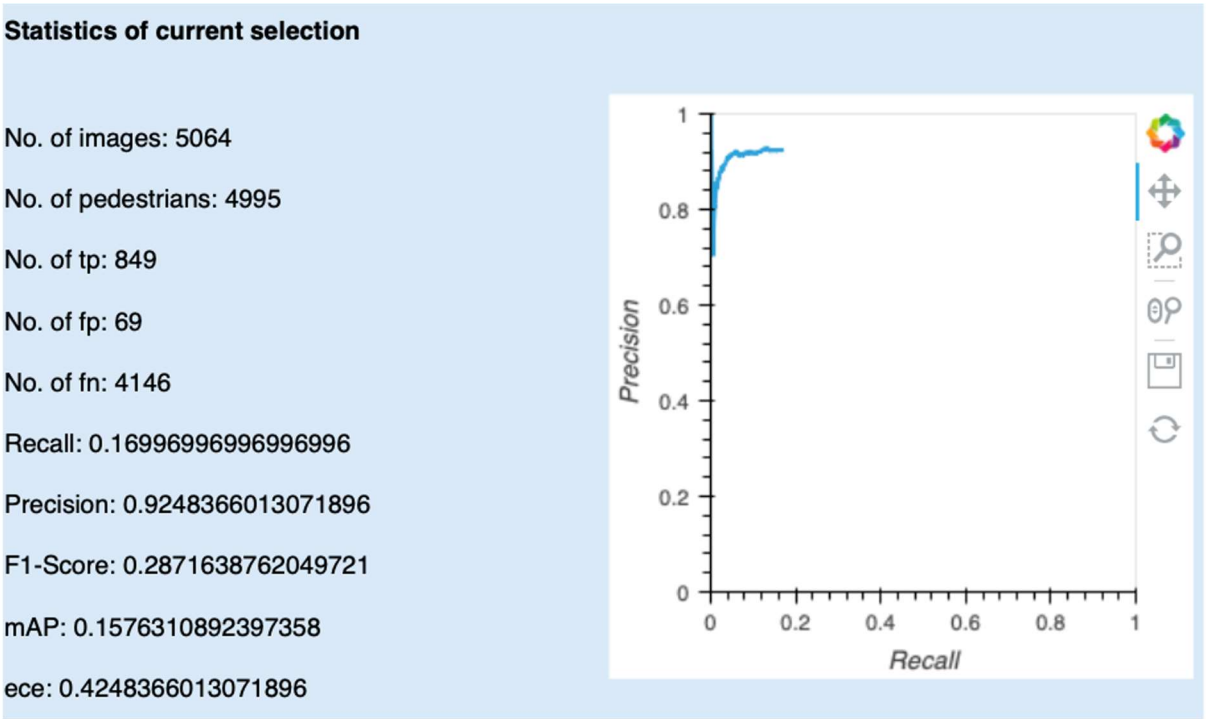


Abbildung 8: *Überblick der Gesamtstatistik für den Testdatensatz*

Der Überblick in Abbildung 8 zeigt, von den 4995 vorhandenen Fußgängern wurden nur 849 erkannt, was zu einer Gesamterkennung von 16,9 % führte. Darüber hinaus wurden 69 FPs generiert, was zu einer Präzision von 92,4 % führt. Ausgehend von diesen Gesamtwerten wurden mit dem Tool weitere Analysen durchgeführt, die zu den folgenden detaillierten Ergebnissen geführt haben. In allen folgend dargestellten Visualisierungen stehen die orangefarbenen Bounding Boxes für FNs, die grünen Bounding Boxes für TPs und die violetten Bounding Boxes für FPs.

Die Mindestgröße eines erkennbaren Fußgängers muss bei diesem Ansatz als Parameter definiert werden, z. B. Höhe > 30 Pixel. Wenn Personen nur wenige Pixel in einem Bild einnehmen, dann ist nicht zu erwarten, dass das DNN sie perfekt erkennt. Andererseits können auch weit entfernte oder kleine Personen sicherheitsrelevant sein. Die Festlegung der Mindestpixelgröße von Personen, die erkannt werden müssen, erfordert somit Diskussionen und Erkenntnisse sowohl von Sicherheitsexperten als auch von KI-Experten. Das Tool ermöglicht dann flexibel die Auswahl von Personengrößen auf der Grundlage der definierten Mindestpixelgröße im Bild. Weiter wurde angenommen, dass Personen mit einer Größe von 36 Pixeln nicht übersehen werden sollten. Andererseits können Personen mit einer Körpergröße von weniger als 30 Pixeln auch unter dem Gesichtspunkt der Sicherheit übersehen werden.

Folgende Beispiel-Ergebnisse und -Erkenntnisse konnten gewonnen werden:

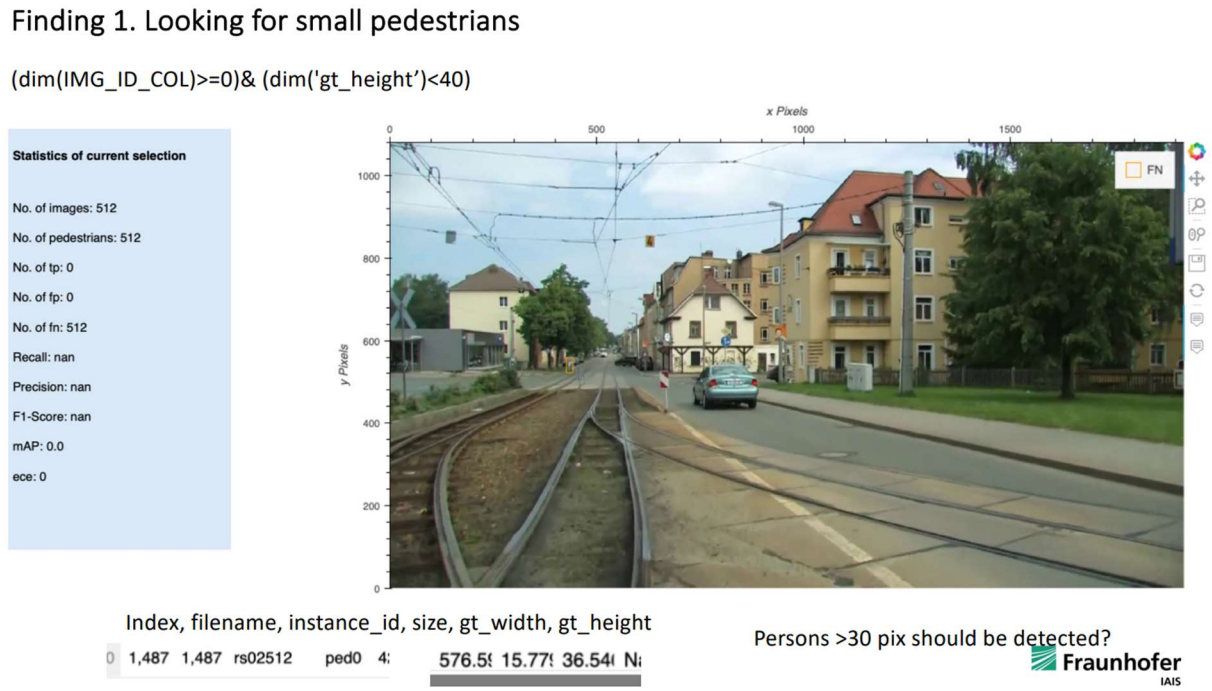


Abbildung 9: *Bestimmung des Schwellenwerts für die Größe der Fußgängerpixel, ab dem Fußgänger erkannt werden sollten*

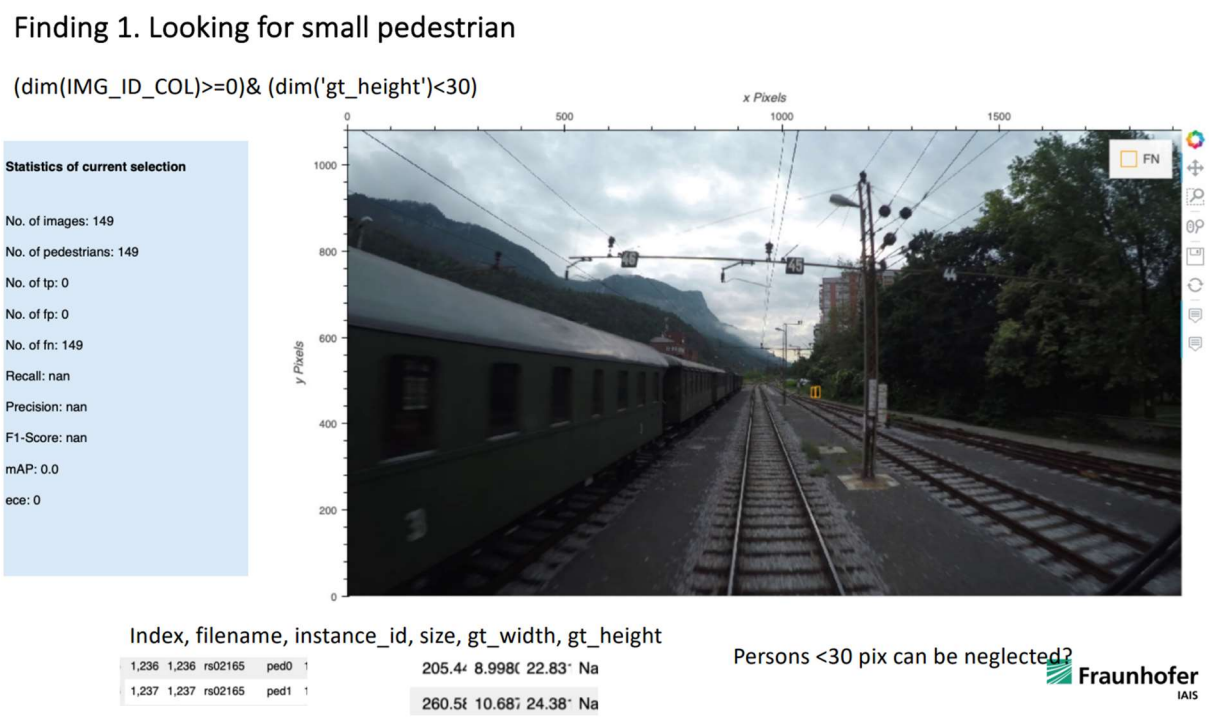


Abbildung 10: *Können kleinere Fußgänger (oder Fußgänger in einer gewissen Entfernung) ignoriert werden?*

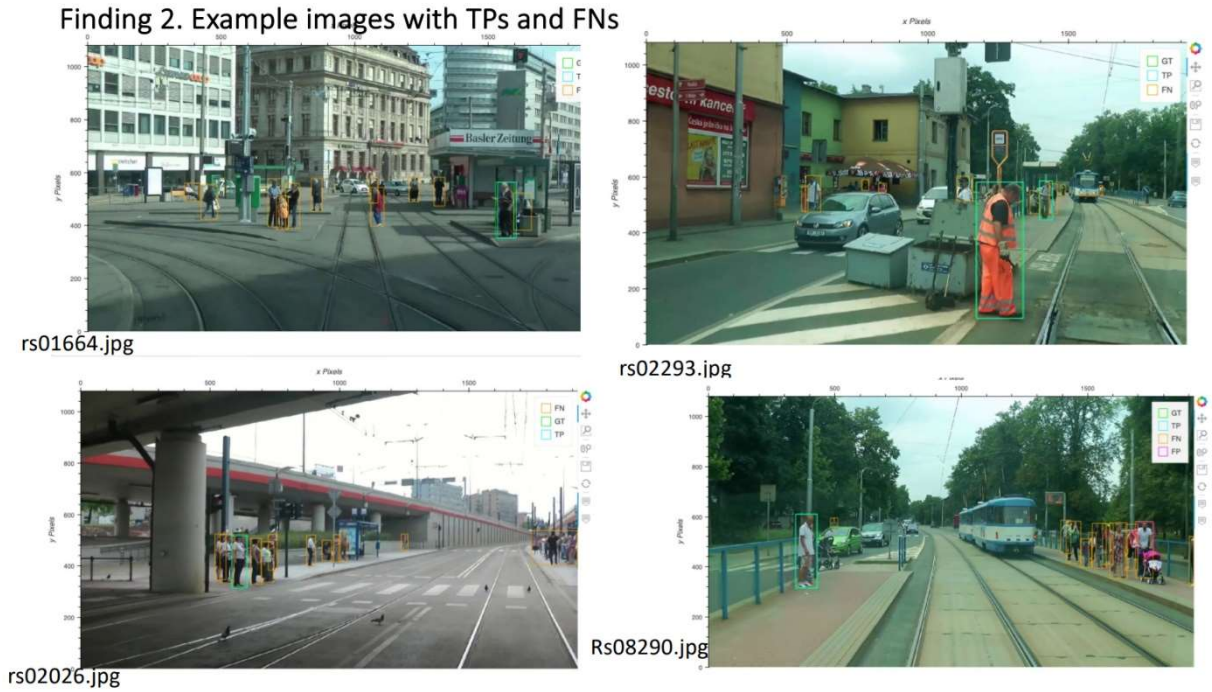


Abbildung 10: *Unterschiedlich gute Erkennung von Personen in Beispielen der Testdaten.*

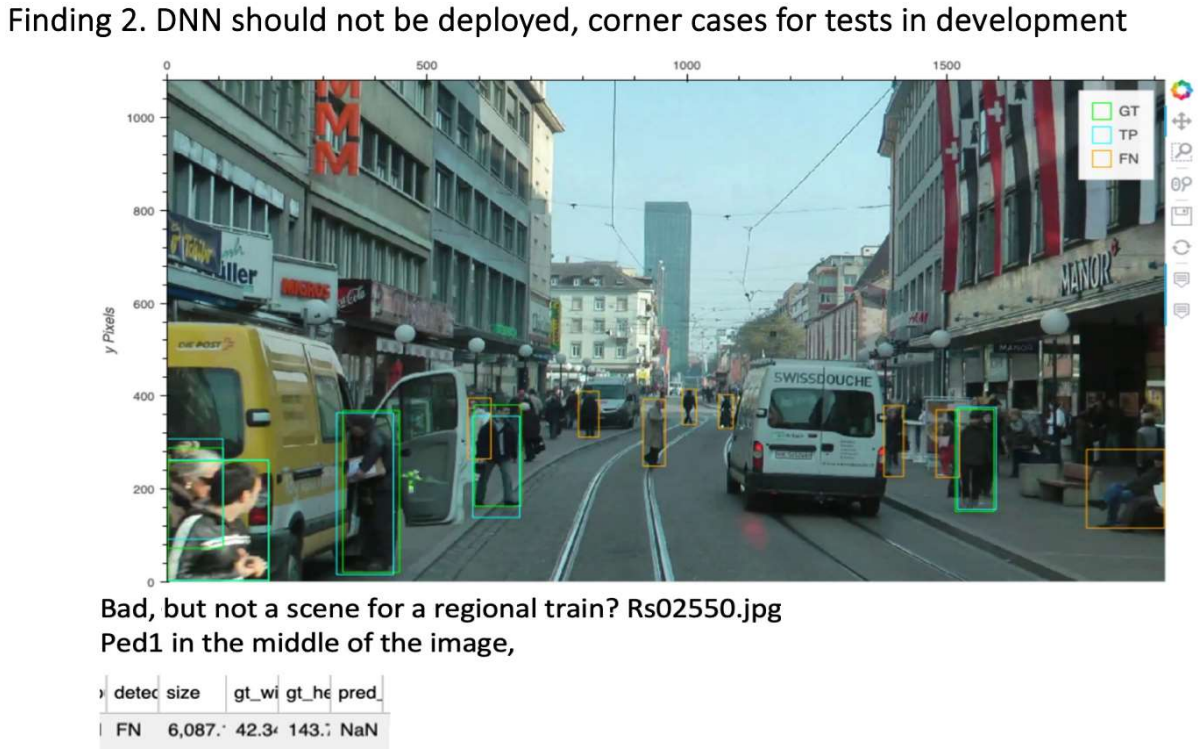


Abbildung 11: *In dicht bevölkerten Straßenszenen wurden mehrere Fußgänger nicht erkannt.*

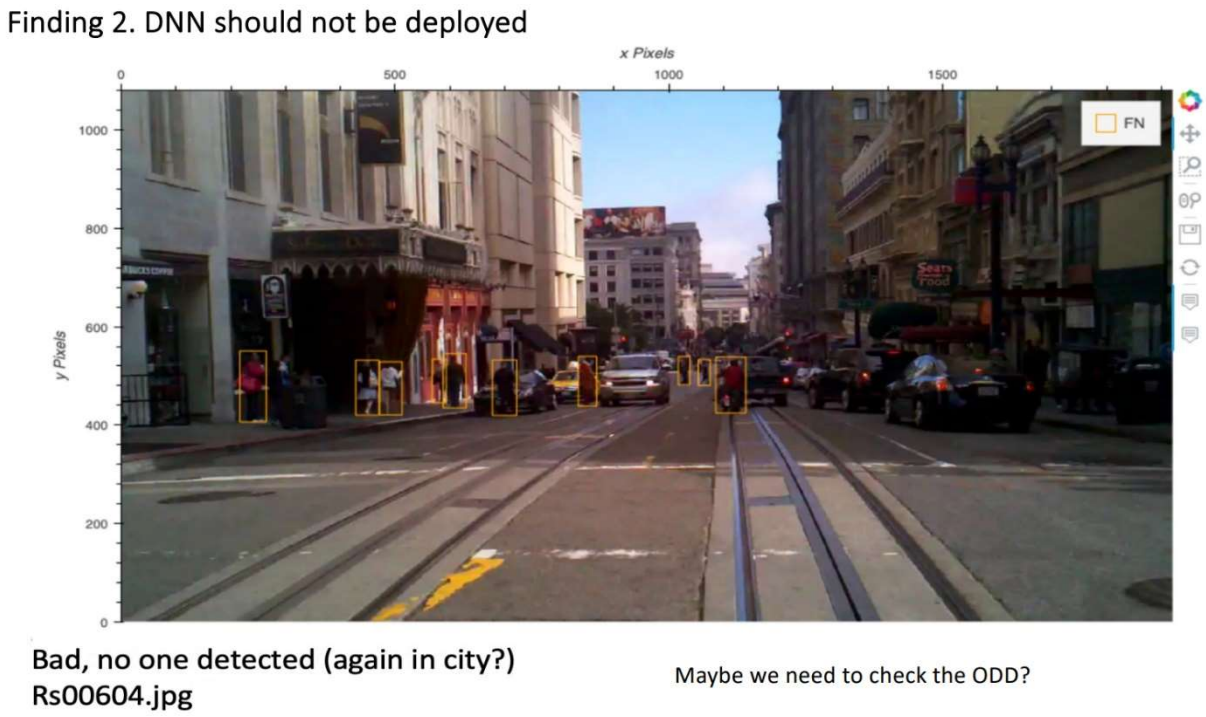


Abbildung 12: *Bei bestimmten Szenen werden keine Fußgänger erkannt*

Insgesamt konnten aus der Analyse mit dem ScrutinAI Tool folgende Schlüsse gezogen und Rückmeldungen gegeben werden:

- Es ist wichtig, zwischen betroffenen Personen und eher unrelevanten Personen, also etwa Personen, die sich weit entfernt vom Bahngleis befinden, in der Berechnung der Metrik zu unterscheiden. Das nötige Domänenwissen müssen Sicherheitsingenieure beisteuern, die beurteilen, welche Personen als „sicherheitsrelevante Fußgänger" erkannt werden sollten.
- Als Rückmeldung und die Entwickler würden Grenzfälle identifiziert, die nächsten Entwicklungsschritt zu berücksichtigen sind. (Feedback an Entwickler)
- Bei der Definition der ODD sollte eventuell berücksichtig werden, dass das Modell im innerstädtisch Bereich Schwächen aufweist. Diese Bereiche könnten aus der ODD ausgenommen werden. (Feedback an Sicherheitsingenieure)

II.1.2.2.2 Fazit

Die Anwendung von ScrutinAI-ähnlichen semantischen Analysewerkzeugen im Eisenbahnbereich bietet ein vielversprechendes Potenzial, um die Verlässlichkeit und Transparenz KI-basierter Systeme zu bewerten. In sicherheitskritischen Anwendungen wie der Zugerkennung, Weichenzustandsdiagnose oder Hinderniserkennung ist es nicht ausreichend, dass ein Modell korrekte Vorhersagen trifft. Vielmehr muss nachvollziehbar sein, ob diese Vorhersagen auf semantisch relevanten Informationen beruhen, z. B. spezifischen technischen Komponenten, sicherheitsrelevanten Objekten oder Umgebungsmerkmalen.

Ein solcher Ansatz kann als zusätzliche Metrik zur Bewertung von KI-Systemen dienen, indem er aufzeigt, ob die Vorhersagequalität mit einer semantisch sinnvollen Entscheidungsgrundlage korreliert. Dies kann insbesondere Audits, Zertifizierungen und kontinuierliche Modellüberwachung im Betrieb unterstützen. Die systematische Überprüfung semantischer Kohärenz bietet somit nicht nur eine Möglichkeit zur Fehleranalyse, sondern trägt auch zur Erhöhung des Vertrauens und der Nachvollziehbarkeit von KI-gestützten Entscheidungsprozessen im Eisenbahnwesen bei.

II.1.2.3 Prototype based Out-of-Domain Detection without Labels

Um verschiedene Sicherheitsaspekte im Zusammenhang mit verschiedenen KI-Komponenten des Systems zu quantifizieren, ist die Erkennung von sogenannten Out-Of-Domain (OOD)-Objekten im Gleisbereich sehr relevant. Die im Folgenden vorgestellte und von Fraunhofer entwickelte Methode hat zum Ziel, eine unabhängige Metrik zur Erkennung von OOD-Objekten zu geben. Die Methode, namens *PROWL (PRototype based OOD detection Without Labels)*, ermöglicht eine Zero-Shot-Erkennung und Segmentierung von OOD-Objekten über die definierten Listenobjekte in der Operation Design Domain (ODD) zur Laufzeit. Die Methode und Ergebnisse wurden auf der internationalen Konferenz WACV 2025 veröffentlicht.

II.1.2.3.1 Überblick über PROWL

Das grundlegende Ziel dieser Forschungsarbeiten war die Entwicklung einer Methode zur Erkennung von Objekten außerhalb der Domäne (OOD). Diese Methode soll leicht auf jede neue Szene angewendet werden können. Dies wird verfolgt, indem die in der Szene erwarteten ODD-Objekte spezifiziert werden, ohne dass jedoch eine große Menge an gelabelten Daten oder überwachtes Training erforderlich ist. Die Methode stützt sich dabei auf sogenannte *Foundation-*

Modell-basierte Merkmalsextraktoren, wie beispielsweise DINOv2[5], und arbeitet mit *Zero-Shot-Inferenz*. Das heißt, das Modell klassifiziert Beispiele aus Klassen, die während des Trainings nicht beobachtet wurden. Die Methode benötigt lediglich *stellvertretende Prototypen* per ODD-Objekt, diese basieren auf nur wenigen Instanzmasken für jede relevante Objektkategorie. Diese Offline-Prototyp-Merkmalsbank wird dann zur Laufzeit mit allen detektieren Objekten abgeglichen.

Das Schema für die Architektur von PROWL ist in Abbildung 13 dargestellt.

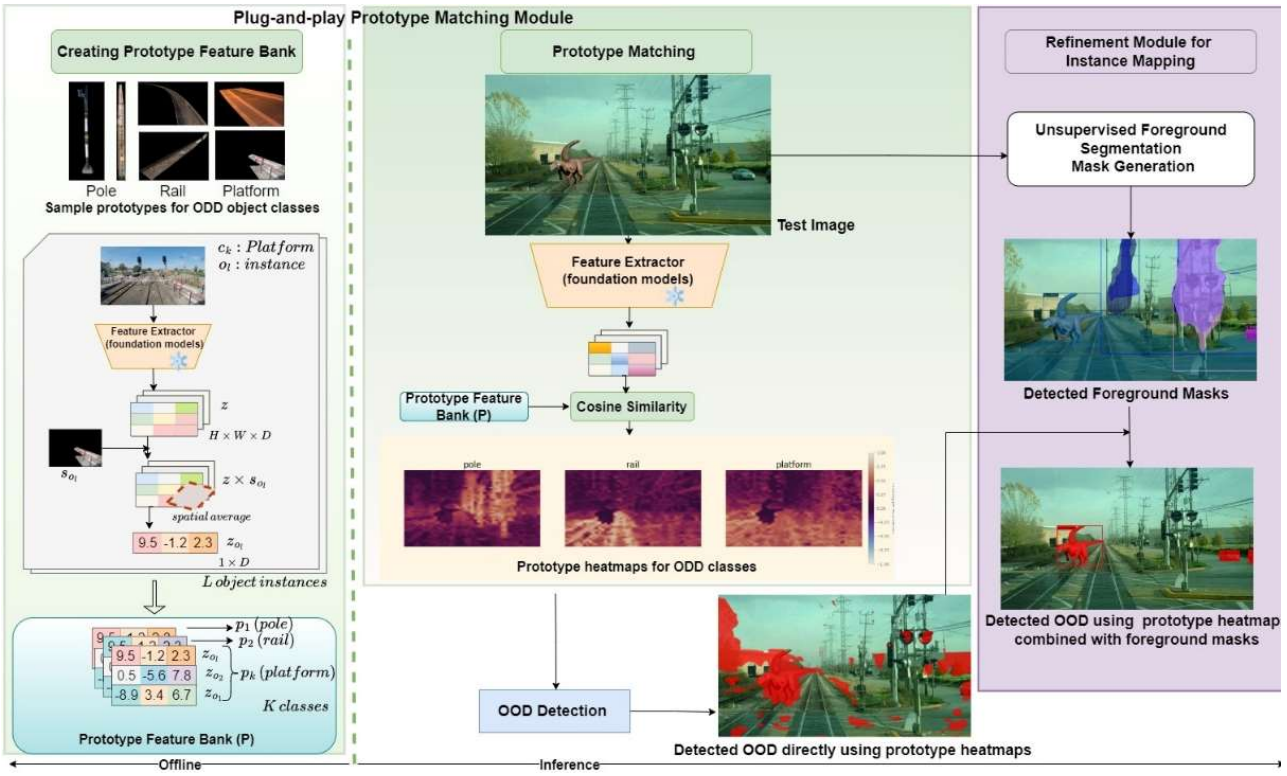


Abbildung 13: *Schematische Darstellung der von Fraunhofer IKS entwickelten Methode PROWL.*

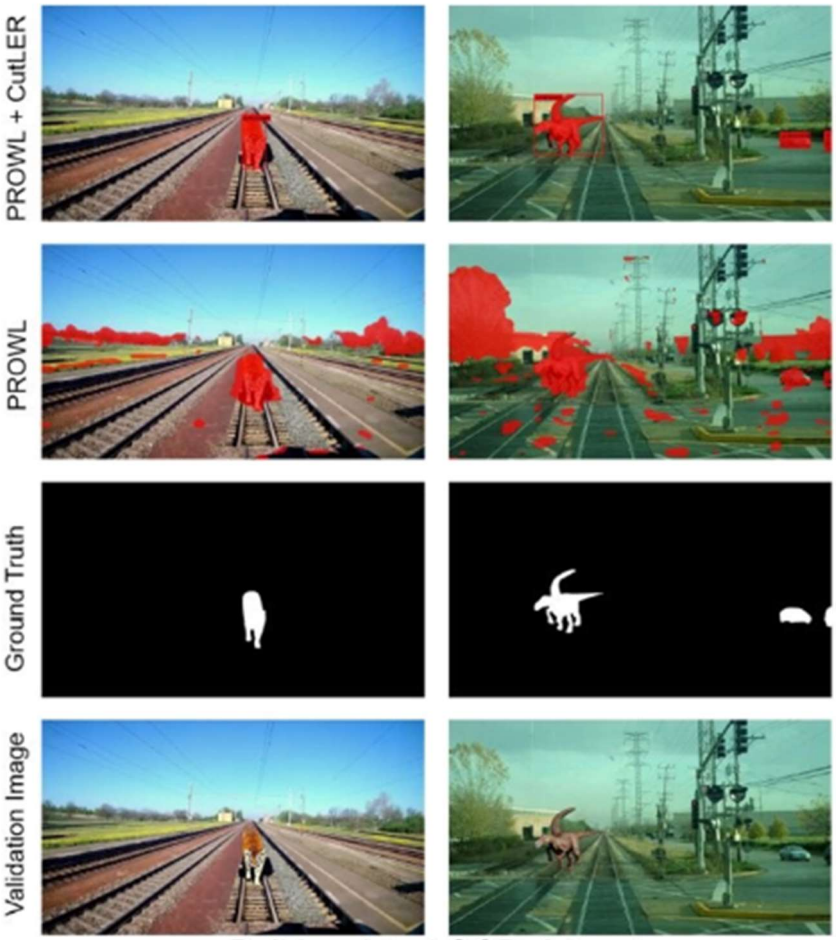
Im Folgenden wird die Modellarchitektur von PROWL zusammenfassend beschrieben:

1. Zunächst wird eine Offline-Prototyp-Merkmalsdatenbank für eine Liste von Objektklassen erstellt, die im ODD definiert sind. Das geschieht, indem wir eine Inferenz auf 5-10 Beispielbildern auf dem Basismodell (DINOv2) durchführen. Dadurch erhalten wir einen 1D-Merkmalsvektor pro Prototypenobjekt (gelernt aus den 5-10 Beispielbildern), welcher eine Instanzsegmentierungsannotation der gegebenen Klasse (z.B. Spurannotation) maskiert und somit deren wichtigsten Merkmale enthält bzw. repräsentiert.
2. Nach der Durchführung der Inferenz auf DINOv2 für das gesamte Bild wird jeder einzelne Pixel einer Objektklasse auf der Grundlage der maximalen Kosinusähnlichkeit mit den Prototypmerkmalen (d.h., 1D-Merkmalsvektor) einer der Objektklassen zugeordnet. Wenn der Ähnlichkeitswert unter einem vorher festgelegten Schwellenwert fällt, wird der Pixel als OOD-unbekannte Klasse eingestuft. In Abbildung 14 kann dies

[5] https://dinov2.metademolab.com/

beispielsweise durch rot markierte Pixel/Objekte erkannt werden. Der Dinosaurier wird somit als OOD-Objekt detektiert.

3. Die Ausgabe wird weiter verbessert, indem die OOD-Objekte auf Maskenebene statt auf Pixelebene bestimmt werden. Das passiert durch einen unüberwachten Segmentierungsmaskengenerator (z. B. CutLER[6]). Dieser hilft dabei, OOD-Objekte zu finden, die zu plausiblen Objektinstanzen gehören. Das finale Resultat des Beispiels ist in Abbildung 14 (oben rechts) zu sehen: Die endgültig erkannten OOD-Objekte sind der Dinosaurier und zwei Autos, da diese im Beispiel nicht im ODD definiert waren und somit keine Prototypenklasse besitzen.



4.

5. Abbildung 14: *Qualitative Ergebnisse für die Erkennung von OOD-Objekten im Schienenverkehr*[7]

[6] https://github.com/facebookresearch/CutLER

[7] Da keine öffentlich zugänglichen realen OOD-Daten verfügbar waren, wurde ein „Inpainted OOD"-Datensatz erstellt, um die Leistung von PROWL zu bewerten. Erkannte OOD-Pixel oder Segmentierungsmasken sind im Bild in roter Farbe dargestellt.

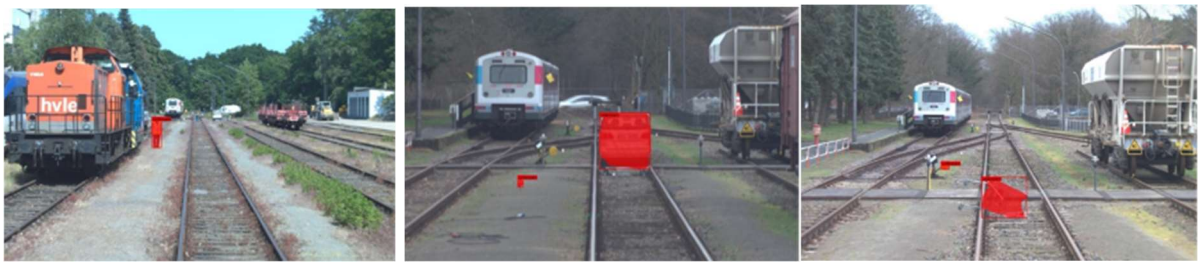
Hinsichtlich der in safe.trAIn betrachteten LAISC kann die Methode die folgenden KI-inhärenten Schwächen verbessern:

- (Nr. 13) **Unzureichende Datenbasis**: PROWL hilft seltene, unbekannte Grenzfälle zu identifizieren und somit ein Fehlerverhalten zu vermeiden
- (Nr. 23) **Datendrift**: PROWL hilft kleine, aber kontinuierliche, Veränderungen in der Einsatzumgebung zu identifizieren

II.1.2.3.2 Ergebnisse und Evaluierung

Um PROWL zu evaluieren, wurden umfangreiche Untersuchungen in mehreren ODD-Domänen auf öffentlich verfügbaren Datensätzen mit echten OOD-Objekten durchgeführt. Im Rahmen von safe.trAIn wurden relevante ODD-Objekte aus RailSem19-Daten extrahiert, und zur Erstellung der prototypischen Feature-Datenbank verwendet. Aufgrund der begrenzten Verfügbarkeit von Datensätzen mit realen OOD-Objekten im Bahnbereich wurden die Evaluierungsergebnisse für die synthetischen OOD-Datensätze zusammen mit intern generierten Datensätzen verwendet.

Zusätzlich wurden auch Auswertungen mit innerhalb von safe.trAIn bereitgestellten HVLE-Daten durchgeführt. Die Untersuchung verwendet einen generierten Datensatz der HVLE-Daten mit entsprechenden OOD-Objekten. Folgende qualitative Ergebnisse wurden mit PROWL erzielt:



(a) (b) (c)

Abbildung 15*: Qualitative Ergebnisse für die Erkennung von OOD-Objekten mithilfe von PROWL auf dem HVLE-Datensatz*[8]

Hindernisse auf oder neben der Strecke, die nicht Teil der ODD sind, wurden mit PROWL zuverlässig als OOD erkannt. Weiter fällt auf, dass auch zusätzliche, nicht in der ODD definierte Objekte erkannt werden. Zum Beispiel der sehr kleine Signalgeber zwischen den beiden Gleisen (vgl. Abbildung 15 (b), (c)). In beiden Fällen ist dieser Signalgeber nicht explizit Teil der definierten ODD-Klassen und wird daher korrekterweise als Out-of-domain Objekt erkannt. Das ist unter anderem darauf zurückzuführen, dass die Prototypen der Merkmalsdatenbank auf realen Bildern aus der ODD basieren. Diese stimmen oft nicht mit den in HVLE-Daten verwendeten Dummy-

[8] .Unbekannte Objekte werden korrekt als OOD-Objekte erkannt. Siehe: (a) Person, (b) Auto auf dem Gleis, (c) Einkaufswagen auf dem Gleis. Erkannte OOD-Objekte sind in Rot dargestellt.

Objekten überein. Dies verdeutlicht insbesondere die Notwendigkeit, realistische Datensätze für die Entwicklung über die Forschung hinaus zu verwenden.

II.1.3 AP3 Fahrzeugarchitektur im GoA4-Betrieb mit dem Fokus auf sichere KI-basierte Funktionen

Im Rahmen des AP3 wurde von Fraunhofer die Entwicklung der Operational Design Domain (ODD) als ein zentrales Element für die Sicherheitsarchitektur fahrerloser Züge vorangetrieben.

II.1.3.1 Operational Design Domain (ODD)

Die ODD beschreibt die Betriebsbedingungen, unter denen ein automatisiertes System oder eine Funktion speziell konzipiert ist, um zu funktionieren. Dazu zählen Umweltbedingungen, geografische Einschränkungen sowie zeitliche Faktoren und die notwendige Präsenz oder Abwesenheit bestimmter Infrastrukturelemente.

ODD ist in diesem Zusammenhang das zusammenfassende Konzept, das eine klare Struktur für die Definition und Verwaltung der Betriebsbedingungen bietet, in denen KI-Systeme sicher und effektiv betrieben werden können. ODD unterstützt dabei, dass KI-Systeme von Anfang an unter Berücksichtigung der Sicherheit (im Sinne von Safety) konzipiert werden und auch bei sich ändernden Betriebsanforderungen zuverlässig bleiben.

Die ODD kann als zentrales Element während des gesamten Entwicklungs- und Betriebsprozesses gesehen werden (s. Abbildung 16). Hierbei ist die Einbeziehung aller Beteiligten während des Entwicklungsprozesses entscheidend für die Transparenz und das Vertrauen. Zu diesem Zwecke werden die Fähigkeiten des Systems mit den Erwartungen der Nutzer und den Einsatzumgebungen abgeglichen. Zudem ist eine kontinuierliche Überwachung auch nach der Inbetriebnahme entscheidend, um Sicherheitsprobleme bzgl. den definierten Betriebsbedingungen zu erkennen und zu beheben. Die Definition von Betriebsumgebungen erleichtert auch das Testen, indem sie die Erzeugung relevanter und vollständiger Testszenarien ermöglicht. Gründliche Tests, die sowohl Simulationen als auch reale Szenarien umfassen, sind notwendig, um potenzielle Risiken vor der vollständigen Einführung eines KI-Systems zu erkennen und in ausreichendem Maß zu verringern. Eine hohe Repräsentativität der Daten hinsichtlich der ODD ist dabei entscheidend für die Qualität des KI-Systems. In safe.trAIn wurde ein iterativer Entwicklungsansatz verfolgt, der Feedback und neue Erkenntnisse während der Entwicklung und Erprobung berücksichtigt. Ziel ist die Performanz und Fähigkeiten des KI-Systems zu erweitern und optimieren und dabei die Sicherheit innerhalb der definierten Betriebsumgebungen zu gewährleisten.

Um Betriebsbedingungen zu beschreiben, hat sich aus dem Automotive-Bereich stammend das Konzept der Operational Design Domain (ODD) etabliert. Grundsätzlich kann bei der Definition und Verwendung der ODD zwischen einer ODD-Taxonomie und ODD-Beschreibung unterschieden werden.

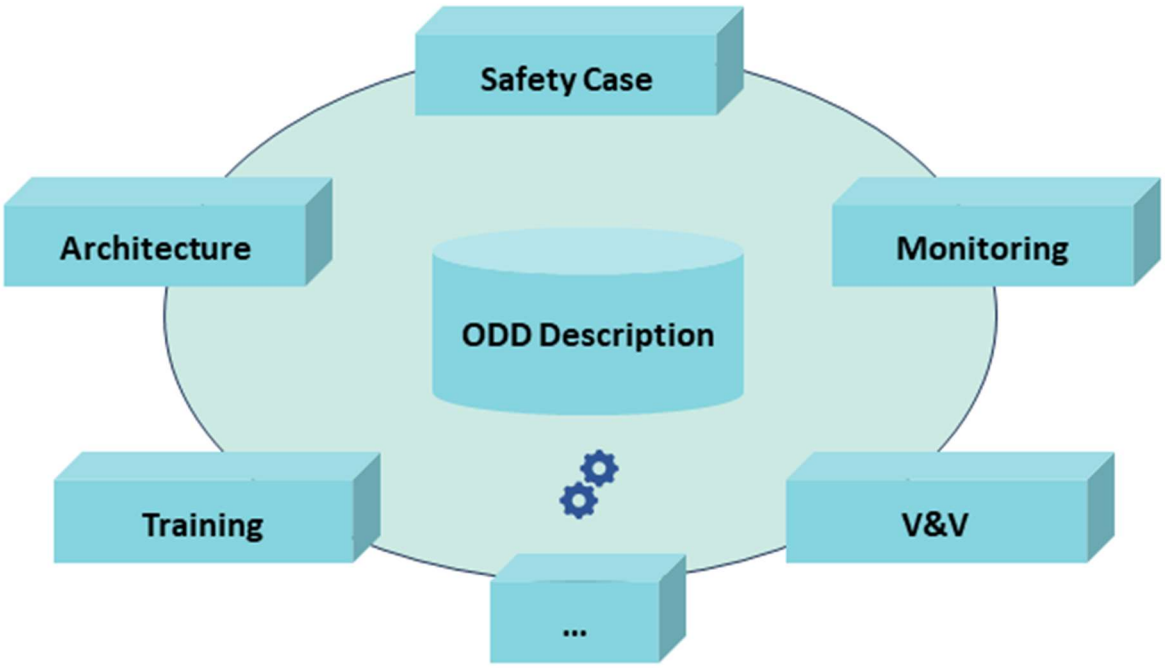


Abbildung 16 *ODD (Operational Design Domain) ermöglicht die Rückverfolgbarkeit und Konsistenz über verschiedene Entwicklungsdisziplinen hinweg.*

Die im Projekt entwickelte ODD-Taxonomie stellt ein Klassifikationssystem dar, das die möglichen Betriebsbedingungen und Einschränkungen von Schienenfahrzeugen kategorisiert und organisiert. Sie bietet einen strukturierten Rahmen zur Definition messbarer Dimensionen der Umgebungen, in denen der vollautomatisierte Zug betrieben werden soll.

Zweck und Merkmale der ODD-Taxonomie:

- **Klassifikation von Betriebsbedingungen**: Die ODD-Taxonomie klassifiziert verschiedene Betriebsbedingungen wie Wetterverhältnisse, Infrastruktur und Geschwindigkeitsbegrenzungen, denen der fahrerlose Zug begegnen kann.
- **Messbare Dimensionen**: Die Taxonomie legt messbare Dimensionen der Kategorien fest, sowohl quantitativ für physikalische Werte als auch qualitativ für allgemeinere Konzepte.
- **Hierarchische Struktur**: Die ODD-Taxonomie folgt einer hierarchischen Struktur, die Betriebsbedingungen und Einschränkungen in Kategorien und Unterkategorien organisiert, um Eindeutigkeit und Verständlichkeit zu gewährleisten.
- **Generalisierte Darstellung**: Die Taxonomie bietet eine generalisierte Darstellung der Betriebsbedingungen und Einschränkungen und definiert allgemein die Betriebsbedingungen für ein Schienenfahrzeug, ohne sich auf spezifische Einzelfälle zu konzentrieren.

Im Gegensatz zur Taxonomie stellt eine ODD-Beschreibung einen spezifischen Kontext dar, der die in der ODD-Taxonomie definierten Kategorien und Dimensionen verwendet. Sie erfasst die einzigartigen Merkmale, Parameter und kontextuellen Informationen eines bestimmten Betriebsumfeldes.

Zweck und Merkmale der ODD-Beschreibung:

- **Spezifische Darstellung**: ODD-Beschreibungen bieten eine spezifische Darstellung einer ODD und erfassen deren relevanten Merkmale und Umweltfaktoren.
- **Instanziierung der ODD**: Eine ODD-Beschreibung repräsentiert eine konkrete ODD und hebt die besonderen Bedingungen, Variationen und Herausforderungen hervor, die in diesem Kontext auftreten oder erwartbar sind.
- **Detaillierte Informationen**: ODD-Beschreibungen enthalten detaillierte Informationen über spezifische Attribute, Parameter und Variationen des jeweiligen Kontexts, wie Wetterbedingungen, Infrastruktur und spezifische Umweltfaktoren.
- **Instanzspezifische Einschränkungen**: ODD-Beschreibungen können zusätzliche Einschränkungen oder Annahmen definieren, die für den spezifischen Kontext gelten und über die generalisierte Darstellung der ODD-Taxonomie hinausgehen.

Zusammenfassend sind die ODD-Taxonomie und die ODD-Beschreibung zwei voneinander zu unterscheidende Konzepte, die für die Erfassung des geplanten Einsatzbereichs autonomer Züge verwendet werden. Die ODD-Taxonomie bietet ein strukturiertes Klassifikationssystem, um die allgemeinen Betriebsbedingungen und Einschränkungen zu definieren, die generell für vollautomatisierte Züge zu erwarten sind. Aufbauend darauf definiert eine ODD-Beschreibung eine spezifische ODD und erfasst die einzigartigen Merkmale, Parameter und Variationen eines bestimmten Betriebsumfeldes. Beispielsweise kann eine ODD-Beschreibung einen konkreten Wertebereich beschreiben. in dem ein sicherer Betrieb gewährleistet wird.

ODD Roadmap und Entwicklung

Fraunhofer hat im Rahmen des safe.trAIn Projekts maßgeblich zur Entwicklung der ODD-Taxonomie und ODD-Beschreibung für den Bahnbereich beigetragen. Zu Beginn wurde eine Roadmap für die Erstellung der ODD für das Projekt aufgestellt (vgl. Abbildung 17) und deren Umsetzung kontinuierlich überwacht und vorangetrieben.

Abbildung 17 Roadmap für die Erstellung der ODD im Rahmen des Projekts.

Eine initiale Sammlung hat einen Überblick über aktuelle Ansätze aus anderen Domänen für die Definition einer ODD geliefert (insbesondere Vorarbeiten aus dem Automotive-Bereich) sowie über Ansätze zur strukturierten Beschreibung des Bahnbereichs (beispielsweise mit Ontologien). Anhand dessen wurde von Fraunhofer ein erster Prototyp der ODD-Beschreibung erstellt. Dabei wurde der Prozess von Erfahrungen und Berichten aus anderen Projekten, die ODDs entwickelt haben abgeleitet (wie z. B. KI-Absicherung[9]). Durch die Organisation und Umsetzung verschiedener Stakeholder-Workshops mit Expertengruppen ermittelte Fraunhofer die Bedürfnisse und Erwartungen an die ODD (vgl. Abbildung 19). Das erhaltene Feedback wurde in iterativen

[9] Projekt KI-Absicherung, https://www.ki-absicherung-projekt.de/

Anpassungen und Erweiterungen der ODD sowie in der Methodik und der dafür erstellten Werkzeuge integriert. Für die iterative Anpassung der ODD wurde darüber hinaus regelmäßig Feedback von Repräsentanten der Expertengruppen eingeholt (vgl. Abbildung 18).

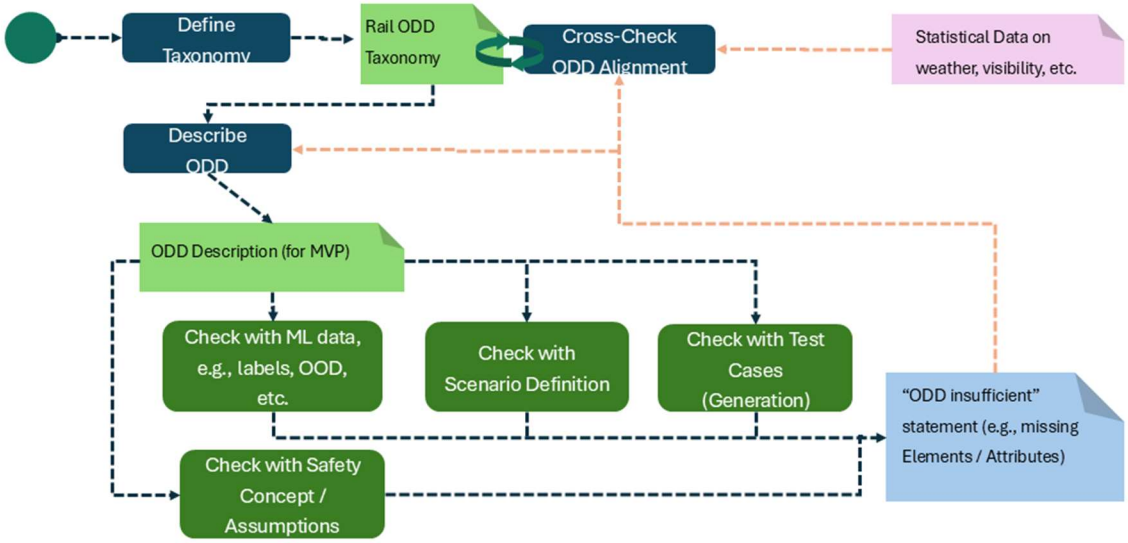


Abbildung 18: *Ablauf der Prüfung und Weiterentwicklung der ODD im Projekt*

Die Arbeiten von Fraunhofer konzentrierten sich zunächst auf die Definition der ODD als Taxonomie, die eine strukturierte Klassifikation der Betriebsbedingungen ermöglicht. Fraunhofer hat die Taxonomie kontinuierlich weiterentwickelt und bahnrelevante Konzepte in Absprache mit den Projektpartnern integriert. Darauf aufbauend wurde eine ODD-Beschreibung, die spezifische Kontexte innerhalb der Taxonomie abbildet, ebenfalls teilweise parallel von Fraunhofer für das Projekt vorangetrieben.

Für die Qualität der Ergebnisse spricht, dass die erarbeitete ODD-Taxonomie als Ausgangpunkt für die Standardisierung als DIN DKE SPEC 99004[10] „Spezifikation von ODD im Schienenverkehr" mit verschiedenen europäischen Experten verwendet werden konnte. Fraunhofer hat in diesem Rahmen die ODD-Taxonomie weiter überarbeitet und mit internationalen Experten aus der Bahn-Forschung und -Industrie außerhalb des Projekts weiterentwickelt.

[10] DIN DKE SPEC 99004, https://dx.doi.org/10.31030/3610746

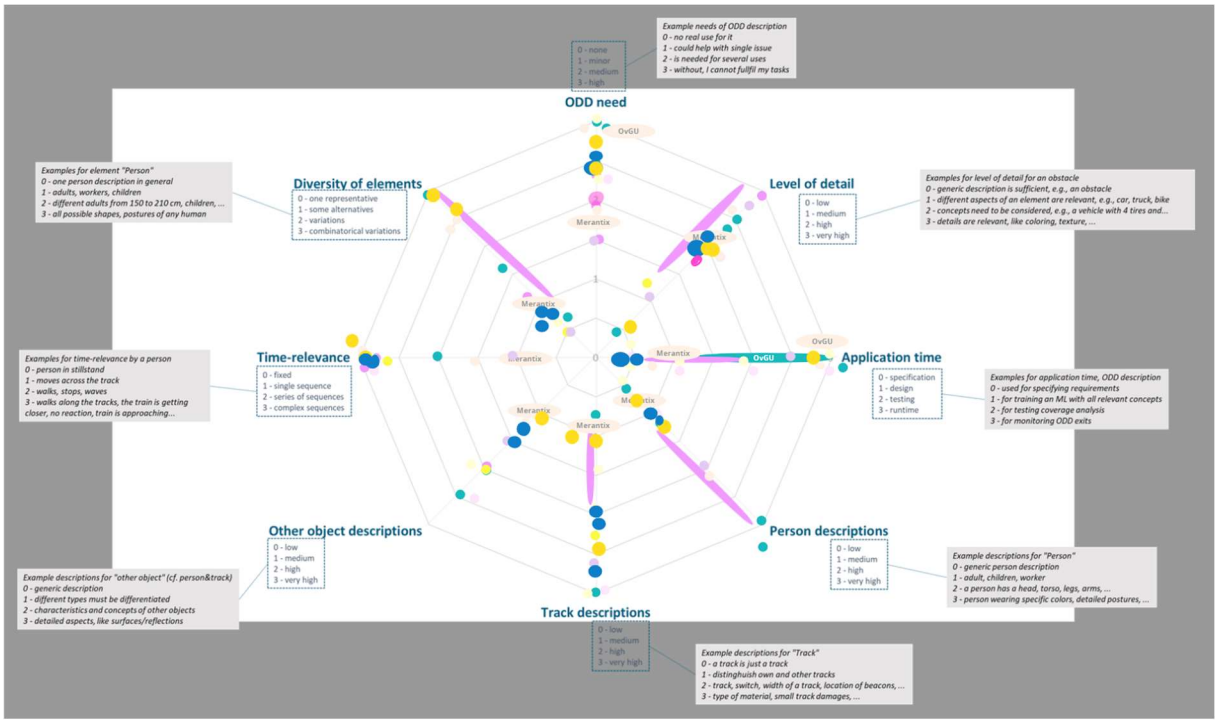


Abbildung 19: *Zusammenfassung des initialen Feedbacks der Stakeholder in den ersten Workshops zur ODD*

ODD Workbench und Visualisierung

Fraunhofer entwickelte im Rahmen von safe.trAIn als Software-Toollösung eine sogenannte ODD-Workbench (s. Abbildung 20), um ODD-Taxonomien und -Definitionen strukturiert festzulegen und zu verwalten – da zum Projektstart keine speziellen Werkzeuge für deren Erstellung existierten. Inzwischen ist mit OpenODD[11] im Automotive Bereich ein Beschreibungskonzept für ODDs entstanden, Werkzeuge sind aber weiterhin nicht allgemein verfügbar und für den spezifischen Automotive-Kontext ausgelegt.

Die Workbench verwendet ein eigenes Modell, um die Maschinenlesbarkeit zu gewährleisten und eine schrittweise Beschreibung verschiedener Aspekte in Abstraktionsebenen zu fördern. Eine solche Taxonomieebene kann auf andere Ebenen verweisen und diese verfeinern, sodass eine modulare Beschreibung möglich ist. So kann beispielsweise eine Ebene allgemeine Informationen über Personen abdecken, während sich eine andere auf Infrastrukturgebäude konzentriert. Eine anwendungsfallspezifische Ebene kombiniert Konzepte aus diesen Ebenen zu einer einheitlichen Taxonomie und fügt Attribute hinzu, die für bestimmte Anwendungen spezifisch sind.

[11] ASAM OpenODD, https://www.asam.net/standards/detail/openodd/

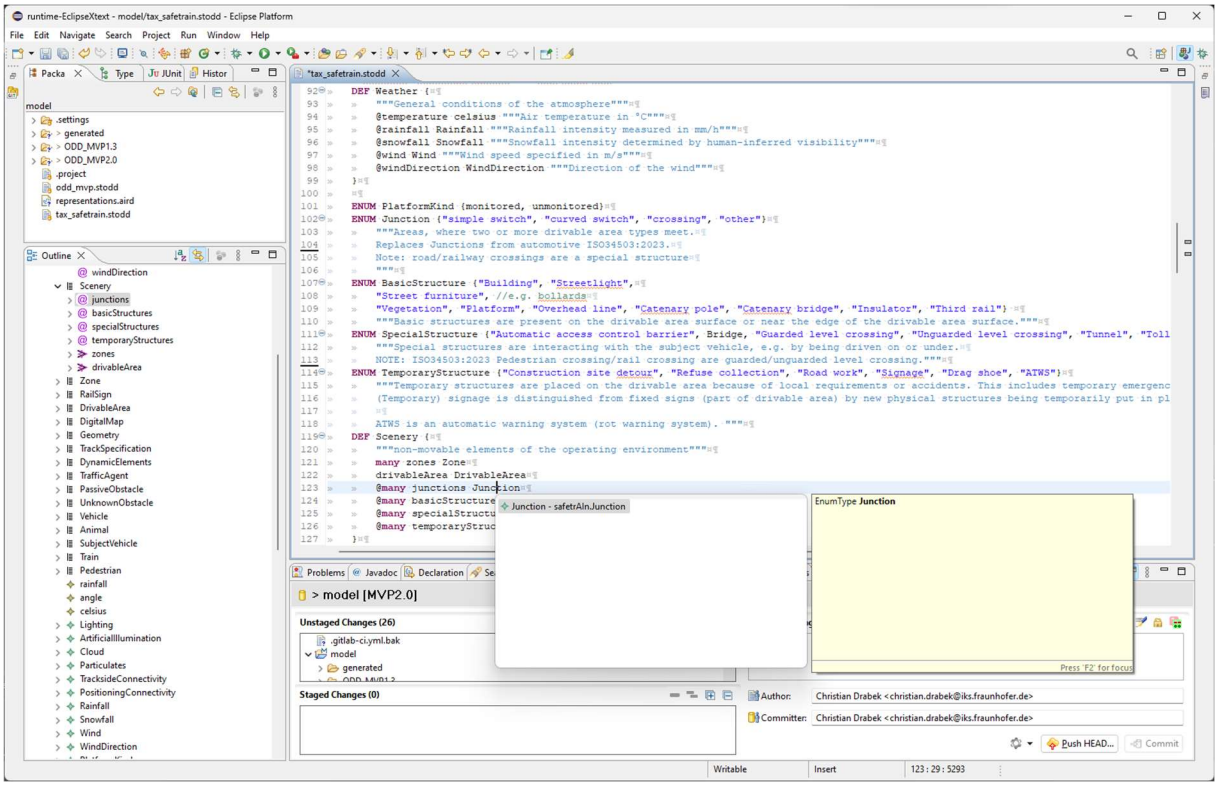


Abbildung 20: *Screenshot der Workbench, die zur Bearbeitung einer ODD-Taxonomie verwendet wird*

Aus Gründen der Modularität verwendet die ODD-Workbench zwei Arten von Ebenen für ODD-Definitionen: Einschränkungs- und Instanzenebenen. Einschränkungsebenen verwenden einfache Wahr/Falsch-Aussagen, um zu definieren, was innerhalb einer ODD festgelegt ist, und verweisen auf bestimmte Attribute oder Bedingungen in der Taxonomie. Instanzenebenen legen die Grenzen der relevanten Taxonomie-Attribute eindeutig fest und lassen definierte Unsicherheiten durch Bereiche oder Optionssätze zu. Dieser zweigeteilte Ansatz gewährleistet, dass Taxonomie-Referenzen sowohl erwartete Bedingungen als auch tatsächlich gemessene Szenarien genau beschreiben können.

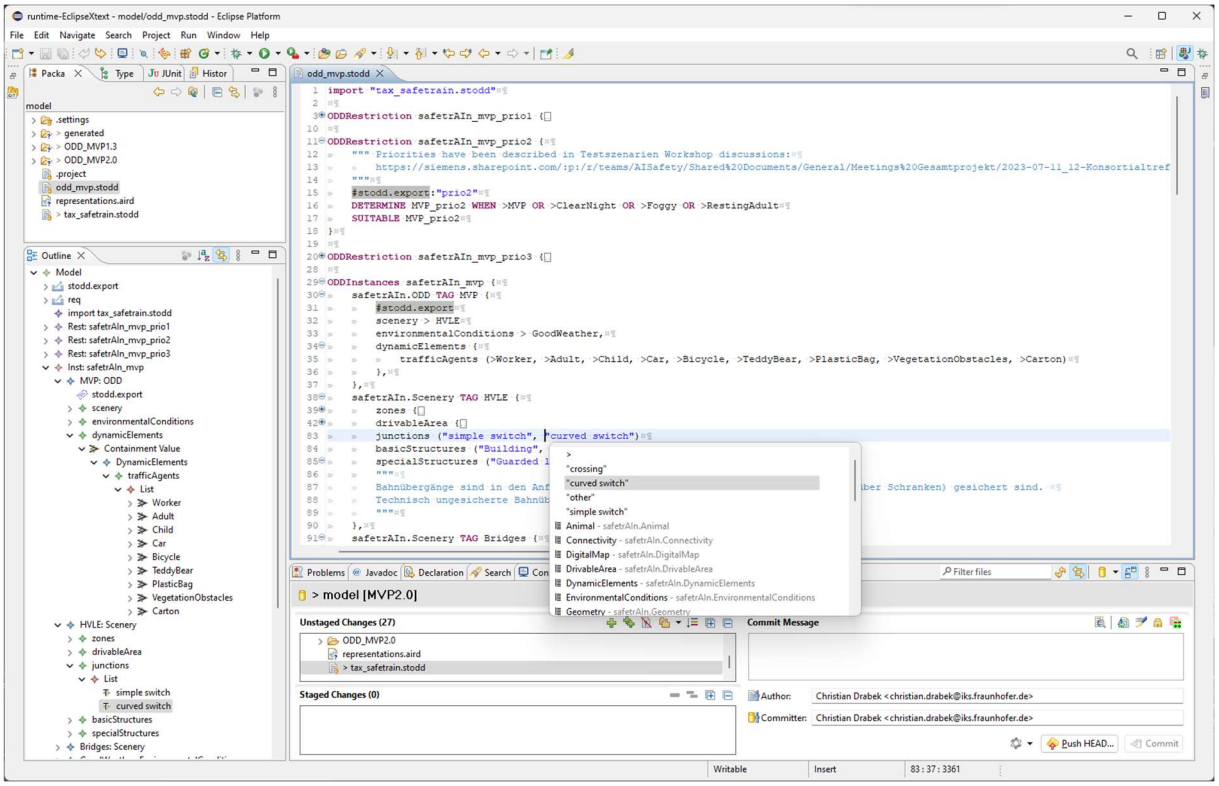


Abbildung 21: *Screenshot der Workbench, die zur Bearbeitung einer ODD-Beschreibung verwendet wird*

Die Workbench wurde möglichst benutzerfreundlich für Entwickler gestaltet und verwendet eine textuelle domänenspezifische Sprache (Domain-Specific Language, DSL). Diese erleichtert das Definieren und Ändern von ODD-Taxonomien und -Definitionen ähnlich wie das Schreiben von Code (vgl. Abbildung 21). Sie unterstützt Versionierung und Versionskontrolle und ermöglicht so die Zusammenarbeit und iterative Entwicklung. Funktionen wie Gliederung, Autovervollständigung, Überprüfung und Syntaxhervorhebung verbessern die Benutzerfreundlichkeit, während die Möglichkeit, Modelle mit einem beliebigen Texteditor anzuzeigen und zu bearbeiten, die Flexibilität, Wiederverwendbarkeit und Integration in andere Werkzeuge erhöht.

Als Teil der ODD-Workbench wurden darüber hinaus in safe.trAIn verschiedene Tools zur Verwendung und Bewertung von ODDs in realen Situationen entwickelt. Diese Tools prüfen, wie gut ODD-Definitionen miteinander und mit vorhandenen Testdaten übereinstimmen. Um Beziehungen zwischen verschiedenen Taxonomien herzustellen, die unterschiedliche Namen oder Darstellungen für dieselben Konzepte verwenden, wurden spezielle Abbildungs-Ebenen definiert. Dies hilft bei der automatischen Überprüfung von Abdeckung und Kompatibilität: Dabei werden verschiedene Taxonomien miteinander verbunden wie Labels aus Algorithmen für Maschinelles Lernen und Kundenerwartungen. Beispielsweise können Bilddaten-Labels, die für maschinelles Lernen verwendet werden, auf der Grundlage von ODD-Konzepten übersetzt werden. Dafür werden in diesem Fall Python-Skripte generiert, welche die Übersetzung von Labels in ODD-Attribute automatisieren. Außerdem erleichtern sie es den Entwicklern, KI-Ergebnisse mit ODD-Spezifikationen zu verbinden. Vergleichstools helfen dabei, Klarheit und Konsistenz zwischen ODD-Taxonomien und -Definitionen sicherzustellen. Diagramme als visuelle Darstellungen

von Abdeckungen, Überschneidungen und Lücken erleichtern die Bewertung der Eignung einer ODD und der Nutzung von Datensätzen (s. Abbildung 22).

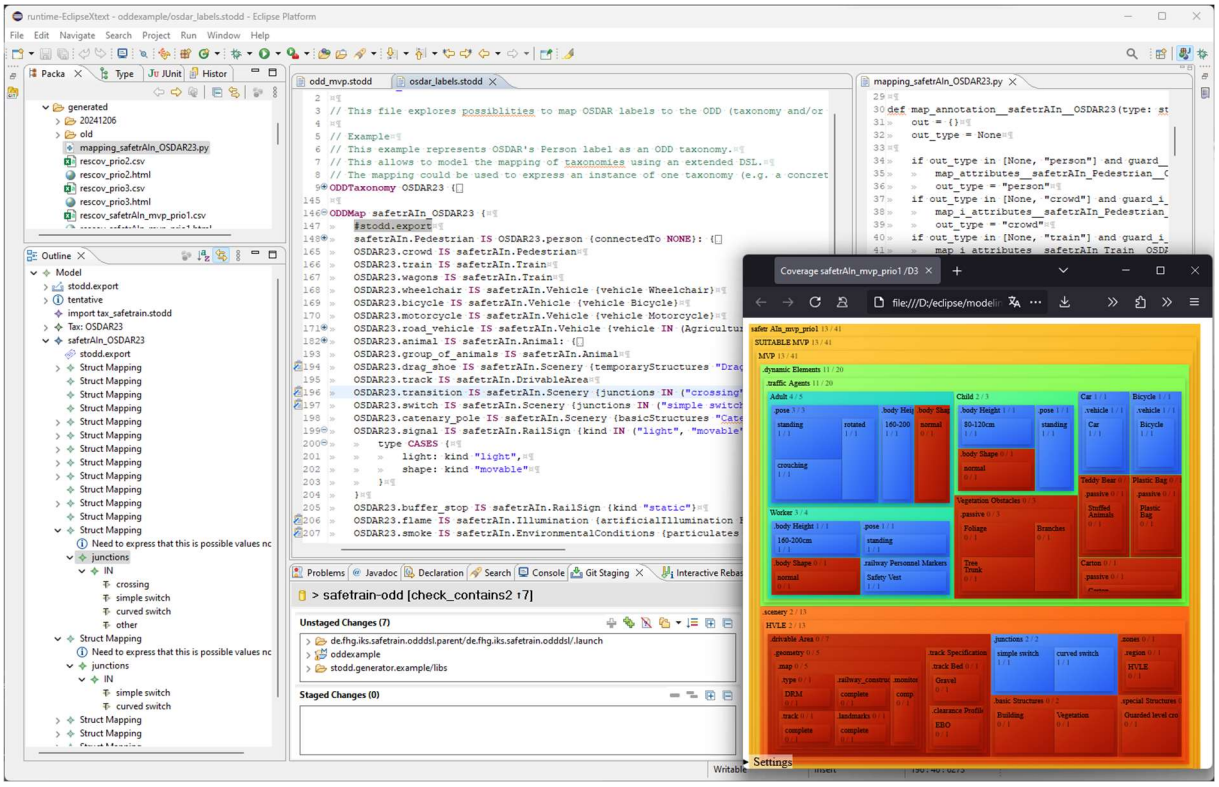


Abbildung 22: *Screenshot der Workbench, die eine ODD-Taxonomie ML-Labels zuordnet und die Überschneidung visualisiert*

Die Reporting-Werkzeuge der Workbench erstellen detaillierte Textberichte oder CSV-Tabellen (vgl. Abbildung 23) und erleichtern so die Dokumentation und Kommunikation, indem sie ODD-Taxonomien und -Definitionen leicht zugänglich machen. Beispielsweise konnte so die Diskussions- und Entwicklungsgrundlage für die standardisierte ODD-Definition der DIN DKE SPEC 99004 erzeugt und kontinuierlich aktualisiert werden.

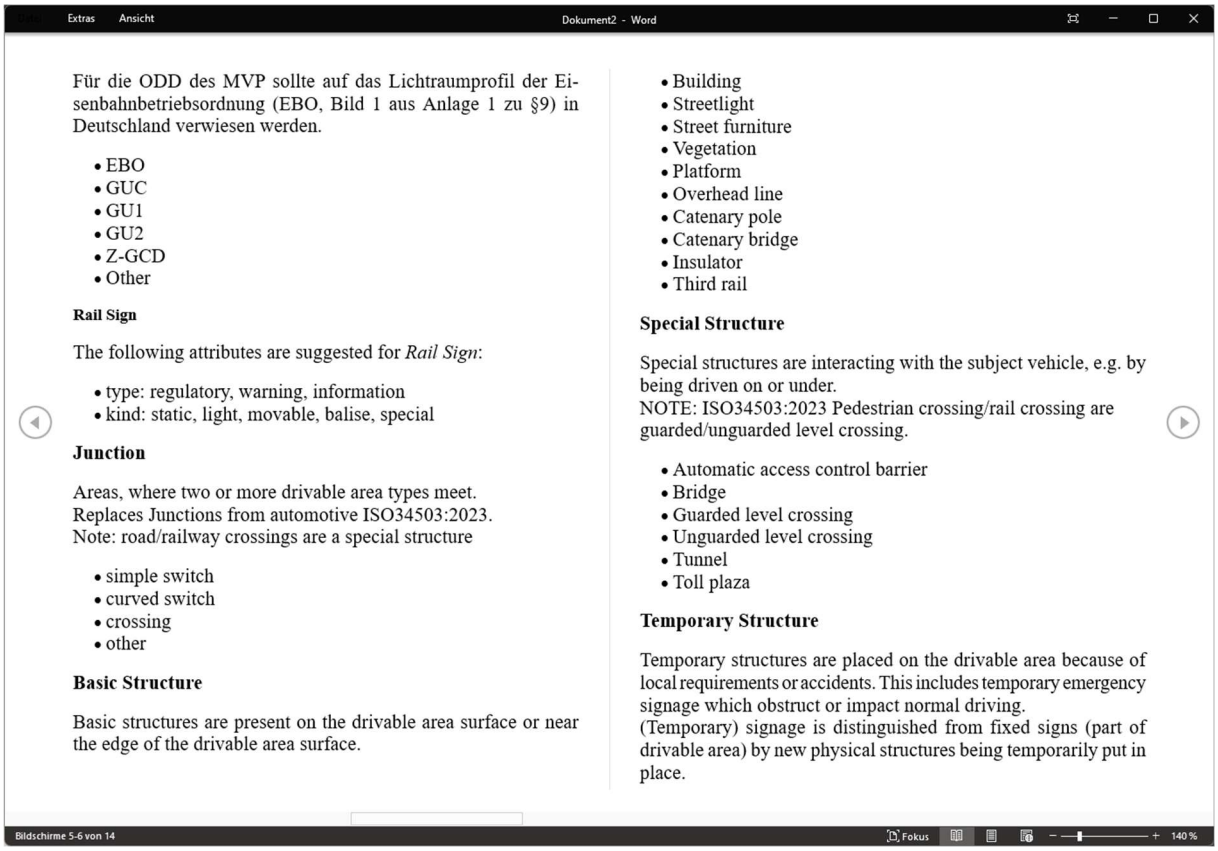


Abbildung 23: *Screenshot eines generierten Berichts für eine ODD-Taxonomie*

Zudem wurde ein Visualisierungs-Tool entwickelt, um die Verwendung von ODD-Taxonomie und -Beschreibung innerhalb des Projekts zu vereinfachen (s. Abbildung 24). Die Visualisierung erfolgt graphisch in einer Baum-Struktur, welche die Zusammenhänge zwischen verschiedenen ODD-Elementen intuitiv darstellt. So kann interaktiv eine ODD-Taxonomie oder -Beschreibung analysiert werden. Beim Navigieren können die Details der einzelnen ODD-Elemente im rechten Bildbereich eingesehen werden. Das Tool erlaubt darüber hinaus verschiedene Ansichten zu aktivieren, die gezielt Elemente ein- und ausblendet. Zum Beispiel um Elemente für bestimmte Aspekte anzuzeigen, wie für Safety oder für Machine-Learning-Funktion relevante. Dies ermöglicht „Sichten" (sogenannte Viewpoints) auf die ODD aus verschiedenen Anwenderperspektiven, bei denen man sich auf die für die eigene Aufgabe relevanten Bereiche fokussieren kann. Auf das Tool kann mithilfe eines Browsers zugegriffen werden, wodurch alle Stakeholder der ODD einen schnellen Überblick erhalten können, ohne Software lokal installieren zu müssen. Im Projekt trug das Visualisierungs-Tool maßgeblich zum Verständnis des ODD-Konzepts mit der ODD-Taxonomie und ODD-Beschreibung über den Entwicklungsprozess hinweg bei. Dies führte auch zur hohen Akzeptanz und Nutzung bei den Projektpartnern bei.

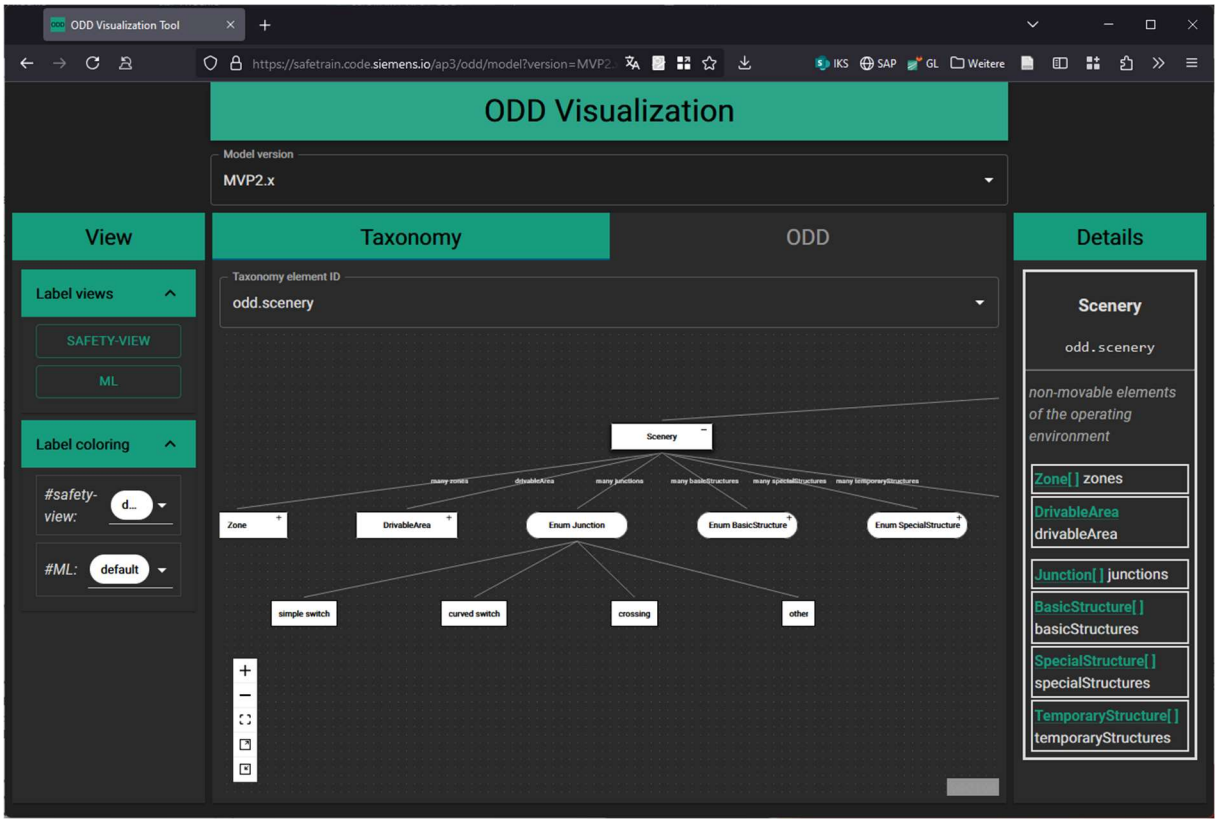


Abbildung 24: *Screenshot des Visualisierungs-Tools zum Durchsuchen von ODD-Taxonomien und -Definitionen*

Darüber hinaus ermöglicht die Workbench den Export von ODDs in benutzerdefinierte Formate, was die Benutzerfreundlichkeit und Integration erhöht. Diese Funktion unterstützt die Anpassung von Ausgaben an spezifische Systemanforderungen und kann in Continuous Integration/Continuous Delivery (CI/CD)-Pipelines integriert werden. Dadurch können ODD-Konzepte auch in vorhandenen Lebenszyklus-Management-Tools verfügbar gemacht werden und mit aktuellen Verfahren bei der Testplanung und Testfallgenerierung verknüpft werden. Dies ermöglicht beispielsweise eine Testabdeckung der ODD zu verfolgen.

II.1.3.2 Sicherheitsnachweis für die GOA3/4 Architektur

Für die Sicherheitsnachweisführung der entwickelten Architektur für die Hinderniserkennung wurden verschiedene grundlegende Beiträge von Fraunhofer geleistet. Das beinhaltet die Entwicklung einer Assurance Case Argumentation auf Basis der Goal Structuring Notation (GSN) und die Entwicklung und Evaluierung der in AP2 erarbeiteten Metriken aller Partner.

Beiträge zum Sicherheitsnachweis

Für die Erarbeitung des Sicherheitsnachweis trug Fraunhofer wesentlich zur Entwicklung der Systematik zum Einsatz der in AP2 entwickelten Metriken in der Sicherheitsargumentation bei. Dafür unterstützte Fraunhofer bei der Entwicklung der Gesamtmethodik zum Aufbau einer evidenz-basierten Sicherheitsargumentation. Diese beruht darauf, dass KI-spezifische Safety Concerns explizit (anhand der LAISC) identifiziert werden, Evidenzen zur Risikobewertung dieser Sicherheitsbedenken über spezifische Metriken gesammelt werden und die Argumentation zur

ausreichenden Risikominderung in einem strukturierten Sicherheitsnachweis, auch mit Hilfe der GSN (Goal Structuring Notation), erbracht wird. Mithilfe speziell im Projekt erarbeiteten Fact Sheets wurden die Erkenntnisse und Ergebnisse der in AP2 entwickelten Methoden erhoben und dokumentiert. Um den Beitrag von Maßnahmen anhand der Metriken zu den Sicherheitszielen zu identifizieren und belegen, wurden diese konkreten Teilzielen des GSN-basierten Sicherheitsnachweis zugeordnet. Gemeinsam mit den Entwicklern der Metriken und Methoden bewertete Fraunhofer, ob das jeweilige Sicherheitsziel bereits erwiesen oder plausibel durch Anwendung der Metriken adressiert werden kann. Hierfür wirkte Fraunhofer auch bei der Ableitung sogenannter „Verifiable Requirements" (überprüfbarer Anforderungen) auf Basis der LAISC und GSN-Bäume mit. Aufgrund nicht definierter Schwellwerte für die Metriken fand die Bewertung deren Beitrag zum Sicherheitsnachweis nur qualitativ statt. Eine quantitative Bewertung sollte dann möglich sein, sobald Schwellwerte für einzelne Metriken oder Sicherheitsziele für ein reales System abgeleitet werden. Regelmäßige Abgleiche zwischen der LAISC und dem GSN-basierten Sicherheitsnachweis aus AP2 unterstützen, die Konsistenz und Abdeckung beider sicherzustellen. Somit konnte iterativ und inkrementell das Sicherheitskonzept entwickelt werden.

ODD-basierte Safety

Zusammen mit dem Projektpartner Siemens wurde ein Ansatz zur Argumentation der Systemsicherheit auf Basis einer ODD erarbeitet und bei der internationalen Konferenz CAIN'24 veröffentlicht und vorgestellt. Zudem wurde ein weiteres Paper zum Themenfeld erarbeitet und bei der internationalen Konferenz EDCC'24 vorgestellt. Es beschreibt, wie eine ODD-Beschreibung für die Teilautomatisierung von Safety-Engineering-Aktivitäten verwendet werden kann.

II.1.4 AP4 Virtuelles Testfeld und Sicherheitsbewertung

Das AP4 hatte die Entwicklung eines virtuellen Testfelds für die Absicherung der in safe.trAIn betrachteten Hinderniserkennung eines Regionalzugs im Fokus. Fraunhofer wirkte durch Beiträge zur Konzeptionierung des virtuellen Testfelds mit. Insbesondere konnten die Erkenntnisse aus dem BMWK Förderprojekt KI-Absicherung mit eingebracht werden. Dies betraf die Entwicklung des Gesamt-Vorgehens im Projekt zur Verzahnung von ODD-Beschreibung, KI-Entwicklung, Entwicklung des Safety-Cases und den Ergebnissen aus der Validierung und dem virtuellen Testfeld.

Folgende Abbildung zeigt den mit Unterstützung von Fraunhofer entwickelten Gesamtzyklus zur Entwicklung des KI-Systems, verzahnt mit der Entwicklung des Sicherheitsnachweises:

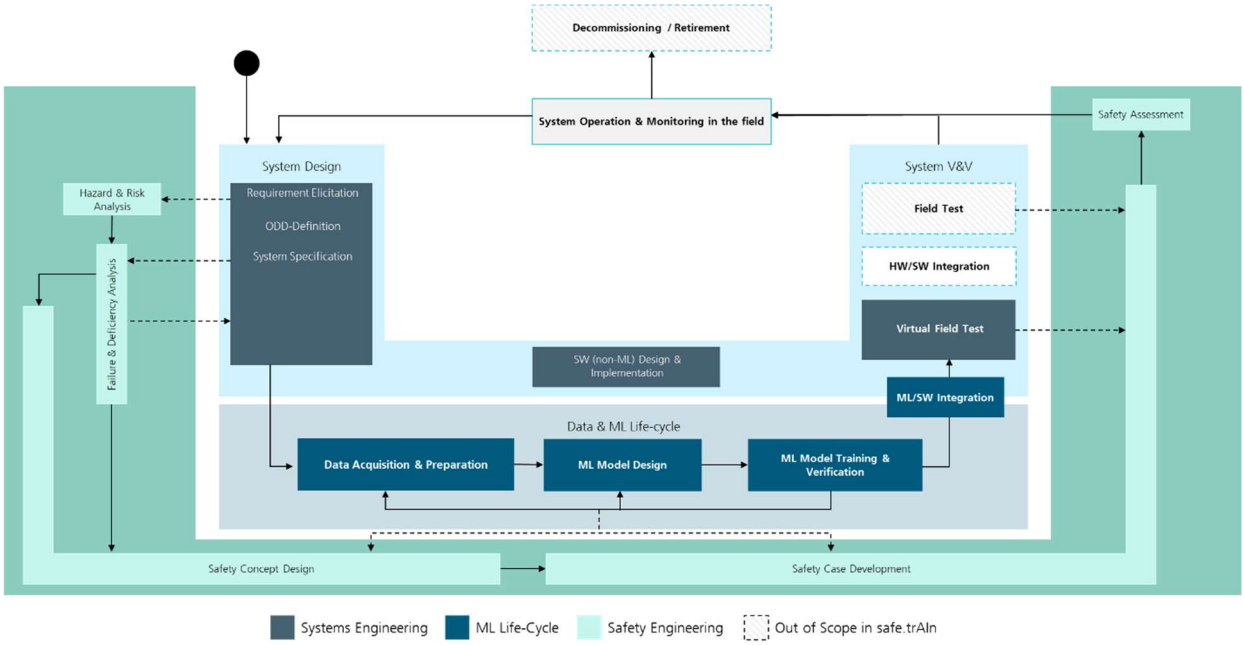


Abbildung 25 Safe MLOps Prozess für ML-basierte Systeme in der Railway Domäne (nach[12])

Die Grundidee der dargestellten Vorgehensweise besteht darin, den Entwicklungszyklus des KI-Systems mit der Entwicklung des Sicherheitsnachweises (grün) im Gesamtprozess zu verzahnen. Dazu gehört die Einbettung des klassischen ML-Zyklus in den Prozess der Erstellung und Testen des Gesamt-Systems (dunkelblau). Auf Basis einer Risiko-Analyse des Gesamtsystems ergeben sich sowohl die Sicherheitsanforderungen an den Sicherheitsnachweis als auch Anforderungen den das Design und Test des ML Systems, insbesondere in Bezug auf die Wahl und Entwicklung von sicherheitsrelevanten Metriken (vgl. Abschnitt II.1.2). Ergebnisse der Auswertung dieser Metriken sowohl auf Ebene des ML-Modells (unten) als auch auf Systemebene im virtuellen Testfeld (rechts) gehen unmittelbar in die Entwicklung des Sicherheitsnachweises ein.

Die Ergebnisse und Erkenntnisse aus dem virtuellen Testfeld wurden sowohl in den Fact Sheets der Metriken aus AP2 als auch in Validierungsreports dokumentiert. Auf deren Basis wurde die Relevanz der sicherheitsrelevanten Metriken für die Sicherheitsnachweisführung bewertet und die Metriken mit dem GSN-basierten Sicherheitsnachweis verknüpft. Schließlich wurden sowohl der GSN-basierte Sicherheitsnachweis, die Fact Sheets als auch die Validierungsreports für die konzeptionelle Begutachtung zur Nachweisführung für das sichere Hinderniserkennungssystem bereitgestellt und genutzt.

II.1.5 AP5-6 Standardisierung, Verwertung und Projektmanagement

Fraunhofer lieferte Beiträge zum Transfer der methodischen Vorgehensweise auf andere Anforderungen der Vertrauenswürdigkeit. Neben der Vorstellung der grundsätzlichen oben beschriebenen Vorgehensweise zum Aufbau einer Sicherheitsargumentation für autonomes Fahren auf verschiedenen nationalen Workshops hat Fraunhofer insbesondere bei der Durchführung eines Anwenderkreis-Workshops gemeinsam mit den DIN zur Vernetzung der Projekte safe.trAIn und

[12] Towards a safe MLOps Process for the Continuous Development and Safety Assurance of ML-based Systems in the Railway Domain, https://arxiv.org/pdf/2307.02867v1

"Zertifzierte KI" beigetragen. In weiteren Anwenderkreisen präsentierte und diskutierte Fraunhofer seine Projektergebnisse, was auch zur Initiierung der beiden aus dem Projekt entstandenen DIN DKE SPECs beigetragen hat. Zudem entstand durch Beiträge von Fraunhofer in Zusammenarbeit mit dem DIN und dem VDE ein Entwurf eines Vorhabens zur Entwicklung einer DIN DKE SPEC mit dem Arbeitstitel "Prozess zur Erstellung von Metriken für sicherheitskritische KI". Fraunhofer definierte den Scope einer möglichen SPEC. Die beabsichtigte DIN DKE SPEC sollte einen Prozess beschreiben, der geeignet ist, die für einen solchen Sicherheitsnachweis geeigneten und relevanten Metriken zu bestimmen. Der Prozess sollte auch umfassen, wie die Akteure mit ihren jeweilen Kompetenzen, insbesondere KI-Experten, Safety-Experten und Domänen-Experten, zusammenwirken, die relevanten Metriken zu bestimmen. Auf dem mit dem DIN und VDE durchgeführten Workshop wurde des Weiteren eine Zielgruppen- und Stakeholder-Analyse durchgeführt. Die weitere Verfolgung der DIN SPEC konnte jedoch aufgrund von Ressourcenknappheit bei wichtigen weiteren Partnern und Priorisierung anderer Projektthemen für die Standardisierung nicht weiterverfolgt werden. Als weiterer Beitrag hat Fraunhofer die Anwenderkreise zur Dissemination und Verzahnung mit externen Experten und Projekten unterstützt.

Im Rahmen von safe.trAIn wurden zwei DIN DKE SPECs als Standardisierungsaktivität erfolgreich umgesetzt. DIN DKE SPEC 99002 „Terminologie – KI in Bahnanwendungen"[13] definiert Begriffe zum Thema Künstliche Intelligenz und Schienenverkehr, um das gemeinsame Verständnis für KI-Anwendungen im Bereich des schienengeführten Transports, einschließlich Eisenbahnen, zu verbessern. Die entwickelte Terminologie dient als Grundlage und zur Verbesserung der eindeutigen Kommunikation im neuen Themenfeld und richtet sich an verschiedene mit dem Bahnbereich involvierten Akteure. Die Definition wichtiger Begriffe für KI in Bahnanwendungen soll zu einer gleichen Interpretation der für den Bereich neuer Technologien und Konzepte führen. Fraunhofer hat insbesondere die Teilgruppe und deren Beiträge zu Definitionen für den Bereich Vollautomatisiertes Fahren und ODD organisiert.

In DIN DKE SPEC 99004 „Spezifikation von ODD im Schienenverkehr"[14] wurde das Thema ODD für den Bahnbereich behandelt, welche Kernbeiträge aus dem safe.trAIn Projekt enthält. Hierfür wurde die ODD-Taxonomie (vgl. Abschnitt II.1.3.1) als Ausgangspunkt für die SPEC verwendet. Fraunhofer initiierte die SPEC, stellte den Geschäftsplan mit auf und übernahm den Vorsitz. So koordinierte Fraunhofer das Standardisierungs-Team aus internationalen Experten und führte das Vorhaben erfolgreich zur Veröffentlichung der SPEC. Diese kann nun als Hilfestellung und Vorlage für die Definition von ODDs im Bahnbereich dienen, welche für viele Arten von KI-Anwendungen notwendig ist. Darüber hinaus hat Fraunhofer die SPEC weiteren Standardisierungsgruppen für eine mögliche internationale / europäische Standardisierung vorgestellt und diskutiert.

Fraunhofer hat sich vielseitig an der Ergebnisverbreitung der safe.trAIn Projektergebnisse beteiligt. Dies beinhaltet neben den Anwenderkreisen beispielsweise auch wissenschaftliche und Fach-Veröffentlichungen sowie Vorträge (s. auch Abschnitt II.6) oder auch verschiedene Diskussionen mit weiteren Forschungspartnern und Industrieunternehmen.

[13] DIN DKE SPEC 99002:2025-03, https://dx.doi.org/10.31030/3600968
[14] DIN DKE SPEC 99004:2025-05, https://dx.doi.org/10.31030/3610746

Hinsichtlich des Projektmanagements wurden die eigenen Arbeiten von Fraunhofer kontinuierlich koordiniert, mit den jeweiligen Partnern, Konsortialführer und Projektträger abgestimmt. Hierunter fällt auch die Mitwirkung am organisatorischen und technischen Projektsteuerkreis sowie viele weitere Koordinierungsmeetings zu spezifischen Projektthemen.

II.2 Wichtigste Positionen des zahlenmäßigen Nachweises

Die beantragten Projektmittel wurden grundlegend gemäß dem beantragten Finanzierungsplan ausgegeben. Der Hauptanteil der Kosten entstand wie geplant für den Personalaufwand. Im Rahmen des Projekts wurden durch Fraunhofer keine Mittel an Dritte gegeben.

II.3 Notwendigkeit und Angemessenheit der geleisteten Arbeit

Die sichere Hinderniserkennung im Fahrweg ist entscheidend für die Entwicklung eines fahrerlosen Regionalzugs, wobei KI-Methoden eine vielversprechende Lösung darstellen. Jedoch fehlen insbesondere robuste Konzepte, um eine erfolgreiche Zulassung zu ermöglichen. Einzelne Forschungsarbeiten bieten zwar Ansätze zur Qualitätssicherung von KI-Komponenten, aber deren Integration in einen umfassenden Prüfprozess und eine Sicherheitsarchitektur mit neuen Sensoransätzen ist ungelöst. Die Technologieentwicklung und Integration stellte ein Risiko dar und benötigte einen Forschungsverbund unterschiedlicher Partner sowie finanzielle Unterstützung durch öffentliche Fördermittel. Ein fahrerloser Regionalzug bietet gesellschaftlichen Nutzen durch attraktive Schienenverkehrsangebote bei reduzierten Betriebskosten. Die von Fraunhofer durchgeführten Forschungsarbeiten im Projekt safe.trAIn sowie die dafür aufgewandten Ressourcen waren notwendig und angemessen, um diese geplanten Ziele zu erreichen und das Gesamtprojekt erfolgreich abzuschließen. Die in das Projekt eingebrachten Arbeiten entsprechen dem Umfang und Komplexität der zu bearbeitenden Fragestellungen. Die für den erfolgreichen Abschluss notwendige Verlängerung des Projekts um drei Monate konnte kostenneutral umgesetzt werden.

II.4 Voraussichtlicher Nutzen und Verwertbarkeit

Die Fraunhofer-Gesellschaft verfolgt traditionell das Ziel, wissenschaftliche Erkenntnisse in praktische Anwendungen zu überführen, um sowohl wirtschaftlichen als auch gesellschaftlichen Mehrwert zu generieren. Der voraussichtliche Nutzen und die Verwertbarkeit der Projektergebnisse aus dem Projekt safe.trAIn sind entscheidend für die weitere Ausrichtung der beiden beteiligten Fraunhofer-Institute. Im Rahmen dieses Projekts wurde bewusst der Fokus auf den Forschungstransfer gelegt. Die Verwertung der Ergebnisse erfolgt durch die Stärkung und Fortführung der Forschungsaktivitäten in Schlüsselbereichen wie Safety Assurance für Künstliche Intelligenz (KI), vertrauenswürdige KI, Datengenerierung, Prüfung von KI-Funktionen sowie resiliente Softwaresysteme. Diese Bereiche sind von herausragender Bedeutung, da sie die Grundlage für die Entwicklung sicherer und effektiver KI-Anwendungen in vielen Bereichen bilden.

Die identifizierten Verwertungsgebiete umfassen unter anderem den Gesundheitssektor, die Mobilität, die Industrieautomatisierung sowie Automated Business Decisions und Business Analytics. Insbesondere die verlässliche Erkennung von Personen und die Definition einer Operational

Design Domain (ODD) sind zentrale Elemente, die in sämtlichen Bereichen mobiler autonomer Systeme, wie Autonomous Mobile Robots, Mining Vehicles, Mobility Shuttles oder Safe Robot Control, entscheidend sind. Diese Themen werden bereits aktiv in verschiedenen Domänen erforscht, was die Grundlage für eine breite Anwendbarkeit und Transferierbarkeit der Projektergebnisse schafft.

Zusätzlich wird angestrebt, die gewonnenen Erkenntnisse in zukünftigen anwendungsorientierten Forschungsprojekten weiterzuentwickeln und im Rahmen von Veröffentlichungen zu publizieren. Ein zudem wichtiger Aspekt ist die weitere Verfolgung der Ergebnisse in Standardisierungsaktivitäten, die zur Schaffung einer einheitlichen Basis in der Branche beitragen können. Durch diese Maßnahmen wird sichergestellt, dass die Ergebnisse des Projekts nachhaltig verwertet werden können und einen signifikanten Beitrag zur Weiterentwicklung der für viele Bereiche relevanten Technologien und Systeme leisten.

II.5 Fortschritts auf dem Gebiet des Vorhabens bei anderen Stellen

Auf dem Gebiet der Forschung und Entwicklung rund um KI für sicherheitskritische Aufgaben und Objekterkennung, wie sie auch für eine Hinderniserkennung notwendig sind, hat es einige allgemeine Aktivitäten gegeben. Für den Anwendungsfall eines fahrerlosen Regionalzugs sind keine konkreten, relevanten Arbeiten bekannt. Allerdings gibt es in Deutschland auf regionaler und nationaler Ebene Projekte mit verwandten Entwicklungszielen. Im Projekt KI-Lok sollten Prüfverfahren für KI-basierte Komponenten im Eisenbahnbetrieb entwickelt werden. Das Projekt ARTE (Automatisiert fahrende Regionalzüge in Niedersachsen) sollte die technische Machbarkeit eines GoA3-Betriebs mit ETCS auf einer Linie in Niedersachsen erforschen. Im Rahmen von safe.trAIn fand ein Austausch mit diesen Projekten statt. Das französische Großprojekt Confiance.ai verfolgt die Erforschung vertrauenswürdiger KI in verschiedenen industriellen Anwendungsfällen. Fraunhofer beteiligte sich als Partner von safe.trAIn am Austausch u.a. beim Confiance.ai Day. Im Rahmen der DIN SPEC Standardisierung konnten zudem europäische Initiativen zum Thema ODD im Bereich Rail zusammengeführt werden. Mit OpenODD von ASAM wurde inzwischen ein offener Standard zur maschinenlesbaren Beschreibung und Nutzung von ODDs mit Fokus auf den Automotive-Bereich veröffentlicht.

II.6 Veröffentlichungen

P. Schleiss, Y. Hagiwara, I. Kurzidem, and F. Carella, "Towards the quantitative verification of deep learning for safe perception," in *2022 IEEE International Symposium on Software Reliability Engineering Workshops (ISSREW)*, Oct. 2022, pp. 208-215.

P. Schleiss, F. Carella, and I. Kurzidem, "Towards continuous safety assurance for autonomous systems," in *2022 6th International Conference on System Reliability and Safety (ICSRS)*, Nov. 2022, pp. 457-462.

F. Schwaiger, A. Matic, K. Roscher, and S. Günnemann, "Preventing Errors in Person Detection: A Part-Based Self-Monitoring Framework," *arXiv preprint arXiv:2307.04533*, 2023.

S. S. Gannamaneni, M. Mock, and M. Akila, "Assessing systematic weaknesses of DNNs using counterfactuals," *AI and Ethics*, pp. 1-9, 2024.

S. S. Gannamaneni, A. Sadaghiani, R. P. Rao, M. Mock, and M. Akila, "Investigating CLIP Performance for Meta-data Generation in AD Datasets," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 3839-3849.

G. Weiss, "How autonomous systems become reality – Operational Design Domains for Highly Automated Functions of Embedded Systems," in *Proceedings of Embedded Software Engineering (ESE) Kongress 2023*, 2023.

M. Trapp, "Certifying Autonomy in Railway Systems," Invited Talk at DATE 2024 Initiative on Autonomous Systems Design (ASD), 2024.

G. Weiss, "Challenge of understanding operational context of AI for trustworthiness," Invited Talk at confiance.AI Day 2024, Paris, 07.03.2024.

Article on safe.trAIn: "Besser unterwegs mit Bus und Bahn," *Fraunhofer-Magazin 1/2024*, Apr. 2024.

A. Kreutz, G. Weiss, and M. Trapp, "Automatic Deduction of the Impact of Context Variability on System Safety Goals," in *Proceedings of the 19th European Dependable Computing Conference*, Leuven, Belgium, Apr. 2024.

G. Weiss, M. Zeller, H. Schoenhaar, A. Kreutz, and C. Drabek, "Approach for Argumenting Safety on Basis of an Operational Design Domain," in *3rd International Conference on AI Engineering – Software Engineering for AI (CAIN 2024)*, 2024.

S. S. Gannamaneni, F. Klein, M. Mock, and M. Akila, "Exploiting CLIP Self-Consistency to Automate Image Augmentation for Safety Critical Scenarios," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, 2024, pp. 3594-3604.

S. S. Gannamaneni, R. P. Rao, M. Mock, M. Akila, and S. Wrobel, "Detecting Systematic Weaknesses in Vision Models along Predefined Human-Understandable Dimensions," In Transactions on Machine Learning Research (TMLR), https://openreview.net/forum?id=yK9pvt4nBX.

P. Sinhamahapatra, F. Schwaiger, S. Bose, H. Wang, K. Roscher, and S. Guennemann, "Finding Dino: A plug-and-play framework for unsupervised detection of out-of-distribution

objects using prototypes," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, 2025.

III. Abkürzungen

Abkürzung	**Begriffserklärung**
AF	Autonomes Fahren
AP	Arbeitspaket
ATO	Automatic Train Operation
CLIP	Contrastive Language Image Pre-training
CSV	Comma-Separated Values
DNN	Deep Neural Network
DSL	Domain-Specific Language
ETCS	European Train Control System
FCN	Fully Convolutional Network
GSN	Goal Structuring Notation
HVLE	Havelländische Eisenbahn
IoU	Intersection over Union
KI	Künstliche Intelligenz
LAISC	Landscape of AI Safety Concerns
MVP	Minimal Viable Product
ODD	Operational Design Domain
OOD	Out-of-Domain
PROWL	PRototype based OOD detection Without Labels
SOTA	State-of the-Art
SPD	Semantic Performance Discrepancy
VA	Visual Analytics