

IOSB

Fraunhofer

visIT

[Erklärbare KI im Praxiseinsatz]



Impressum

visIT erscheint etwa drei Mal pro Jahr und informiert über ausgewählte Forschungsthemen des Fraunhofer IOSB. Für die Bestellung von Einzelheften, ein kostenloses visIT-Abo sowie für Adressänderungen und Abbestellungen schicken Sie bitte eine E-Mail an publikationen@iosb.fraunhofer.de

Herausgeber

Prof. Dr.-Ing. habil. Jürgen Beyerer

Redaktion

Dipl.-Phys. Ulrich Pontes

Layout

Anja Wollfarth M.A.

Druck

Stork Druckerei GmbH
76646 Bruchsal

Anschrift der Redaktion

Fraunhofer-Institut für Optronik,
Systemtechnik und Bildauswertung IOSB
Fraunhoferstr. 1
76131 Karlsruhe
Telefon +49 721 6091-300
Fax +49 721 6091-413
presse@iosb.fraunhofer.de

© Fraunhofer IOSB
Karlsruhe 2023

Ein Institut der Fraunhofer-Gesellschaft
zur Förderung der angewandten
Forschung e. V. München

24. Jahrgang
ISSN 1616-8240

Bildquellen

Titelbild: Stefan Riel, Fraunhofer IOSB
Seite 4: Indigo
Seite 8, 9 & 10: Stanford Cars Dataset
Seite 11: VisDrone Dataset

Alle anderen Abbildungen:
© Fraunhofer IOSB

Nachdruck, auch auszugsweise,
nur mit vollständiger Quellenangabe und
nach Rücksprache mit der Redaktion.
Belegexemplare werden erbeten.

Inhalt

Themen

- | | | |
|-------|----|--|
| Seite | 4 | XAI in der Qualitätssicherung
Mathias Anneken, Manjuntha Veerappa |
| Seite | 6 | Qualität sichern und Ressourcen schonen: XAI in der industriellen Kunststoffproduktion
Norbert Greifzu, Andreas Wenzel |
| Seite | 8 | XAI in der Bildauswertung
Arne Schumann |
| Seite | 10 | Semantische XAI-Verfahren für natürlichsprachliche Erklärungen
Achim Kuwertz, Almuth Müller |
| Seite | 12 | Erklärbare KI in der Medizin aus Sicht von User Interface Design und Usability Engineering
Henrik Mucha |
| Seite | 14 | Mit der XAI-Toolbox KI-Modelle durchleuchten
Mathias Anneken, Arne Schumann |

Liebe Freunde des Fraunhofer IOSB,

Künstliche Intelligenz (KI) kann den Menschen vielfältig unterstützen, etwa bei der Optimierung industrieller Fertigungsabläufe, beim Autofahren oder in der Medizin bei der Auswahl von Therapiemöglichkeiten. Da KI-Lösungen eine immer größere Rolle im Alltag spielen, sind ethische Grundsätze – wie bspw. durch die EU im derzeit verhandelten AI Act beschrieben – wichtige Grundpfeiler für ihre Entwicklung. Nicht zuletzt um diese Grundsätze zu beachten, sind Nachvollziehbarkeit und Transparenz von KI-Modellen wichtig: Sie unterstützen die Entwicklung von KI-Modellen und stärken das Vertrauen in KI-Ergebnisse.

Eine lange Tradition im Bereich KI haben von Natur aus nachvollziehbare White-Box-Modelle wie kleine Entscheidungsbäume oder einfache Bayes'sche Netze. Durch den einfachen und übersichtlichen Aufbau ermöglichen sie es direkt, die algorithmischen Zusammenhänge zu verstehen.

Seit über zehn Jahren sind Verfahren aus dem Bereich des Maschinellen Lernens, insbesondere des Deep Learnings, in den Fokus gerückt. Diese werden auf großen Datenmengen trainiert und liefern für spezielle Aufgaben KI-Modelle mit hoher Qualität. Solche Modelle sind sehr komplex; wir können ihre Vorgehensweise nicht intuitiv durchdringen. Sie werden daher als Black-Box bezeichnet. Sogenannte Erklärmodelle aus dem Forschungsbereich Explainable Artificial Intelligence (XAI) ermöglichen es, die Vorgänge in der Black-Box besser zu verstehen. Diese können explizit für spezifische KI-Modelle oder allgemein für unterschiedliche KI-Modelle entwickelt werden. Die Erklärungen werden über das gesamte KI-Modell hinweg oder zu bestimmten Ergebnissen der KI generiert. Durch den Einsatz speziell entwickelter Vorgehensweisen zum User Interface Design und Usability Engineering wird eine optimale Nutzbarkeit gewährleistet.

Das Fraunhofer IOSB beherrscht die Entwicklung von KI-Lösungen und dazugehörigen Erklärmodellen und wendet diese in unterschiedlichen Einsatzgebieten an. Als Partner für den Mittelstand und die Industrie bringen wir diese Erfahrung gerne auch in Ihren Use-Case ein. In diesem visIT geben wir Einblick in konkrete XAI-Anwendungsbeispiele aus den Bereichen Qualitätssicherung und Produktion, sensorgestützte Aufklärung und Überwachung, sowie Medizin. Abschließend stellen wir unsere XAI-Toolbox vor, die es uns ermöglicht, auch für Ihr KI-Verfahren verschiedene Erklärungen zu erzeugen – und damit für zielgerichtete Entwicklung, Vertrauen und Transparenz auch in Ihrer Anwendung zu sorgen.

Wir wünschen viel Freude beim Lesen unserer Ausgabe »Erklärbare KI im Praxiseinsatz«.

Fraunhofer IOSB, im April 2023

Prof. Dr.-Ing. habil. Jürgen Beyerer

Dr. Elisabeth Peinsipp-Byma

Editorial



Prof. Dr.-Ing. habil. Jürgen Beyerer



Dr. Elisabeth Peinsipp-Byma

Themen



Mathias Anneken M.Sc.

Interaktive Analyse und
Diagnose (IAD)
Fraunhofer IOSB Karlsruhe

Telefon +49 721 6091-174
mathias.anneken@iosb.fraunhofer.de
www.iosb.fraunhofer.de/IAD



Manjunatha Veerappa M.Sc.

Interaktive Analyse und
Diagnose (IAD)
Fraunhofer IOSB Karlsruhe

Telefon +49 721 6091-619
manjunatha.veerappa@iosb.fraunhofer.de
www.iosb.fraunhofer.de/IAD

XAI in der Qualitätssicherung

In Produktionsanlagen werden zur Optimierung der Prozessführung, aber auch zur Qualitätssicherung, immer häufiger maschinelle Lernverfahren (ML) eingesetzt. Diese Verfahren liefern zwar sehr gute Ergebnisse, es ist für den Menschen aber nicht unbedingt möglich zu verstehen, wie diese zustande kommen. Insbesondere in der Qualitätssicherung – wenn bspw. Nachweispflichten zum Tragen kommen – ist dies ein Problem. Hier kommt Explainable Artificial Intelligence (XAI) ins Spiel: XAI ermöglicht eine transparente und nachvollziehbare Entscheidungsfindung für ML-Verfahren. Sie hilft somit, Fehler und Probleme in Produktionsprozessen zu identifizieren und die Effizienz und Genauigkeit der Qualitätskontrolle zu verbessern.

BEISPIEL SCHÜTTGUTSORTIERANLAGE

Als Beispielanwendung haben wir am Fraunhofer IOSB eine Schüttgutsortieranlage betrachtet, wie sie in der Qualitätssicherung zum Einsatz kommt. Ein solches sensorbasiertes Sortiersystem klassifiziert einzelne

Partikel in einem Materialstrom mittels Sensorik und Algorithmen. Über schnell-schaltende, in einem Array angeordnete pneumatische Ventile werden dann einzelne Partikel ausgestoßen, um das Material in vordefinierte Klassen zu trennen. So lassen sich etwa minderwertige Entitäten aus einem Materialstrom entfernen. Mögliche Anwendungen sind die Qualitätskontrolle in Produktionsprozessen, Landwirtschaft, Lebensmittelindustrie oder Mineralienverarbeitung sowie Vorbehandlungsprozesse für das Recycling, z. B. die Sortierung von Kunststoffen oder Bau- und Abbruchabfällen.

Durch eine Überwachung der Sortierqualität basierend auf der Auslastung der pneumatischen Ventile ist es möglich, die korrekte Funktionsweise der Anlage zu überprüfen. Sollten bei dieser Überwachung Anomalien auftreten, ist es wichtig, schnell den Grund zu erkennen. So kann eine Minderung der Sortierqualität durch ein Fehlverhalten der Anlage, aber auch eine abnormale Abnutzung der Komponenten frühzeitig abgestellt werden.

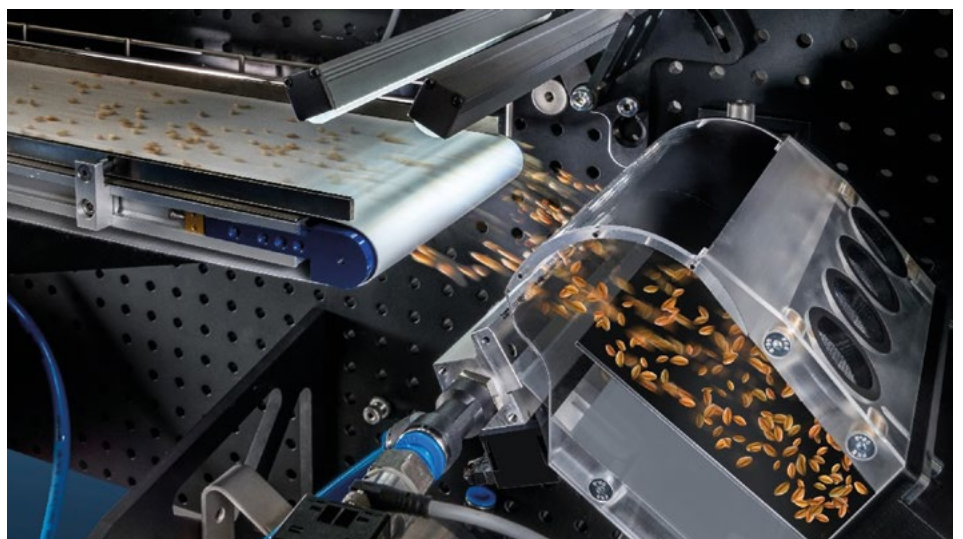


Abb. 1: Das vom Fraunhofer IOSB entwickelte sensorbasiertes Sortiersystem TableSort. Über ein Array an pneumatischen Ventilen können Fehlteile aus einem Materialstrom entfernen werden.

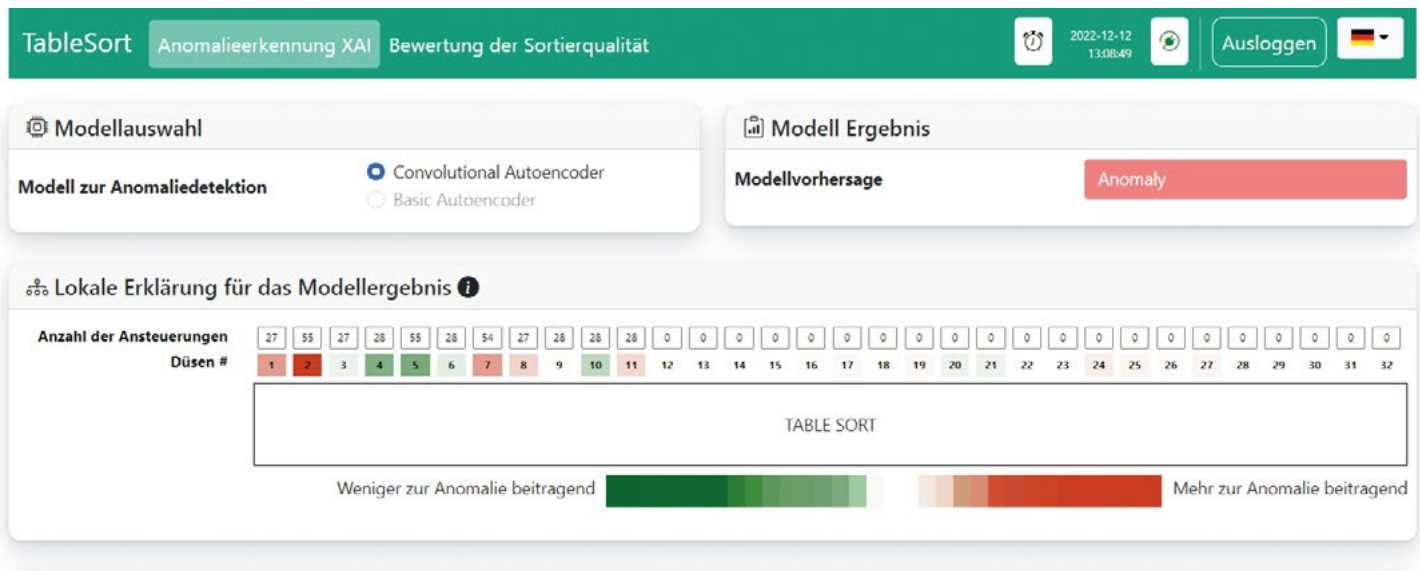


Abb. 2: Dashboard zur Erklärung von Klassifikationsergebnissen. Hier wurde eine Anomalie erkannt, die insbesondere durch die ersten pneumatischen Ventile im Array ausgelöst wurde (lokale Erklärung für das Modellergebnis)

TRANSPARENTE ANOMALIEERKENNUNG DANK XAI

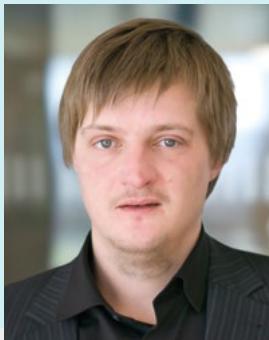
Wie das konkret aussehen kann, veranschaulicht unser voll funktionsfähiger Demonstrator für diesen Use Case: Sobald die Anlagenüberwachung mittels KI eine Anomalie erkennt, sorgen XAI-Methoden für eine für den Menschen verständliche Darstellung, welche Faktoren ausschlaggebend waren für die Einstufung als Anomalie. Hierbei kommen bspw. sogenannte Feature Attribution Verfahren wie SHAP zum Einsatz, die bewerten, wie wichtig die einzelnen Eingangsmerkmale für die Klassifikation waren. Da hier als Eingangsparameter die Schaltsequenzen der einzelnen pneumatischen Ventile analysiert werden, kann ein Werker direkt erkennen, welche Ventile sich ungewöhnlich verhalten, und daraus Rückschlüsse ziehen, welche nächsten Schritte eingeleitet werden müssen.

Der Einsatz von XAI-Methoden in der Qualitätssicherung unterstützt den Menschen also bei seinen Aufgaben und trägt dazu bei, den Aufwand und die Kosten für manuelle Überprüfungen zu reduzieren. Insbesondere bei der Prüfung großer Datenmengen können XAI-Methoden helfen, die wichtigsten Probleme zu identifizieren und somit eine schnelle Fehlererkennung und -behebung zu ermöglichen. So können XAI-Methoden in Kombination mit ML-Verfahren Schwachstellen in Prozessen identifizieren und damit zur Verbesserung der Qualität beitragen. Der Einsatz von XAI-Methoden kann also Prozesse effizienter und Entscheidungen transparenter machen.

Da im Rahmen der Industrie 4.0 der Einsatz von ML-Verfahren an Bedeutung gewinnt, um wettbewerbsfähig zu bleiben, wird auch die Unterstützung durch XAI-Methoden immer wichtiger. Indem sie Entscheidungs-

prozesse optimieren und die Fehlererkennung/-behebung beschleunigen, können sie entscheidend dazu beitragen, Produkte und Dienstleistungen auf einem hohen Qualitätsniveau zu halten.

Themen



Norbert Greifzu M. Sc.

Eingebettete Intelligente
Systeme (EIS)
Fraunhofer IOSB-AST

Telefon +49 3677 461-145

norbert.greifzu@iosb-ast.fraunhofer.de
www.iosb.fraunhofer.de/EIS



Prof. Dr.-Ing. Andreas Wenzel

Eingebettete Intelligente
Systeme (EIS)
Fraunhofer IOSB-AST

Telefon +49 3677 461-144

andreas.wenzel@iosb-ast.fraunhofer.de
www.iosb.fraunhofer.de/EIS

Qualität sichern und Ressourcen schonen: XAI in der industriellen Kunststoffproduktion

Ob beim Tippen auf der Tastatur, im Auto oder in der Küche: Gegenstände aus Kunststoff begegnen uns überall. Um Kunststoff-Objekte kostengünstig zu produzieren, setzt die Industrie verschiedene Verfahren ein. Dominant ist das Kunststoffspritzgießen: Das Verfahren ist am besten skalierbar, bezüglich der Kosten und der möglichen Ausstoßmenge. Das Ausgangsmaterial liegt hierbei in Form von Granulat vor. Dieses wird verflüssigt und unter Verwendung produktspezifischer Druck- und Temperaturprofile in eine ebenfalls temperierte Form gespritzt. Nach dem Aushärten wird die Form geöffnet und das Produkt entnommen. Die Verflüssigung des Ausgangsmaterials, das Einspritzen, das Öffnen und Schließen der Form wird durch die Kunststoffspritzgießmaschine geführt. Sie nimmt das Werkzeug mit der entsprechenden Produktform auf und erlaubt die Auswahl einer Reihe von Prozessparametern für die anschließende Produktion.

FLEXIBLE PRODUKTION HAT NICHT NUR VORTEILE

Durch den Austausch des Werkzeugs kann eine Maschine verschiedene Produkte herstellen. Das hat den Vorteil, dass die Anzahl der Maschinen und der benötigten Stellflächen gesenkt werden können, was Kosten und Ressourcen spart. Die gewonnene Effizienz und Flexibilität bergen jedoch auch Risiken. Da die Flexibilitätsreserven nach und nach ausgeschöpft wurden, orientiert sich die Produktion heute am »Just in Time«-Prinzip: Dabei werden nacheinander mehrere Werkzeuge auf eine Maschine gerüstet, um jeweils eine bestimmte Menge von einem Produkttyp nach dem anderen herzustellen. Bei jeder dieser Umrüstungen muss der Prozess erneut auf das herzustellende Produkt justiert und Kühlzeiten, Temperatur- und Druckprofile des neuen Produkts an der Maschine eingestellt werden. Zusätzlich unterliegen sowohl

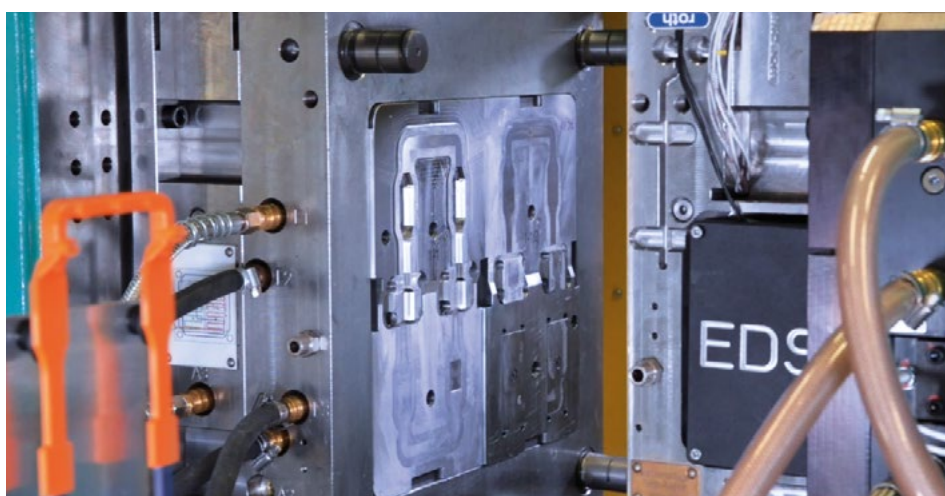


Abb. 1: Ein geöffnetes Spritzgießwerkzeug. In der Bildmitte ist die Form für das Produkt erkennbar. Die kompakte Mess- und Auswerteelektronik befindet sich in dem grauen Gehäuse mit der Aufschrift EDS in der rechten Bildhälfte.

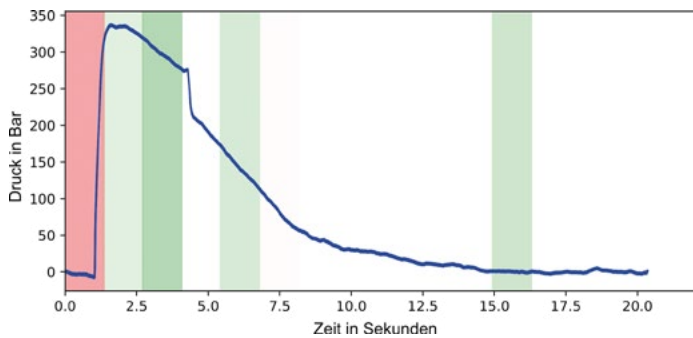


Abb. 2: Erklärung mittels LIME. Das Sensorsignal wird hierfür entlang der Zeitachse in Abschnitte aufgeteilt und der Beitrag jedes Abschnitts zur Prognose untersucht. Auf einer Farbskala von Rot über Weiß bis Grün wird der Beitrag des Abschnitts auf die Prognose von entgegengesetzt über neutral bis zustimmend angegeben.

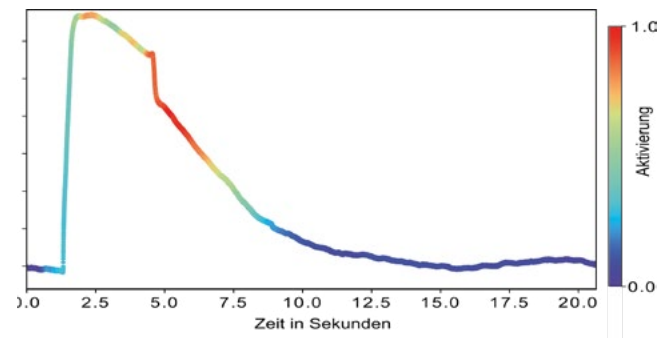


Abb. 3: Erklärung mittels CAM. Im Gegensatz zu LIME wird das Sensorsignal nicht in feste Abschnitte geteilt, sondern der Beitrag zur Prognose ist kontinuierlich auf einer Farbskala von Blau mit der Bedeutung »kein Beitrag« bis Rot mit der Bedeutung »hoher Beitrag« dargestellt.

die Spritzgießmaschinen als auch die Werkzeuge einem gewissen Verschleiß; befinden sie sich in unterschiedlichen Phasen ihres Lebenszyklus, können abweichende Produktionsparameter nötig sein, um eine gleichbleibende Qualität zu gewährleisten. Die Einrichtung und Überwachung dieser Produktionsprozesse erfordern ein hohes Maß an Expertise. Angesichts des wachsenden Fachkräftemangels und steigender Qualitätsanforderungen wird die Situation für Produzenten in Deutschland zunehmend problematisch.

RESSOURCEN SCHONEN DURCH QUALITÄTS PROGNOSEN

Um kunststoffverarbeitende Betriebe bei diesem umfangreichen Aufgabenspektrum zu unterstützen, haben wir mit der Hochschule Schmalkalden eine praktische Lösung entwickelt: Über Sensoren in der Spritzgießform werden Prozessdaten wie Druck- und Temperaturverläufe erfasst, um die Qualität der hergestellten Produkte vorherzusagen. Dazu wird eine kompakte Mess- und Auswertelektronik am Werkzeug angebracht

(siehe Abbildung 1), die das Ergebnis der Qualitätsprognose durch eine Ampel anzeigt. Eigens zugeschnittene KI-Modelle lernen direkt aus den Prozessdaten und benötigen keine zusätzliche Merkmalsextraktion. Mit diesem System werden kritische Qualitätsparameter für viele Situationen sehr genau vorhergesagt. Jedoch kann die abgegebene Prognose nicht unter allen Umständen exakt sein. Die Akzeptanz solcher KI-Lösungen in der Anwendungsdomäne kann durch Erklärungsansätze für einzelne Prognosen deutlich gesteigert werden. Für die konkrete Problemstellung wurde eine Toolbox mit mehreren Algorithmen entwickelt, welche dem Nutzenden die Relevanz einzelner Zeitsignalabschnitte für das Prognoseergebnis graphisch darstellt. Dabei kommen unter anderem modellagnostische Ansätze wie LIME (Local Interpretable Model-agnostic Explanations) [1] und modellspezifische Ansätze wie CAM (Class Activation Mapping) [2] zum Einsatz. In Abbildung 2 und 3 sind beispielhafte Annotationen mittels der genannten Methoden gegenübergestellt. Hierdurch können Anwendende mit Expertise entscheidungs-

relevante Signalabschnitte sehr gut erkennen und deren Plausibilität mit dem eigenen Wissen und der eigenen Erfahrung abgleichen. Für die Validierung der Modelle ist dies eine wesentliche Funktionalität, welche gleichzeitig dem Anwendenden ein Bild über die Verlässlichkeit des KI-Modells vermittelt. Durch ein frühzeitiges Erkennen von Qualitätsschwankungen wird zudem der Verbrauch von Ressourcen wie Energie und Kunststoffgranulat minimiert.

[1] S. M. Lundberg und S.-I. Lee, »A Unified Approach to Interpreting Model Predictions«, in Proceedings of the 31st International Conference on Neural Information Processing Systems, Red Hook, NY, USA, 2017, S. 4768–4777.

[2] B. Zhou, A. Khosla, A. Lapedriza, A. Oliva, und A. Torralba, »Learning Deep Features for Discriminative Localization«, ArXiv151204150 Cs, Dez. 2015. Verfügbar unter: <http://arxiv.org/abs/1512.04150>.

Die Bildauswertung ist ein wichtiges Anwendungsgebiet der Künstlichen Intelligenz. Dominierend in diesem Bereich sind Modelle des Deep Learnings, da diese besser mit dem umfangreichen Informationsgehalt von Bildern umgehen können als andere Formen von KI-Modellen. Dies liegt primär daran, dass man Deep-Learning-Modelle beliebig komplex entwerfen und somit in die Lage versetzen kann, komplexe semantische Muster in Bildern zu erkennen und darauf basierende Entscheidungen zu treffen.

Diese Leistungsfähigkeit hat jedoch auch ihre Kosten: Zum einen werden große Mengen aufbereiteter Bilddaten zum Trainieren der KI benötigt. Entsprechende Datensätze sind heutzutage für viele Anwendungen verfügbar. Zum anderen führt der massiv datengetriebene Einsatz komplexer Modelle dazu, dass die Entscheidungsfindung bei der erlernten Aufgabe selbst von den Modellentwickler*innen nicht mehr klar nachvollziehbar ist.

Es ist somit nicht auszuschließen, dass Fehlverhalten eintrainiert wurde und in der Anwendung zu unerwünschten Konsequenzen führt.

MARKIERUNG ENTSCHEIDUNGS-RELEVANTER BEREICHE

Sowohl Verfahrensentwicklerinnen wie Anwender profitieren deshalb von Erklärverfahren, die Entscheidungen dieser KI-Modelle nachvollziehbarer machen. Hierfür gibt es verschiedene Ansätze. Die in der Bildverarbeitung geeignetste Gruppe von XAI-Methoden sind Verfahren zur Attribuierung: Das Erklärverfahren ermittelt, welche Aspekte der Eingabedaten für ein KI-Modell von hoher Relevanz für die getroffene Entscheidung sind. Als konkretes Ergebnis werden Bereiche im Eingabebild oder -video gekennzeichnet, die zur Entscheidung beigetragen haben.

Beispielsweise kann ein Verfahren der visuellen Attribuierung das Verhalten eines Fahr-



Dr.-Ing. Arne Schumann

Videoauswertesysteme (VID)
Fraunhofer IOSB Karlsruhe

Telefon +49 721 6091-614
arne.schumann@iosb.fraunhofer.de
www.iosb.fraunhofer.de/VID



Abb. 1: Markierung der relevanten Regionen, die zur Klassifikation des Fahrzeugs führen. Relevante Regionen sind rot markiert, irrelevante Regionen blau. Somit wird deutlich: Das unterliegende KI-Verfahren stützt sich auf aussagekräftige Bereiche vorne am Fahrzeug.



Abb. 2: Detektion von Fahrzeugen (PKW mit roten, Lieferwagen mit schwarzen Rahmen) und dazugehörige Erklärungen. Erwartet war die Beobachtung, dass im Hintergrund die Auflösung der Objekte nicht mehr ausreichend ist, um verlässliche Detektionen zu liefern. Die Erklärungen im Vordergrund sind jedoch ebenfalls schwach, auch wenn die Objekte in diesem Fall erfolgreich detektiert sind. Hier empfiehlt sich dem Entwickler also, das Training bzgl. größerer Objekte anzupassen, um die sich abzeichnende Schwäche auszugleichen.

zeugklassifikators nachvollziehbarer machen. Solche KI-basierten Klassifikatoren können anhand von Bildern von Fahrzeugen feingranular den Hersteller, das Modell und teilweise sogar das Baujahr bestimmen. Bei derart spezialisierten Aufgaben können Anwender*innen der Verfahren die Ergebnisse nur mit umfangreichem Hintergrundwissen verifizieren. Die Information, anhand welcher Bereiche des Bildes bzw. Fahrzeuges die Entscheidung getroffen wurde, schafft hier neue Möglichkeiten: Sie versetzt auch Menschen ohne umfangreiche Kenntnisse bezüglich Automarken und -modelle in die Lage, die Plausibilität des Ergebnisses zu prüfen – etwa indem sie sicherstellen, dass die markierten Bildbereiche tatsächlich auf dem Fahrzeug und dort ggf. auf relevanten Teilen, wie dem Symbol des Herstellers, liegen.

QUALITÄTSSICHERUNG UND TARNBEWERTUNG

Erklärverfahren schaffen also Vertrauen in den Einsatz der KI-Modelle. Ein weiterer

wesentlicher Einsatzbereich ist die Qualitätssicherung im Prozess der KI-Verfahrensentwicklung. So können Entwickler*innen leicht erkennen, wenn sich ein KI-Verfahren bei der Klassifikation oder Detektion von Objekten in Bildern auf nur zufällig korrelierte Information wie Wasserzeichen im Bild stützt, wenn es nur bestimmte Teile von Objekten erlernt, die später im Anwendungsfall womöglich verdeckt sein könnten, oder wenn es Schwächen bei bestimmten Bildauflösungen oder Bildcharakteristiken aufweist. Ermittelt man derartige Fehlverhalten oder Schwächen, lassen diese sich in einem iterierten Entwicklungszyklus aus den KI-Verfahren herausbilden.

Ein dritter Aspekt, unter dem XAI im Bereich von KI-Verfahren zur Bildauswertung wertvolle Dienste leistet, betrifft die Täuschung von bzw. Tarnung vor KI-Verfahren sowie die Ermittlung von Angriffen auf KI-Verfahren. So werden am Fraunhofer IOSB auf XAI basierende Methoden verwendet, um zu bewerten, wie effektiv Tarnungen gegenüber

KI-basierter Objektdetektion wirken. Ebenso können Erklärverfahren eingesetzt werden, um Angriffe auf KI-Modelle zu erkennen, die auf Manipulation der Eingabedaten basieren. Solche Angriffe führen zu starken Veränderungen in den Charakteristiken der resultierenden Erklärungen, die wiederum die Identifikation des Angriffs ermöglichen.

Erklärverfahren unterstützen den Entwicklungsprozess und den Einsatz von KI-Verfahren in der Bildauswertung somit in vielerlei Hinsicht. Sie erhöhen die Qualität der Ergebnisse, die Robustheit gegenüber Störeinflüssen und das resultierende Vertrauen der Anwender*innen. Ihr Einsatz und ihre Weiterentwicklung ist und bleibt daher ein wichtiger begleitender Prozess bei der Erforschung und Integration KI-basierter Bildauswertungsverfahren am Fraunhofer IOSB.

Themen



Dr. Achim Kuwertz

Interoperabilität und
Assistenzsysteme (IAS)
Fraunhofer IOSB Karlsruhe

Telefon +49 721 6091-231
achim.kuwertz@iosb.fraunhofer.de
www.iosb.fraunhofer.de/IAS



Dr. Almuth Müller

Interoperabilität und
Assistenzsysteme (IAS)
Fraunhofer IOSB Karlsruhe

Telefon +49 721 6091-627
almuth.mueller@iosb.fraunhofer.de
www.iosb.fraunhofer.de/IAS

Semantische XAI-Verfahren für natürlichsprachliche Erklärungen

Im aktuell sehr aktiven Forschungsgebiet eXplainable AI (XAI) existieren viele Ansätze zur Erklärung von maschinellen Lernverfahren. Die eingesetzten Methoden unterscheiden sich dabei genauso wie die Arten der erzeugten Erklärungen: Ihre Bandbreite reicht von annotierten Bildern über Diagramme relevanter Merkmale bis hin zu natürlichsprachlichen Erklärungen. Natürlichsprachliche Erklärungen bieten einen Vorteil, da sie die Ergebnisse von Lernverfahren einfach verständlich erläutern. Sie sind jedoch nicht leicht zu erzeugen. Abhilfe schafft ein Verfahren, dass nicht auf der Ebene von Merkmalen oder Pixeln arbeitet, sondern auf einer semantischen Ebene. Für den Menschen intuitiv verständliche Begriffe stehen im Fokus der erzeugten Erklärungen. Die Erklärungen erfolgen daher direkt in natürlichsprachlicher Form.

SEMANTISCHE ERKLÄRVERFAHREN ORIENTIEREN SICH AM MENSCHEN

Grundlage dieses semantischen Ansatzes für XAI [1] sind formale Modelle, die zu erklär-

ende Realweltzusammenhänge abstrahiert darstellen. Als Anwendungsbeispiel wird die Detektion von Fahrzeugen in Bildern betrachtet, wie in Abbildung 1 dargestellt. Semantische Erklärverfahren orientieren sich daran, wie ein Mensch überprüfen würde, ob ein Detektionsergebnis plausibel ist. Für das Anwendungsbeispiel sind dabei Fragen relevant, wie »Besitzt das erkannte Objekt die intuitiv erwarteten Teile?« – hier z. B. Räder, Türen oder Scheinwerfer, und »Tritt das erkannte Objekt im erwarteten Zusammenhang auf?« – hier z. B. auf einer Straße und nicht im Wasser. Derartige Zusammenhänge können in einem Wissensmodell hinterlegt werden (siehe Abbildung 2). Wissensmodelle werden von Wissensingenieur*innen erstellt; für eine semi-automatisierte Erstellung können auch öffentlich verfügbare Wissensdatenbanken (in Sinne von Linked Open Data) wie DBPedia genutzt werden. Auch der Einsatz kundenspezifischer Modelle ist möglich.

Im semantischen Erklärverfahren wird das Wissensmodell genutzt, um die Ergebnisse des maschinellen Lernverfahrens etwa auf



Abb. 1: Semantische Plausibilitätsprüfung für eine Fahrzeugdetektion: Für die detektierte Klasse Fahrzeug werden anhand des Wissensmodells die erwarteten erkennbaren Teile ermittelt und anschließend im Bild detektiert (vgl. [2]). In Abhängigkeit ihres Vorhandenseins wird die Plausibilität des Detektionsergebnisses bewertet und mittels einer natürlichsprachlichen Erklärung beschrieben (Testbild: [3]).

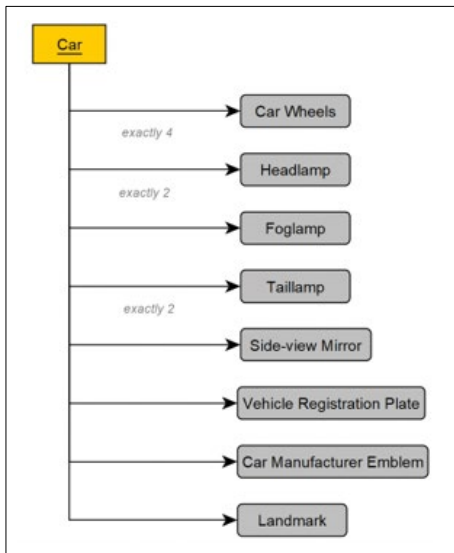


Abb. 2: Beispiel eines Wissensmodells für semantische Erklärverfahren zur Fahrzeugdetektion. Im Modell werden die für Fahrzeuge optisch erkennbaren Teile beschrieben. Dieses Wissen kann in Erklärverfahren eingebunden werden.



Abb. 3: Semantische Relevanzprüfung für die Fahrzeugdetektion: Anhand des Wissensmodells werden die relevanten Teile für die Klasse Fahrzeug im Bild ermittelt (vgl. [4]) und entfernt. Für das derart veränderte Bild wird erneut das Detektionsergebnis berechnet und mit dem Ergebnis des Originalbilds verglichen. Aus dem Unterschied ergibt sich die Relevanz des entfernten Teiles für das Detektionsverfahren (Testbild [3]).

Plausibilität zu prüfen oder um zu beurteilen, wie relevant einzelne Teile eines erkannten Objektes bzw. sein Kontext für das Ergebnis waren. Als Ergebnis einer Plausibilitätsprüfung (siehe Abbildung 1) kann das semantische Erklärverfahren z. B. feststellen, dass die Detektion eines Fahrzeuges plausibel ist (oder nicht), da auch Teile wie Räder, Türen etc. (oder eben nicht) auf dem Bild erkannt wurden. Für die Erkennung der Teile und des Kontextes sind weitere Bilderkennungsverfahren (z. B. [2] für Schlüsselpunkte an Fahrzeugen) und deren Verknüpfung zum Wissensmodell notwendig. Ist diese Verknüpfung erfolgt, verläuft die semantische Plausibilitätsprüfung automatisch. Als Ergebnis wird eine natürlichsprachliche Erklärung erzeugt und durch Informationen, wie im Bild erkannte Teile, ergänzt.

NATÜRLICHSPRACHLICHE ERKLÄRUNGEN FÖRDERN VERSTÄNDNIS FÜR MASCHINELLE LERNVERFAHREN

Für eine semantische Relevanzprüfung werden dagegen entsprechende Teile im Originalbild entfernt (z. B. durch Schwärzen),

um im Sinne eines kontrafaktischen Erkläransatzes feststellen zu können, wie wichtig diese Teile für das zu prüfende KI-Ergebnis waren (siehe Abbildung 3). Auch hier kann das Ergebnis in Form einer natürlichsprachlichen Erklärung dargestellt werden: »Die Frontscheibe im betrachteten Bild ist relevant für die Erkennung als ein Objekt der Klasse Fahrzeug, da die Konfidenz der Erkennung um sechs Prozent absinken würde, wenn sie nicht im Bild wäre«. Eine Relevanzprüfung kann dabei nicht nur auf einzelnen Bildern durchgeführt, sondern auch über ganze Datensätze; so können etwa verschiedene Verfahren zur Fahrzeugdetektion besser verglichen werden.

Natürlichsprachliche Erklärungen besitzen über diese Anwendungsfälle hinaus ein großes Potential, um maschinelle Lernverfahren im Allgemeinen Menschen leichter zugänglich zu machen. Sie sind anpassbar an ein breites Nutzer*innen-Spektrum mit unterschiedlichem Wissenshintergrund zu maschinellen Lernverfahren.

- [1] Sander, Jennifer, and Achim Kuwertz. »Supplementing Machine Learning with Knowledge Models Towards Semantic Explainable AI« in: International Conference on Human Interaction and Emerging Technologies - Future Applications (IHET-AI), 2021.
- [2] Wang, Zhongdao et al. »Orientation Invariant Feature Embedding and Spatial Temporal Regularization for Vehicle Re-Identification« in: The IEEE International Conference on Computer Vision (ICCV), 2017.
- [3] Krause, Jonathan, and Stark, Michael, and Deng, Jia, and Fei-Fei Li. »3D Object Representations for Fine-Grained Categorization,« in: 4th IEEE Workshop on 3D Representation and Recognition at ICCV 2013 (3dRR-13), 2013.
- [4] Wu, Yuxin, and Kirillov, Alexander, and Massa, Francisco, and Lo, Wan-Yen, and Girshick, Ross. »Detectron2«, URL: <https://github.com/facebook-research/detectron2>, 2019.

Themen

Erklärbare KI in der Medizin aus Sicht von User Interface Design und Usability Engineering

Ein potenzielles Anwendungsfeld für XAI ist die Medizin: Insbesondere sollen zunehmend klinische Entscheidungsunterstützungssysteme (clinical decision support systems, kurz CDSS) in medizinische Arbeitsabläufe integriert werden. Dabei spielt KI eine wichtige Rolle. So können mit KI z.B. einzelne Patientengeschichten analysiert und mit großen Kohorten verglichen werden, um individualisierte und damit effektivere Behandlungen zu entwickeln. Bei zunehmend datenintensiven Diagnoseinstrumenten wie z.B. Genomsequenzierung sind KI-Methoden sogar die einzige Möglichkeit, diese zu bewältigen.

Allerdings wird die Einführung dieser Technologie schwierig oder sogar unmöglich sein, wenn diese Systeme keine verständliche Begründung ihrer errechneten Empfehlungen in Form von benutzerfreundlichen und gebrauchstauglichen Schnittstellen und nachvollziehbaren Erklärungen liefern können.

Ein KI-basiertes CDSS muss zum einen also effizient und effektiv arbeiten, aber zwingend auch nutzerzentriert gestaltet sein, d.h. für die spezifischen Anwender*innen nachvollziehbar darstellen können, wie es zu Schlussfolgerungen und Empfehlungen kommt. Denn in der Medizin werden Entscheidungen über Diagnose und Therapie auch in absehbarer Zukunft von menschlichen Ärzt*innen getroffen und verantwortet werden müssen. Die Gebrauchstauglichkeit (Usability) ist daher ein ebenso entscheidender Erfolgsfaktor wie die technische Funktion, z.B. die Interoperabilität solcher Systeme. Ein weiterer Faktor ist die Akzeptanz. Nur wenn diese Systeme einen echten Mehrwert in der

ärztlichen Arbeit bringen und zudem im anspruchsvollem klinischen Arbeitsalltag gut nutzbar sind, werden Ärzt*innen sie in die Entscheidungsfindung einbeziehen. Diese Fragen sind noch nicht gelöst und ein akutes Forschungsthema.

DER BEITRAG VON INTERACTION DESIGN UND USABILITY ENGINEERING

Wir haben es mit lernenden Systemen zu tun, d.h. mit einem völlig neuen Gestaltungsparadigma. Es lernen nicht nur die Benutzer*innen und verändern ggf. ihr Verhalten im Laufe der Zeit, sondern auch die Maschinen selbst. Wenn wir mittels XAI das Verhalten der KI-Systeme nachvollziehbar machen wollen, augmentieren wir eine komplexe Technologie mit einer noch komplexeren Technologie. Dies auf Interfaceebene zu gestalten, erfordert neue Designansätze und Methoden. Daran forschen wir hier am IOSB.

Designer*innen müssen zunächst verstehen, wie Maschinenerklärungen funktionieren und diese dann in eine Form übersetzen, die eine möglichst unbeeinflusste und informierte Entscheidungsfindung begünstigt. Neueste Forschung auf diesem Gebiet zeigt, dass der bisher gewählte Ansatz, die innere Arbeitsweise der KI-Modelle offenzulegen, oft nicht zielführend ist. Es bedarf also umfassender Designforschung, um die Schnittstellen intelligenter Systeme im Sinne ihrer hochspezialisierten Anwender*innen zweckdienlich zu gestalten.

Der erste Schritt zu solchen Interfaces ist, die sog. mentalen Modelle der Nutzer*innen zu



Dipl.-Des. Henrik Mucha M.Sc.

Interaktive Analyse und
Diagnose (IAD)
Fraunhofer IOSB Karlsruhe

Telefon +49 721 6091-603
henrik.mucha@iosb.fraunhofer.de
www.iosb.fraunhofer.de/IAD

KONTAKT

Themen



Mathias Anneken M.Sc.

**Interaktive Analyse und
Diagnose (IAD)
Fraunhofer IOSB Karlsruhe**

Telefon +49 721 6091-174
mathias.anneken@iosb.fraunhofer.de
www.iosb.fraunhofer.de/IAD



Dr.-Ing. Arne Schumann

**Videoauswertesysteme (VID)
Fraunhofer IOSB Karlsruhe**

Telefon +49 721 6091-614
arne.schumann@iosb.fraunhofer.de
www.iosb.fraunhofer.de/VID

Mit der XAI-Toolbox KI-Modelle durchleuchten

Die Forschung im Bereich der Künstlichen Intelligenz (KI) hat in den letzten Jahren in vielen Anwendungsgebieten enorme Fortschritte gemacht. Mit ihrer Fähigkeit komplexe Muster in Daten zu erkennen und Entscheidungen zu treffen, haben KI-Systeme das Potenzial, viele Bereiche zu revolutionieren, von der Produktion über die Gesundheitsversorgung bis hin zur Verkehrssicherheit. Trotz dieser Fortschritte bleibt das Vertrauen in KI-Systeme oft gering, da ihre Entscheidungen und Vorhersagen schwer zu verstehen oder nachzuvollziehen sind.

Hier kommt Explainable Artificial Intelligence (XAI) ins Spiel. XAI-Systeme zielen darauf ab, die Entscheidungen und Vorhersagen von KI-Systemen transparent und verständlich zu machen. Dies ist besonders wichtig in sicherheitskritischen Anwendungen, in denen unerklärliche Entscheidungen eines KI-Systems schwerwiegende Konsequenzen haben können.

Um die Entwicklung von XAI-Systemen zu erleichtern, wurde am Fraunhofer IOSB eine Toolbox entwickelt, die es ermöglicht, verschiedene XAI-Verfahren schnell und einfach in bestehende KI-Systeme zu integrieren. Außerdem lassen sich neue KI-Verfahren leicht integrieren, verschiedene Parameter können schnell erprobt werden und die Übertragbarkeit auf eigentliche Anwendungen verbessert.

Die Toolbox ist bereits für verschiedene Datenformate wie Bilder, tabellarische Daten und Zeitreihen (mit und ohne Georeferenzierung) implementiert, und für verschiedene Anwendungsdomänen erprobt, einschließlich maritimer Überwachung, Produktion und Bildklassifikation. Durch die Toolbox werden verschiedene ML-Frameworks wie TensorFlow, PyTorch und Scikit-learn unterstützt; so können Entwickler*innen die Toolbox effizient in ihre vorhandene KI-Infrastruktur integrieren.

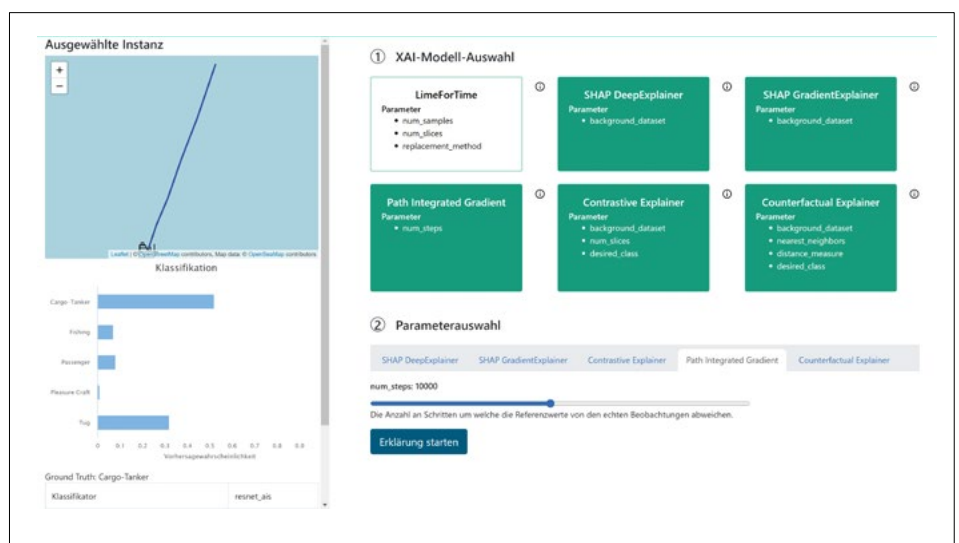


Abb. 1: Links: Automatisch erkannte Schiffsklasse basierend auf einer ausgewählten Trajektorie. Rechts: Auswahl und Parametrisierung der XAI-Modelle.

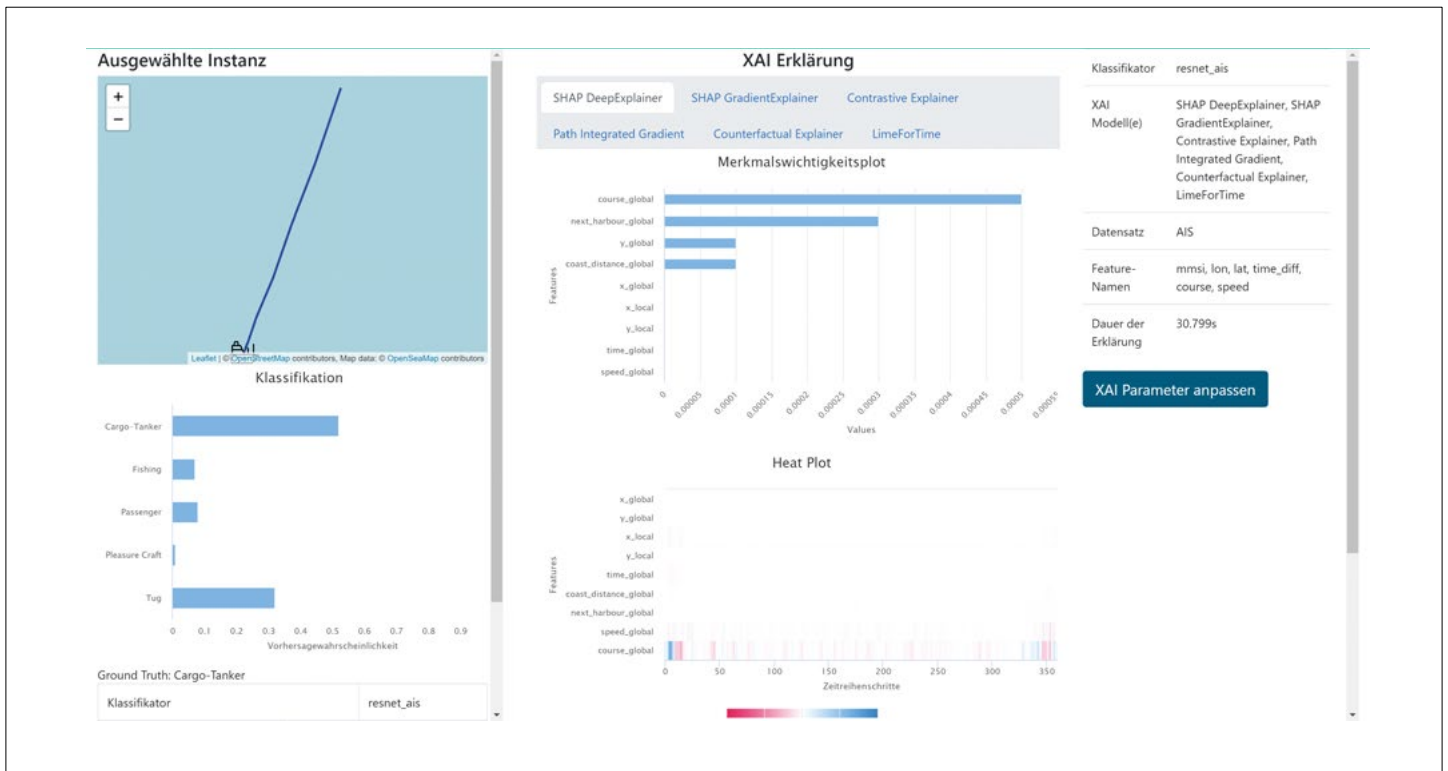


Abb. 2: Erklärungen basierend auf XAI-Modellen für räumlich-zeitliche Daten.

Eines der zentralen Merkmale der Toolbox ist ihre Fähigkeit, eine Vielzahl von Erklärungsverfahren anzubieten. Sie umfassen Interpretierbarkeit, Visualisierung, Attribution und Erklärung von KI-Modellen. Einige Beispiele hierfür sind LIME, SHAP, Integrated Gradients und kontrafaktische Erklärungen: Diese Verfahren ermöglichen es Benutzer*innen, das Verhalten von KI-Modellen besser zu verstehen, indem sie visuelle und textbasierte Erklärungen liefern. Auf Basis dieser Erklärungen können Menschen die Entscheidungsfindung von KI-Systemen überprüfen und validieren.

Darüber hinaus ermöglicht die Toolbox eine schnelle Verknüpfung von XAI-Verfahren mit KI-Modellen. Dies bedeutet, dass Benutzer*innen ohne großen Aufwand verschiedene XAI-Verfahren auf ihre KI-Modelle anwenden können, um die Entscheidungen der KI besser zu verstehen. Auf diese Weise können Probleme identifiziert und entsprechende Änderungen vorgenommen werden, um die Vorhersagegenauigkeit und -verlässlichkeit zu verbessern.

Ein weiteres Merkmal ist eine einfache Integration neuer Datensätze. Dies ist besonders nützlich für Anwender*innen, die spezielle Datensätze verwenden oder ihre eigenen Daten generieren möchten. Mit der Toolbox können Benutzer*innen ihre eigenen Datensätze importieren und nahtlos integrieren, um die Leistung ihrer KI-Modelle und die dazugehörigen XAI-Erklärungen auf diesen Daten effizient zu testen und zu verbessern.

Zusätzlich bietet die Toolbox eine einfache Wiederverwendung von Verfahren, Datensätzen und Visualisierungskomponenten: Benutzer*innen können bereits vorhandene Verfahren und Komponenten verwenden und an ihre eigenen Anforderungen anpassen. Auf diese Weise können Benutzer*innen wertvolle Zeit und Ressourcen sparen und sich auf die Entwicklung und Optimierung ihrer KI-Modelle konzentrieren.

Zusammenfassend bietet die XAI-Toolbox des Fraunhofer IOSB eine umfassende Lösung zur Interpretation von KI-Modellen in

verschiedenen Anwendungsdomänen. Die Toolbox wurde entwickelt, um den Prozess der Integration von KI-Modellen zu erleichtern und gleichzeitig sicherzustellen, dass die Modelle transparent und interpretierbar bleiben. Die Verknüpfung von XAI-Verfahren mit KI-Modellen ermöglicht es den Anwender*innen, schnell und einfach verschiedene Parameter zu erproben und die optimale Konfiguration für ihr spezifisches Szenario zu finden. Dies ermöglicht uns auf einfache Weise neue Anwendungsfälle von Kund*innen zu untersuchen und dadurch neue KI-Ansätze zu erproben bzw. Verbesserungsvorschläge für bestehende Ansätze zu entwickeln.

Nähere Informationen zum Thema auf: <https://www.iosb.fraunhofer.de/de/projekte-produkte/XAIToolbox.html>

**Fraunhofer-Institut für
Optronik, Systemtechnik
und Bildauswertung IOSB**

info@iosb.fraunhofer.de

www.iosb.fraunhofer.de

FRAUNHOFER IOSB KARLSRUHE

Fraunhoferstraße 1

76131 Karlsruhe

Telefon +49 721 6091-0

FRAUNHOFER IOSB ETTLINGEN

Gutleuthausstraße 1

76275 Ettlingen

Telefon +49 7243 992-0

**INSTITUTSTEIL FÜR ANGEWANDTE
SYSTEMTECHNIK IOSB-AST**

Am Vogelherd 90

98693 Ilmenau

Telefon +49 3677 461-0

info@iosb-ast.fraunhofer.de

www.iosb-ast.fraunhofer.de

**INSTITUTSTEIL FÜR INDUSTRIELLE
AUTOMATION IOSB-INA**

Campusallee 1

32657 Lemgo

Phone +49 5261 94290-22

juergen.jasperneite@iosb-ina.fraunhofer.de

www.iosb-ina.fraunhofer.de

Zweigstellen & Kontaktbüro

**FORSCHUNGSGRUPPE AKTIVE
LASERFASERN**

c/o HENSOLDT Optronics GmbH

Carl-Zeiss-Str. 22

73447 Oberkochen

**FORSCHUNGSGRUPPE IT-SICHERHEIT
FÜR KRITISCHE INFRASTRUKTUREN
ENERGIE UND WASSER**

Wilhelmsplatz 11

02826 Görlitz

**FORSCHUNGSGRUPPE SMART
OCEAN TECHNOLOGIES SOT**

Alter Hafen Süd 6

18069 Rostock

**FRAUNHOFER-ZENTRUM FÜR DIE
SICHERHEIT SOZIO-TECHNISCHER
SYSTEME SIRIOS**

c/o Fraunhofer FOKUS

Kaiserin-Augusta-Allee 31

10589 Berlin

KONTAKTBÜRO BEIJING

Unit 0602G, Tower D1

DRC Liangmaqiao Office Building

19 Dongfangdonglu, Chaoyang District

100600 Beijing, PR China

muh@fraunhofer.com.cn