Universität Konstanz

Schlussbericht

Verbundprojekt: Vertrauenswürdige Künstliche Intelligenz für polizeiliche Anwendungen (VIKING)

Teilvorhaben: Visual Analytics für vertrauenswürdige und erklärbare Künstliche Intelligenz

VIKING

Stand 11.2025

FKZ: 13N16242

Gefördert durch:

Bundesministerium für Forschung, Technologie und Raumfahrt

Schlussbericht

Vertrauenswürdige Künstliche Intelligenz für polizeiliche Anwendungen (VIKING)

Visual Analytics für vertrauenswürdige und erklärbare Künstliche Intelligenz

Metadaten

Vorhabenbezeichnung	Vertrauenswürdige Künstliche Intelligenz für polizeiliche Anwendungen
Projekt-Akronym	VIKING
Förderkennzeichen	13N16242
Auftraggeber	Bundesministerium für Bildung und Forschung (BMBF)
Projektträger	VDI Technologiezentrum GmbH (VDI)
Projektkoordinator	Dr. Michael Dose (IDEMIA)
Projektlaufzeit	01.01.2022 –31.03.2025
Teilvorhaben	Visual Analytics für vertrauenswürdige und erklärbare Künstliche Intelligenz
Antragssteller	Universität Konstanz (UKON)
Anschrift	Universität Konstanz Universitätsstr. 10 78464, Konstanz
Projektleitung	Prof. Dr. Daniel A. Keim

Autor: Yannick Metz, Dr. Maximilian T. Fischer, Prof. Dr. Daniel A. Keim

Ausgabe: 1

Stand: November 2025

Danksagung

Wir bedanken uns bei dem Bundesministerium für Bildung und Forschung (BMBF), mittlerweile umbenannt zu dem Bundesministerium für Forschung, Technologie und Raumfahrt (BMFTR), für die Förderung des Forschungsprojekts VIKING, Teilprojekt UKON (FKZ: 13N16242) und bei dem Projektkoordinator VDI, insbesondere Herrn Dr. Röhrig und Herr Hasenauer, für ihre Unterstützung bei der Durchführung des Projekts. Wir möchten zudem dem Koordinator IDEMIA unter Leitung von Herrn Dr. Michael Dose sowie allen Projektpartnern für die sehr gute und stets professionelle Zusammenarbeit unseren herzlichen Dank aussprechen.

Inhaltsverzeichnis

1. Kurze Darstellung

1.1. Aufgabenstellung

Das Hauptziel des Verbundprojekts "*Vertrauenswürdige Künstliche Intelligenz für polizeiliche Anwendungen*“ (VIKING) ist die Entwicklung von KI-Methoden zur Unterstützung der polizeilichen Ermittlungsarbeit unter besonderer Berücksichtigung ethischer und rechtlicher Aspekte. Im Rahmen des Projekts wurde die Eignung aktueller KI-Methoden überprüft und in Demonstratoren umgesetzt. Neben der Genauigkeit und Robustheit wurden für die eingesetzten Methoden die Attribute Transparenz, Erklärbarkeit und Fairness bzw. Debiasing untersucht und entsprechende Lösungen implementiert. Gemeinsam wurden übergreifende Standards und Richtlinien für die Sicherstellung rechtlicher und ethischer Aspekte entworfen und umgesetzt.

Der vorliegende Schlussbericht behandelt das Teilprojekt der Universität Konstanz (UKON), das sich mit der technischen Umsetzung von Transparenz- und Erklärbarkeitsmethoden beschäftigt. Als technische und wissenschaftliche Leitung unterstützt die UKON bei der Umsetzung aller technischen Teilprojekte. Außerdem wirkte die UKON im Rahmen des Teilprojekts direkt an der Erstellung von Dokumenten zur DIN-Standardisierung und an Prüfkatalogen mit.

1.2. Voraussetzungen, unter denen das Vorhaben durchgeführt wurde

Die Arbeitsgruppe für Datenanalyse und Visualisierung der Universität Konstanz unter Leitung von Prof. Dr. Daniel A. Keim beschäftigt sich mit verschiedenen Forschungsthemen im Bereich Informationsvisualisierung, Data-Mining und Künstlicher Intelligenz (KI), mit einem Fokus auf Visual Analytics und Human-AI Teaming. Erklärbare Künstliche Intelligenz integriert die Forschungsschwerpunkte in einem praxisnahen Kontext. Der Fokus der Arbeitsgruppe liegt auf der Untersuchung skalierbarer visueller Analyseansätze, die eine Kombination aus automatisierten und interaktiven Ansätzen nutzen. Das Hauptziel besteht darin, Entscheidungsprozesse in großen, komplexen Informationsräumen durch interaktive Einbindung menschlicher Expertise zu unterstützen. Die Arbeitsgruppe hat in diesem Kontext im Laufe des letzten Jahrzehnts eine besondere Expertise in dem von Dr. Maximilian T. Fischer verantworteten Bereich der Zivilen Sicherheitsforschung erlangt. In diesem Rahmen war die Arbeitsgruppe erfolgreich an zahlreichen europäischen (VALCRI 608142/FP7, ASGARD 700381/H2020, VICTORIA 740754/H2020, SPARTA 830892/H2020) und nationalen BMBF-Projekten (VASA, FLORIDA, PEGASUS) im Themenfeld von Terrorismusbekämpfung, Organisierte Kriminalität und Kommunikationsanalyse sowie Signal- und Massendatenanalyse im Dark Web beteiligt. Insbesondere ist die Einbindung durch Erklärbarkeitsmethoden für komplexe Modelle wie Neuronale Netzwerke und deren Interaktionskonzepte durch Human-AI Teaming ein zentraler Schwerpunkt der Arbeitsgruppe, der durch diese Vorerfahrungen in VIKING eingebracht werden konnte.

1.3. Planung und Ablauf des Vorhabens

Während die ursprüngliche Planungsphase noch überschattet von der Corona-Pandemie stattfand, fand das Projekt größtenteils wieder unter normalen Bedingungen in Präsenz statt.

Am Beginn des Projekts stand eine Bedarfsanalyse für die einzelnen Teilprojekte und ein darauf aufbauender Entwurf der technischen Demonstratoren. Im Rahmen des ersten Konsortialmeetings fand ein gemeinsamer Workshop zu technischen Grundlagen von Erklärbarer KI statt. Als

technischer und wissenschaftlicher Koordinator war UKON in der Bedarfsanalyse der technischen Arbeitspakete eingebunden. Im Speziellen wurde der Einsatz von Attributionsmethoden und konzept-basierten Erklärbarkeitsmethoden als übergreifende Lösungen diskutiert und vorbereitet. Weiterhin wurden relevante rechtliche und technische Aspekte mit Partnern wie der DIN und dem IZEW diskutiert, um die Grundlagen für weitere Spezifikationen zu legen.

Im weiteren Verlauf des Projekts begleitete das Teilprojekt die Umsetzung der verschiedenen technischen Arbeitspakete. Im Rahmen monatlicher Treffen sowie während der Konsortialmeetings, wurde zur technischen Umsetzung und Bewertung von Designentscheidungen beigetragen. Das *explAIner* Framework wurde zu einer domänenagnostischen Software-Bibliothek erweitert, die unter Einhaltung strenger Datenschutzstandards die Erzeugung lokaler Erklärungen (z.B. Heatmaps von visuellem Input) ermöglicht. Das Werkzeug wurde unter anderem im Rahmen der Gesichts- und Spracherkennung getestet. Parallel zu den technischen Arbeitspaketen wurden die Dokumente zur Standardisierung ausgearbeitet. Dazu wurden Vorlagen zu Datasheets, Model Cards und XAI Fact Sheets im gegenseitigen Austausch der Partner definiert.

Die Ergebnisse der Arbeiten mündeten, neben der erfolgreichen Umsetzung innerhalb der technischen Teilprojekte, in einer Reihe von übergreifenden Implementierungen und Spezifikationen. Es wurde ein Demonstrator zur Erstellung von Erklärungen fertiggestellt und den anderen Projektpartnern zur Verfügung gestellt. Die entwickelten Vorlagen wurden in diesem Demonstrator als interaktive Check-Listen für alle Projektpartner zur Verfügung gestellt. Außerdem wirkte das Teilprojekt an der Evaluation von Demonstratoren mit assoziierten Ermittlern mit. Schließlich wurde unter Federführung von UKON die DIN SPEC 91517:2025-05 entworfen und veröffentlicht. Sie bildet einen einheitlichen Beitrag zur Standardisierung der Projektergebnisse.

Durch das Projekt wurden insgesamt zehn internationale wissenschaftliche Fachpublikationen, eine DIN-SPEC und je eine Bachelor- und Masterarbeit sowie eine Dissertation gefördert.

1.4. Angeknüpfter Wissenschaftlicher und technischer Stand

Das Teilprojekt baut auf dem Einsatz von domänenagnostischen Erklärbarkeitsmethoden und deren Einbettung in interaktive visuelle Demonstratoren auf. Diese Bereiche sind ein Hauptfokus der Arbeitsgruppe, die in diesem Gebiet international führend ist. Hierfür wurden auf Forschungsergebnisse der Gruppe im Bereich der erklärbaren KI, z.B. das *explAIner* Framework, sowie zahlreichen Sicherheitsforschungsprojekten wie VALCRI (EU), ASGARD (EU), FLORIDA (BMBF), VICTORIA (EU), SPARTA (EU) und PEGASUS (BMBF) aufgebaut.

1.5. Zusammenarbeit mit anderen Stellen

Neben intensiver Zusammenarbeit innerhalb des Konsortiums erfolgte ein reger Austausch mit den assoziierten Partnern und Dienststellen der teilnehmenden Polizeibehörden.

1.6. Zusammenfassung der Projektergebnisse

Der vorliegende Abschlussbericht umfasst die Ergebnisse des von UKON verantworteten Teilprojekts zur wissenschaftlichen und technischen Leitung des Verbundprojekts. In dieser Funktion trug die UKON sowohl zur erfolgreichen Umsetzung der einzelnen Demonstratoren bei als auch zur Entwicklung übergreifender Demonstratoren sowie Spezifikationsdokumenten wie der DIN SPEC.

2. Eingehende Darstellung

Im Folgenden wird eingehender über die Ergebnisse der Forschung berichtet.

2.1. Verwendung der Zuwendung und des erzielten Ergebnisses

Im Rahmen der eingehenden Darstellung fokussieren wir uns auf die Demonstratoren und Dokumente, die für die übergreifende Analyse entwickelt worden sind. Die durch dieses Teilprojekt unterstützten Demonstratoren der anderen APs sind ausführlich in den Abschlussberichten der anderen Teilprojekte beschrieben.

2.1.1. Kontext und Zielsetzung des VIKING-Projekts

Parallel zur fortschreitenden Digitalisierung der Gesellschaft wachsen auch die Anforderungen an Polizeibehörden mit diesen Entwicklungen Schritt zu halten. Im Rahmen der Beweisfindung für polizeiliche Ermittlungen müssen exponentiell wachsende Datenmengen ausgewertet werden, häufig verteilt über verschiedene Medien. Dies erfordert eine (teil-)automatisierte Auswertung, wobei Methoden der Künstlichen Intelligenz bzw. des maschinellen Lernens eingesetzt werden können. Moderne Methoden können komplexe, nichtlineare Zusammenhänge modellieren und sind somit für ein breites Anwendungsfeld geeignet. Der Einsatz im polizeilichen Kontext setzt dabei aber besonders strenge Anforderungen an die Zuverlässigkeit von KI-Methoden sowie die Einhaltung von ethischen und rechtlichen Standards. Neue KI-Methoden, basierend auf Deep-Learning-Verfahrenen, haben die Genauigkeit und Robustheit von automatisierter Datenauswertung erheblich verbessert. Allerdings bestehen insbesondere bei diesen komplexen Modellen Fragen bezüglich der Erklärbarkeit, Transparenz sowie Fairness bzw. des Debaisings. Diese Attribute sind jedoch notwendig für vertrauenswürdige Künstliche Intelligenz.

Für eine polizeiliche und insbesondere gerichtsverwertbare Anwendung solcher Modelle sind die Attribute dieser Methoden unerlässlich. Während in der aktuellen Forschung hauptsächlich Erklärbarkeitsmethoden für Entwickler und KI-Experten entwickelt werden, fallen solche für Endanwender selbst oft in den Hintergrund. Jedoch müssen diese Methoden auf die jeweiligen Benutzergruppen angepasst werden, um effektiv zu sein bzw. die Transparenz überhaupt fördern zu können. In VIKING werden in vier Anwendungsfeldern spezifische KI-Methoden entworfen, wobei parallel an Methoden zur Erklärbarkeit, Transparenz, und des Debiasings für die speziellen Anwendungsfälle und Modelle geforscht wird. Diese spezifischen Erklärbarkeits- und Transparenzmethoden werden schließlich in einem übergreifenden Framework zusammengefasst. Hierin fließt die Übertragbarkeit der Konzepte sowie die ethische und rechtliche Begleitforschung ein, um Prüfkataloge, Standardisierungsverfahren sowie Best-Practices für KI-Methoden im polizeilichen Anwendungsfeld erarbeiten zu können.

Die Beiträge dieses Teilprojekts zum Verbundprojekt VIKING sind:

- Die technische und wissenschaftliche Leitung des VIKING-Projektes, insbesondere für die Entwicklung von Methoden der erklärbaren KI (XAI)
- Leitung der Bedarfsanalyse und Szenarien-bezogenen Spezifikation der einzelnen Teilprojekte (AP1)

- Forschung an spezifischen Erklärbarkeits-, Transparenz-, und Debiasingmethoden für spezifische Anwendungsfälle (AP4 - AP7)
- Erstellung eines Moduls für die Objektdetektion in Zusammenarbeit mit dem LKA NRW (AP7)
- Prüfung der Übertragbarkeit der Erklärungs-, Transparenz-, und Debiasingmethoden und deren technische Integration auf ein einheitliches Framework sowie die Standardisierung, sowie Erstellung von Prüfkatalogen sowie Best Practices (AP8)

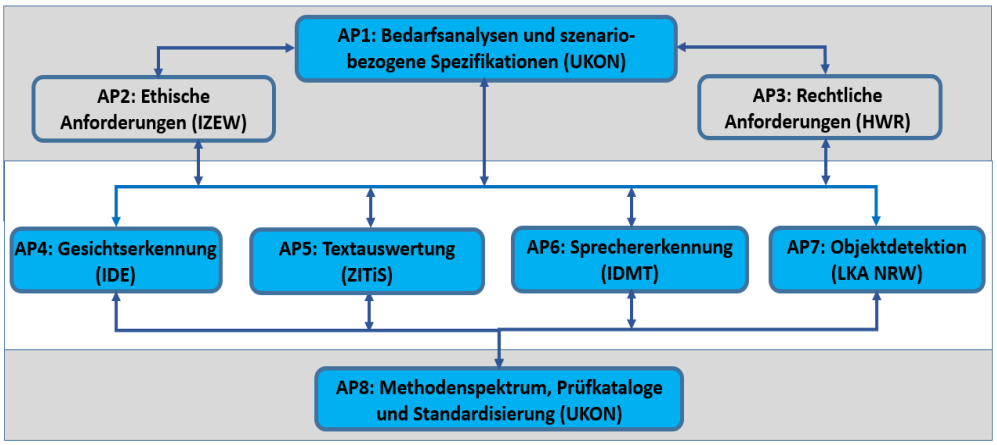

Abbildung 1: Übersicht der Arbeitspakete. Die Uni Konstanz (UKON) war an den blau gekennzeichneten APs beteiligt.

Im Folgenden werden die erreichten Ergebnisse vertieft dargestellt, sowohl innerhalb der Arbeitspakete als auch im Rahmen geförderter Forschung.

2.1.2. Bedarfsanalyse und Szenario-bezogene Spezifikationen

Im Rahmen des ersten Arbeitspaket AP1 wurde in Zusammenarbeit mit allen beteiligten Partnern eine Bedarfsanalyse für die technische Umsetzung, insbesondere bezüglich des Einsatzes von Erklärbarkeitsmethoden, entwickelt.

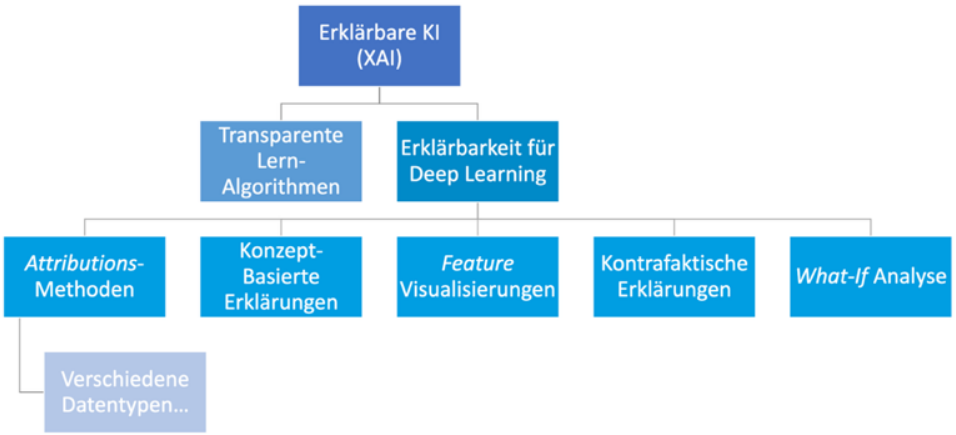

Abbildung 2: Ontologie von Erklärbarkeitsmethoden, wie in VIKING-Workshops als Material verwendet.

Als erster Schritt wurde dazu im Rahmen des ersten Statustreffens ein Workshop durch die Universität Konstanz durchgeführt. In dem Workshop erstellte die UKON für alle technischen Partner einen Überblick über verfügbare Erklärbarkeitsmethoden. Abbildung 2 zeigt eine Übersicht über verschiedene Ansätze von Erklärbarkeitsmethoden. Dazu wurden Ansätze zu transparenten Algorithmen, Attributionsmethoden für neuronale Netzwerke (z.B. Heatmaps bei Objektdetektion

in Bildern oder für Zeitreihen (siehe Abbildung 3)), konzeptbasierte Erklärungen, *Feature Visualization*, und kontrafaktische Erklärbarkeit vorgestellt. Alle vorgestellten Methoden sind generalisierbar auf verschiedenen Typen von Eingabedaten und Vorhersagen.

Abbildung 3: Konzept von Attributionen: Eine Attributionsmethode weißt Eingabewerten eine Wichtigkeit für eine Vorhersage zu, hier innerhalb einer Zeitreihe (Schlegel et al., Introducing…, 2023)

Als motivierendes Beispiel für einen visuellen, interaktiven Prototyp wurde das explAIner-Framework vorgestellt, das von der Universität Konstanz entwickelt wurde. Das Framework zeigt, wie lokale Erklärbarkeitsmethoden, wie Attributions-Heatmaps, in eine komplexe Anwendung eingebettet werden, um die zielgerichtete Untersuchung von bestimmten Eingaben zu unterstützen. In weiterführenden Diskussionen, aufgeteilt auf die technischen Arbeitspakete, wurden mögliche Lösungen für die einzelnen Anwendungen diskutiert.

Ergebnis der Bedarfsanalyse war ein Spezifikationsdokument für das VIKING-Projekt. In diesem Dokument werden alle relevanten Anforderungen an die Technikgestaltung der verschiedenen Teildemonstratoren zusammengefasst. Diese Anforderungen stammen aus dem involvierten Anwenderkreis, wobei ethische und rechtliche Aspekte (z. B. Gerichtsverwertbarkeit) berücksichtigt werden. Weiterhin wurden technische Spezifikationen der Teildemonstratoren, wie etwa gewählte Methoden und Evaluierungsansätze, spezifiziert.

Details zur Bedarfsanalyse und Umsetzung findet sich in den Ergebnisberichten der technischen Teilprojekte (AP4-AP7).

2.1.3. Unterstützung der Technischen Teilvorhaben: Gesichtserkennung, Textauswertung, Sprechererkennung und Objektdetektion

Als wissenschaftlicher und technischer Leiter innerhalb des Konsortiums stand die Universität Konstanz als Partner während der technischen Entwicklung der Demonstratoren zur Verfügung. Insbesondere unterstützte die Konstanz die Demonstratorentwicklung für die Objektdetektion in AP7. Hier wirkte die UKON von Beginn an beim Entwurf, der Entwicklung sowie der Evaluation mit. Die detaillierten Ergebnisse werden in den zugehörigen Teilvorhabens-Berichten beschrieben.

Für den Demonstrator zur Objektdetektion (AP7) wirkte die Universität Konstanz in monatlichen Meetings mit, in denen der aktuelle Fortschritt und Details zur Implementation besprochen wurden. Insgesamt war die Universität Konstanz zu drei Veranstaltungen beim durchführenden Partner, dem LKA NRW, zu Besuch. Einmal für ein Planungsmeeting und in zwei Evaluationsmeetings. Die Evaluation fand im Rahmen des Arbeitspakets AP 8.4 statt. In beiden Evaluationsmeetings wurde der Demonstrator mit Ermittlern getestet. Gesammeltes Feedback floss dann wiederum in die weitere Entwicklung ein. Die Uni Konstanz unterstützte hier bei der Durchführung und Protokollierung der Evaluation.

Ein wichtiges Ergebnis der Evaluation war, dass die Nutzer den Einsatz des interaktiven visuellen Demonstrators zur Einbettung von Erklärungen und Schulung von Nutzern als sehr positiv bewerteten. Einzelne Ergebnisse von Attributionsmethoden ohne weiteren Kontext wurden häufig als

wenig aussagekräftig bewertet. Bei den Ergebnissen von Erklärbarkeitsmethoden bleiben hier Herausforderungen bestehen, die durch die zugrundeliegende Intransparenz von neuronalen Netzwerken verursacht werden. Deshalb ist eine Kontextualisierung, und Kombination verschiedener Methoden zielführender. Dieser Grundsatz wurde bei der Entwicklung der interaktiven Demonstratoren berücksichtigt.

Wie im vorherigen Abschnitt beschrieben, stand die Universität Konstanz darüber hinaus bei allen technischen Teilvorhaben, besonders zu Beginn, in der Design- und Spezifikationsphase, beratend zur Seite. Des Weiteren war die Uni Konstanz regelmäßig bei Update-Treffen und bei Konsortialtreffen mit den durchführenden Partnern in Kontakt (z. B. für AP4 und AP6), um Strategien und Lösungsansätze zu besprechen und auszuarbeiten.

2.1.4. Methodenspektrum, Verfahrensempfehlungen und Standardisierung (DIN-SPEC *91517*)

Im Arbeitspaket AP 8 zur Standardisierung und Entwicklung eines übergreifenden Frameworks konnten mehrere Erfolge erzielt werden:

Zusammen mit dem IZEW, verantwortlich für AP2, wurden Kataloge für die Dokumentation von technischen, ethischen und rechtlichen Anforderungen entwickelt. Diese Kataloge wurden in mehreren Runden detailliert besprochen, analysiert und angepasst. Um eine Übertragung der Anforderungsanalyse zu gewährleisten und einen einfachen Zugang zu ermöglichen, wurde ein Demonstrator erstellt (Abbildung 4). Der Demonstrator ermöglicht das einfache Erstellen, Bearbeiten, Ausfüllen und Teilen von Anforderungskatalogen. Vorlagen können dabei einfach als maschinenlesbare JSON-Dokumente ausgetauscht werden.

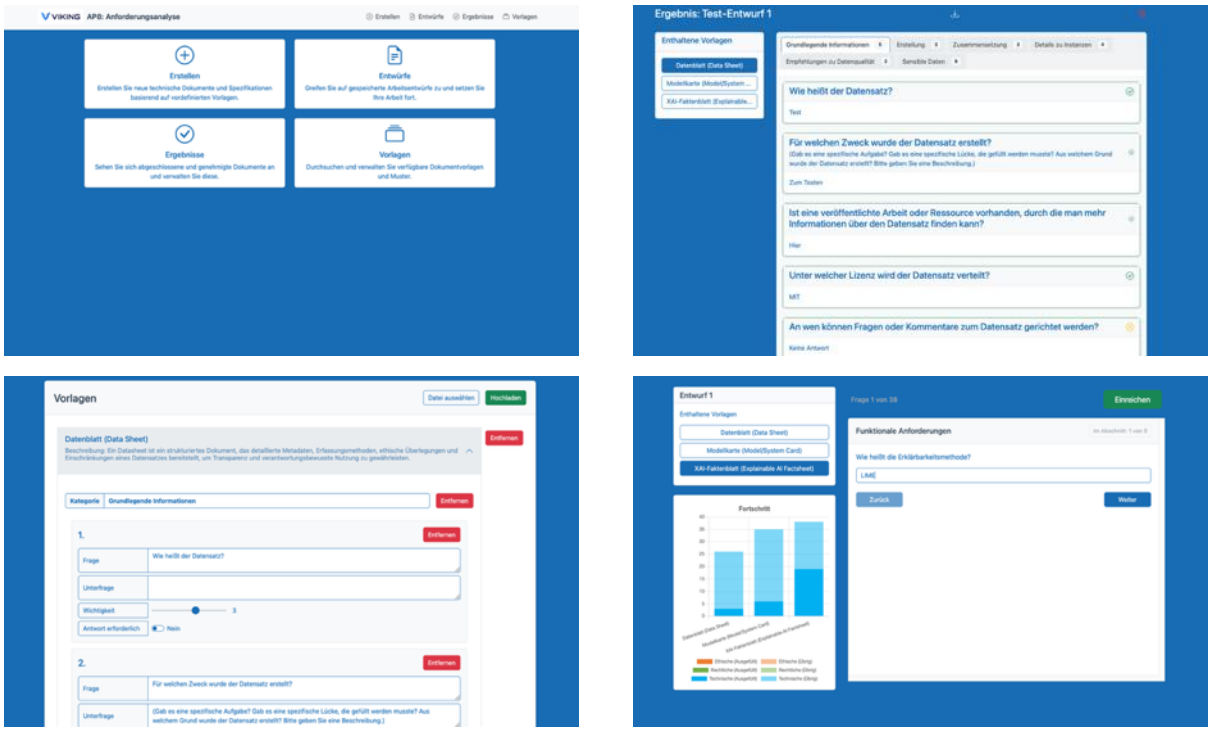

Abbildung 4: Demonstrator zur Interaktiven Erstellung von Fragebögen zur Übertragbarkeit/Dokumentation

Ein weiteres Ziel war die Entwicklung eines technischen Frameworks zur Generierung von Erklärungen. Dieses Framework soll die Komplexität bei der Erstellung von Erklärungen bündeln und

für Anwender eine einfache und effiziente Schnittstelle bieten. Abbildung 5 zeigt die Architektur des entwickelten Demonstrator-Systems. Das System ist eine Erweiterung des explAIner-Systems und basiert auf der Captum-Softwarebibliothek, um automatische Attributionen oder konzeptbasierte Erklärungen zu generieren. Die Bibliothek unterstützt dafür generische PyTorch-Modelle (z.B. ONNX oder ein PyTorch-eigenes Format). Zur automatischen Erstellung von Erklärungen werden die zu verwendende Modell-Art und der relevante Datensatz an den Berechnungs-Server im Backend gesendet. Die erzeugten Erklärungen werden dann an den Nutzer in der Web-Applikation zurückgegeben. Entscheidend, um die rechtlichen Anforderungen zu erfüllen, ist, dass weder Modelle noch Datensätze dauerhaft auf dem Server gespeichert sind. Modell und Daten bleiben dauerhaft nur bei dem eigentlichen Endnutzer. Trotzdem kann eine zentrale Infrastruktur zur Erzeugung genutzt werden.

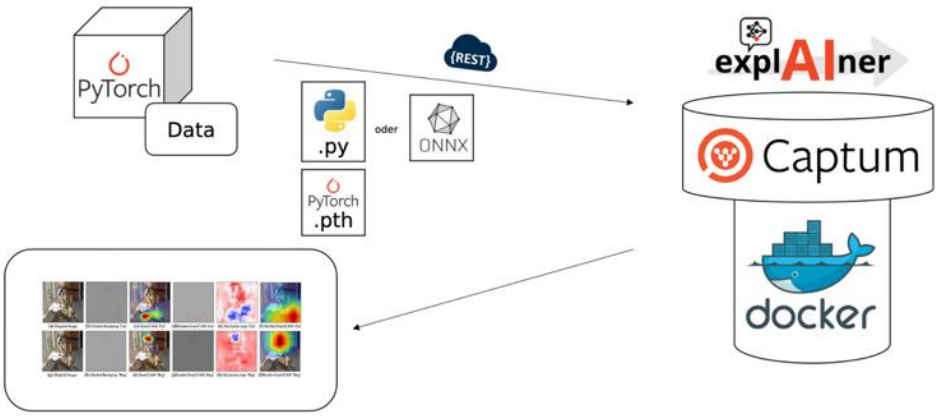

Abbildung 5: Infrastruktur des technischen Frameworks für Generierung von Erklärungen (AP8.2)

Um den langfristigen Erfolg des Projektes zu sichern und die Projekterkenntnisse aus VIKING einer breiten Fachöffentlichkeit zur Verfügung zu stellen, wurde neben den Paper-Veröffentlichungen (siehe Publikationsliste) eine Normierung im Rahmen einer DIN SPEC verfolgt. Das Ziel dieses Standardisierungsprozesses war die Überführung der im VIKING-Projekt entwickelten konzeptionellen, rechtlichen und technischen Erkenntnisse sowie deren Zusammenspiel im polizeilichen Anwendungskontext in eine strukturierte, anwendbare Spezifikation und die langfristige Sicherung des erworbenen Wissens. Im Mittelpunkt stand die Definition belastbarer und praxistauglicher Anforderungen an vertrauenswürdige KI-Methoden in sicherheitsbehördlichen Kontexten. Die Ausarbeitung fand in enger Kooperation zwischen den Projektpartnern sowie Vertretern und Projektpartnern der DIN statt. Es wurden acht DIN SPEC-Konsortialmeetings abgehalten, die der strukturierten Abstimmung und Fortschrittskontrolle dienten. Die Arbeitspakete umfassten:

Systematische Normenrecherche innerhalb des DIN-Regelwerks, Beiträge zu Workshops und Fachsitzungen zur Konzeptentwicklung, iterative Ausgestaltung allgemeiner und domänenspezifischer Anforderungen und Analyse realer Use-Cases und praktischer Anwendungsszenarien. Besondere Schwerpunkte lagen auf den KI-Domänen Textauswertung, Objekterkennung, Gesichtserkennung und Sprechererkennung sowie dem Einsatz moderner KI-Technologien in polizeilichen Use-Cases bei gleichzeitiger Schaffung und Berücksichtigung von Transparenz, Verantwortlichkeit und Rechtssicherheit.

Die Universität Konstanz (UKON) initiierte den Standardisierungsprozess im Juni 2024. Herr Dr. Maximilian T. Fischer (Universität Konstanz) wurde aufgrund seiner ausgewiesenen Expertise einstimmig als Konsortialleiter für die DIN-SPEC-Normierung gewählt und Frau Dr. Lou Brandner (IZEW, Uni Tübingen) als Vertretung und Co-Leiterin bestimmt. Beide haben anschließend konzeptionell die Ausarbeitung und Konsortialführung übernommen. Die DIN SPEC wurde von Juni 2024 bis Dezember 2024 erarbeitet und im Abschluss als Entwurfsfassung auf dem VIKING-Konsortialmeeting im Dezember 2024 in Oldenburg vorgestellt. Im Januar 2025 erfolgten noch redaktionelle Überarbeitungen und Feinkorrekturen, und Anfang Februar 2025 wurde von Konsortialseite die endgültige Version finalisiert und von Dr. Fischer und Dr. Brandner freigegeben. Von Seiten der DIN erfolgt dann bis März 2025 eine interne abschließende Prüfung sowie eine Vorab-Online-Veröffentlichung. Die DIN-SPEC-Spezifikation (siehe auch Abbildung 6) wurde schlussendlich final im Mai 2025 unter Berücksichtigung entsprechender Vorankündigungs-Fristen veröffentlicht:

DIN SPEC 91517:2025-05 Anforderungen an vertrauenswürdige KI-Methoden in polizeilichen Anwendungen, DIN Media GmbH, 2025. doi: https://dx.doi.org/10.31030/3612025

Mit Abschluss des Standardisierungsprozesses liegt ein anwendungsorientiertes, methodisches und normatives Fundament für den verantwortungsvollen Einsatz von KI in polizeilichen Kontexten vor. Die enge Verzahnung von Forschungs-, Praxis- und Normungsakteuren hat einen maßgeblichen Beitrag zur Qualität und Anwendbarkeit des Dokuments geleistet.


Mai 2025

DIN SPEC 91517

DIN

ICS 35.240.01

Anforderungen an vertrauenswürdige KI-Methoden in polizeilichen Anwendungen

Requirements for trustworthy AI methods in police applications

Exigences relatives aux méthodes d'IA fiables dans les applications policières

Gesamtumfang 68 Seiten

Dieses Dokument wurde durch die im Vorwort genannten Verfasser erarbeitet und verabschiedet.

Alleinverkauf durch DIN Media GmbH, 10772 Berlin

www.din.de
www.dinmedia.de

3612025

Abbildung 6: Standardisierungsdokument DIN SPEC 91517:2025-05, 68 Seiten (Fischer et al., DIN SPEC 91517: Anforderungen an vertrauenswürdige KI-Methoden in polizeilichen Anwendungen. DIN Media GmbH, 2025. doi: 10.31030/3612025)

2.1.5. Zusammenarbeit und wissenschaftliche Beiträge

Die Zusammenarbeit mit den Konsortialpartnern und assoziierten Partnern spielte eine entscheidende Rolle für den Erfolg des Projekts. Diese Kooperationen trugen wesentlich zur Entwicklung und Evaluation der technischen Lösungen bei. Darüber hinaus war die Präsentation der Ergebnisse bei verschiedenen Konsortialtreffen ein wichtiger Aspekt, der eine rege Beteiligung von Vertretern der Sicherheitsbehörden nach sich zog. Wissenschaftlich betrachtet führte das Projekt zu zahlreichen Publikationen in internationalen Fachzeitschriften und Konferenzbänden, wodurch die Forschungsarbeit der Universität Konstanz in diesem Bereich unterstrichen wurde.

Durch das Projekt VIKING wurde an der Universität Konstanz Grundlagenforschung erfolgreich gefördert. Forscher der UKON haben während der Laufzeit des Projekts erfolgreich zehn teilfinanzierte Publikationen veröffentlicht (siehe Abschnitt 2.6). Die Publikationen lassen sich dabei in drei Gruppen unterteilen:

- Erklärbarkeitsmethoden für Zeitreihen, inklusive der Evaluierung von Attributionsmethoden, neuen Darstellungen via Pixel-basierter Visualisierungen und interaktiver Erstellung von konterfaktischen Erklärungen
- Visuelle Untersuchung von Biases und semantischen Veränderungen in Sprachmodellen und Entwicklung neuartiger Projektionsmethoden
- Entwicklung von graphischer Benutzeroberfläche für Suche in großen Video-Datensätzen („known-item search in video retrieval“)

Zum ersten Themengebiet wurden insgesamt sechs Publikationen veröffentlicht. Neben einem umfassenden Survey zum Thema „Erklärbare KI für Zeitreihen“ (*Theissler et al., 2022*) wurden Beiträge und eine Reihe von Implementierungen und empirischen Untersuchungen vorgestellt. Dabei decken diese Arbeiten mehrere Gruppen von Erklärbarkeitsmethoden ab: Attributionsmethoden, kontrafaktische Erklärungen und *Feature Visualization*.

In einer weiteren Arbeit, wurde eine interaktive pixel-basierte Visualisierung vorgestellt (*Schlegel et al., Interactive dense pixel visualizatons..., 2023)*. Diese Form der Darstellung, siehe Abbildung 7, ermöglicht die gleichzeitige visuelle Analyse einer großen Anzahl von Zeitreihen. Die Analyse wird dabei durch ein automatisiertes Clustering und Ordnen der Zeitreihen und Attributionsdaten unterstützt. In der Publikation wird die Anwendung auf mehrere Datensätze demonstriert und werden einige Ergebnisse diskutiert.

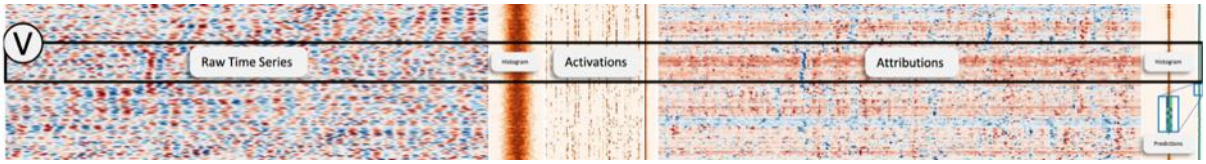

Abbildung 7: Pixel-basierte Darstellung von Zeitreihen und Attributionen (Schlegel et al., Interactive dense pixel visualizatons for time series and model attribution explanations, 2023)

In einer dritten Arbeit wurden Pertubationen als Methode zur Evaluation von Attributionen untersucht (*Schlegel et al., A Deep Dive into Perturbations as Evaluation Technique..., 2023*). Da für Attributionen keine Ground-Truth-Daten verfügbar sind, ist die Evaluation solcher Methoden nicht trivial. Die Arbeit nutzt Pertubationen, um gezielt bestimmte Vorhersagen zu generieren,

und somit bestimmte Bereiche von Zeitreihen als relevant zu bestimmen. Attributionsmethoden können dann evaluiert werden, indem die Fähigkeit, diese Bereiche zu erkennen, gemessen wird.

Diese Arbeit zur Evaluation von Attributionsmethoden für Zeitreihen wurde in einer anschließenden Publikation noch erweitert (*Schlegel et al., Introducing the Attribution Stability Indicator: A Measure for Time Series XAI Attributions, 2023*). In dieser Arbeit wurde ein Stabilitäts-Indikator für Attributionen eingeführt, der sich zur generellen Untersuchung von Attributionsmethoden eignet, und sowohl Robustheit als auch Vertrauenswürdigkeit von Attribution quantifizieren kann.

Eine weitere Forschungsarbeit widmet sich der interaktiven Erstellung von Counterfactuals, also Eingabedaten die innerhalb eines Vorhersagemodels ein anderes Ergebnis erzeugt (*Schlegel et al., Interactive Counterfactual Generation for Univariate Time Series, 2024)*. Wie in Abbildung 8 zu sehen, kann ein Expertennutzer dabei interaktiv die Vorhersagen für einen Zeitreihen-Datensatz explorieren. Für ausgewählte Zeitreihen kann der Nutzer alternative kontrafaktische Datenpunkte erzeugen. Durch Auswählen einer alternativen Klasse für die Vorhersage, wird eine Zeitreihe erzeugt, die vom Klassifikator entsprechend klassifiziert wird, sich dabei aber möglichst wenig von der originalen Zeitreihe unterscheidet. Dies ermöglicht insbesondere Expertennutzern, die Robustheit des Klassifikationsmodells zu untersuchen, und relevante Teile der Eingabedaten zu identifizieren.

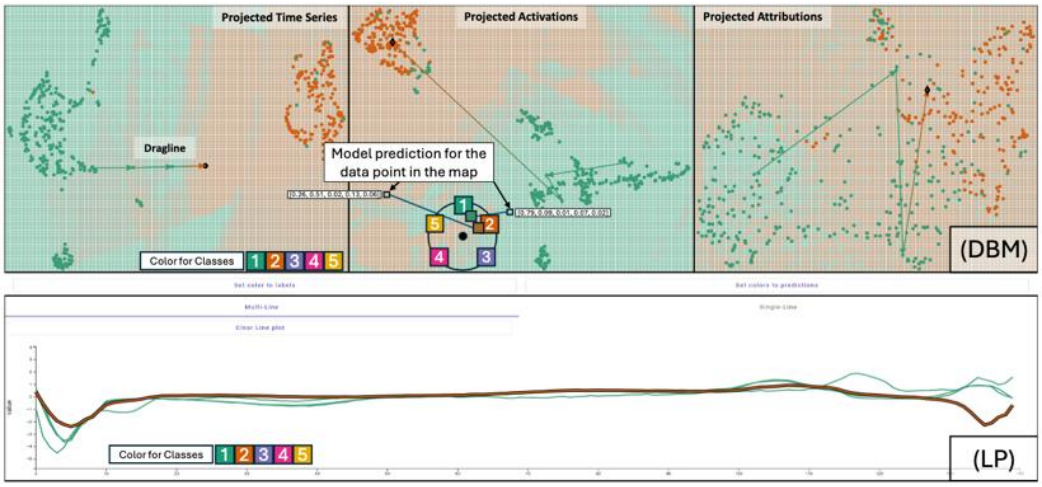

Abbildung 8: Screenshot des Interfaces zur interaktiven Erstellung von kontrafaktischen Zeitreihen-Daten (Schlegel et al., Interactive Counterfactual Generation for Univariate Time Series, 2024)

Die abschließende Arbeit (*Schlegel et al., Finding the DeepDream for Time Series: Activation Maximization for Univariate Time Series, 2024*) im Themengebiet Zeitreihen beschäftigt sich mit Feature Visualization. Hierfür wird auf der sogenannten DeepDream-Methode für neuronale Netzwerke aufgebaut: Dabei werden künstliche Daten erzeugt, die zu maximalen Aktivierungen des Netzwerks führen. Dies kann genutzt werden, um die Features zu visualisieren, die bestimmte Gewichte in einem Netzwerk aktivieren. Das Paper stellt eine Methode vor, um ähnlich wie bei Bildern künstliche Zeitreihen zu erzeugen, die beispielsweise die Vorhersagen für eine bestimmte Klasse maximieren.

Drei weitere Arbeiten widmen sich der Analyse von Biases und semantischen Veränderungen in Sprachmodellen sowie dafür geeigneten Visualisierungsmethoden.

Eine einfache Arbeit führt eine neue 2D-Projektionsmethode ein: cPro ist ein gradientenbasierter Optimierungsalgorithmus für runde Projektionen (*Buchmüller et al., cPro: Circular Projections*

Using Gradient Descent, 2024). Die Methode erlaubt es, hochdimensionale Daten auf einen Kreis zu projizieren und dabei Artefakte bei der Distanzberechnung mit einzubeziehen, um Relationen einfacher darzustellen.

Anwendung finden die kreisförmigen Projektionen in einer Anwendung zur Exploration von Argumentations-Präferenzen, d. h. in der Analyse von verschiedenen Präferenzgruppen in Text-Corpora. Dies ist eine Methode, die in der Linguistik-Forschung genutzt wird (*Buchmüller et al., Exploration of Preference Models using Visual Analytics, 2024*). Abbildung 9 zeigt zwei kreisförmige Projektionen zur Analyse von verschiedenen Nutzern und deren Präferenzen.

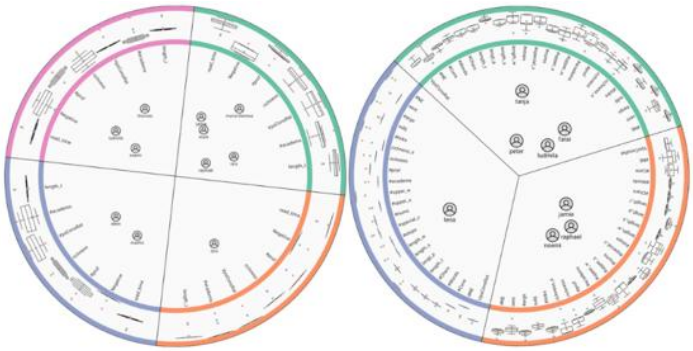

Abbildung 9: Visualisierungen zur Untersuchung von Nutzer-Präferenzen für Argumentationen, Daten (Buchmüller et al., Exploration of Preference Models using Visual Analytics, 2024)

Ein kleines Position Paper diskutiert Ansätze, um semantische Veränderungen beim Umgang mit großen Sprachmodellen zu untersuchen (*Buchmüller et al., Seeing the Shift: Keeping an Eye on Semantic Changes in Times of LLMs, 2024*). Insbesondere werden Analyse- und Visualisierungsmethoden im Ansatz verglichen, die sprachliche Veränderungen über einen Zeitverlauf zeigen können. Diese Forschung ist einzuordnen in eine Reihe von Arbeiten, die potentielle Biases in Sprachmodellen untersuchen, in diesem Fall durch die Auswirkungen auf Nutzer.

In einer weiteren Studie geht es um die Entwicklung effektiver Oberflächen für die Suche in großen Video-Korpora. Speziell geht es darum, Nutzern die Suche nach bestimmten Objekten oder Frames in größeren, relativ homogenen Video-Datensätzen zu ermöglichen. Dies könnte z. B. für das Auswerten von Beweisdaten oder Videoüberwachungsvideos relevant sein. Um eine möglichst schnelle und treffsichere Suche zu ermöglichen, sollten die grafischen Oberflächen möglichst gut gestaltet sein. In der Studie, wird mittels einer Eye-Tracking-Analyse die beste Layout-Anordnung von Video-Standbildern untersucht, siehe Abbildung 10. Insbesondere werden dabei Faktoren wie die Anzahl der gleichzeitig angezeigten Spalten evaluiert.

Abbildung 10: Interface der Eye-Tracking Studie & Heatmap der Sichtpunkte von Studienteilnehmern mit einem öffentlichen Datensatz (Fish Reidentification)
(Joos et al., Known-Item Search in Video: An Eye Tracking-Based Study, 2024)

Zusammenfassend, wurde durch das Projekt VIKING eine Reihe von Themen in der Grundlagenforschung, im Bereich Erklärbare KI, Datenvisualisierung, und Mensch-Computer-Interaktion sehr erfolgreich gefördert.

2.1.6. Zusammenfassung und Ausblick

Das VIKING-Projekt leistet einen wichtigen Beitrag zum verantwortungsvollen, ethisch und rechtlich geprüften Einsatz von KI für Polizei und Ermittlungsbehörden. Die Universität Konstanz hat federführend zu neuen technischen Lösungen und einer umfassenden Anforderungsanalyse beigetragen. Insbesondere durch die Veröffentlichung des DIN SPEC-Standarddokuments wird sichergestellt, dass die Erkenntnisse des VIKING-Projekts über die Laufzeit hinaus zur Verfügung stehen. Auch die entwickelte Anwendung zum Erstellen von Fragebögen und Checklisten kann in Zukunft als eine Referenz dienen. Weiterhin können die entwickelten Demonstratoren, insbesondere durch die erfolgte Evaluation im Polizei-Kontext, als Referenz für die Erstellung und Auswahl von Software in der Zukunft dienen. Bei der Entwicklung und Evaluation hat sich gezeigt, dass isolierte Ergebnisse von Erklärbarkeitsmethoden, wie z.B. Heatmaps für Eingabe-Attributionen, keine zufriedenstellende Lösung darstellen. Das Projekt VIKING als Ganzen, und die Forschung der Universität Konstanz, zielt stattdessen auf eine Einbettung von Modellen und Erklärungen in interaktive, visuelle Anwendungen. Diese ermöglichen eine Kontextualisierung von Vorhersagen und Erklärungen durch Vergleich von verschiedenen Eingabedaten und Erklärbarkeitsmethoden. Gleichzeitig können Nutzer durch den Einsatz dieser Systeme effektiver geschult werden. VIKING leistet deshalb einen wichtigen Beitrag für den zukünftigen Einsatz vertrauenswürdiger KI. Die durch VIKING finanzierte Grundlagenforschung an der Universität Konstanz trägt darüber hinaus zur weiteren nationalen und europäischen Souveränität in der KI-Forschung bei.

2.1.7. Dissemination und Austausch

Im Rahmen des Projektes kam es zu einem regen Austausch innerhalb des Konsortiums sowie mit der wissenschaftlichen Gemeinschaft im Allgemeinen:

Konsortialtreffen

Die Universität Konstanz hat zwei Konsortialtreffen organisiert und ausgerichtet, darunter das initiale Treffen. Bei allen weiteren Treffen war das Teilprojekt mit mindestens einem Vertreter, in der Regel zwei bis drei, bei allen Konsortialtreffen in München, Berlin, und Oldenburg (jeweils im Wechsel) vertreten.

Workshops zu Arbeitspaketen

Die Universität Konstanz war zusätzlich bei einer Reihe von Workshops zum Entwurf und der Evaluierung von Demonstratoren vor Ort, etwa beim LKA NRW in Düsseldorf.

Dissemination der Projektergebnisse

Das Projekt und Zwischenergebnisse wurden bei mehreren Gelegenheiten einer Fachöffentlichkeit vorgestellt:

Internationale und Nationale Forschungskonferenzen und Besuche:

- *European Police Congress 2023*, Berlin, Deutschland, Vorstellung des VIKING-Projekts und Methoden zur erklärbaren KI

- *World Conference on Explainable Artificial Intelligence (XAI) 2023*, Lissabon, Portugal, Vorstellung von zwei Publikationen
- *European Conference on Machine Learning and Principles and Practice of Knowledge Discovery 2024*, Vilnius, Litauen, Vorstellung von zwei Publikationen im Rahmen des Workshop-Programms
- *MLViz Workshop auf der 2023 Eurographics Conference on Visualization (EuroVis 2023),* Leipzig, Deutschland, Vorstellung eines Workshop-Papers im Rahmen des Workshop-Programms
- *EuroVA Workshop auf der 2024 Eurographics Conference on Visualization (EuroVis 2024)*, Odense, Dänemark, Vorstellung einer Publikation im Rahmen des Workshop-Programms

Veröffentlichungen

siehe Abschnitt 2.6

2.2. Wichtigste Positionen des zahlenmäßigen Nachweises

Die summarischen Kostenpositionen sind im Folgenden aufgelistet.

Zeitraum 01-12/2022:

- Personalkosten (0812):	0,00 €
- Hiwikosten (0822):	4.953,23 €
- Sonst. Allg. Verwalt. (0843)	103,29 €
- Dienstreisen (0846):	2.061,33 €

Zeitraum 01-12/2023:

- Personalkosten (0812):	59.653,20 €
- Hiwikosten (0822):	5.753,85 €
- Sonst. Allg. Verwalt. (0843)	0,00 €
- Dienstreisen (0846):	5.977,36 €

Zeitraum 01-12/2024:

- Personalkosten (0812):	235.729,22 €
- Hiwikosten (0822):	13.951,02 €
- Sonst. Allg. Verwalt. (0846)	493,54 €
- Dienstreisen (0846):	8.452,37 €

Zeitraum 01-03/2025:

- Personalkosten (0812):	44.058,48 €
- Hiwikosten (0822):	11.309,85 €
- Sonst. Allg. Verwalt. (0846)	0,00 €
- Dienstreisen (0846):	1.088,16 €

Für den detaillierten Nachweis der einzelnen Kostenpositionen verweisen wir auf den zahlenmäßigen Nachweis.

2.3. Notwendigkeit und Angemessenheit der geleisteten Arbeit

Die Durchführung der Arbeiten folgte überwiegend der Planung des Projektantrags. Alle wesentlichen Ziele wurden sowohl im Gesamtvorhaben, als auch im Teilvorhaben erreicht. Um die Arbeiten des Gesamtvorhabens optimal zu unterstützen, und auf einen internen Personalwechsel zu reagieren, fand eine kostenneutrale Verlängerung um drei Monate statt.

2.4. Nutzen und Verwertung im Sinne des Verwertungsplans

Durch die erfolgreiche Entwicklung und Veröffentlichung der Anforderungen und Standardisierungsdokumente ist eine Grundlage für weitere wirtschaftliche Verwendung gelegt. Insbesondere können die Projektergebnisse zukünftige Entwicklungen für polizeiliche Ermittlungen ermöglichen.

Die entwickelten und mit relevanten Benutzergruppen evaluierten Demonstratoren dienen ebenfalls als Basis für zukünftige, potenziell auch kommerzielle Anwendungen, z. B. durch Industriepartner im Konsortium.

Aus wissenschaftlicher Sicht ist das Projekt sehr positiv zu bewerten. Die Universität Konstanz konnte eine Reihe von Forschungsarbeiten abschließen und erfolgreich publizieren. Die Arbeiten waren wesentlich dabei, den Status der Universität Konstanz als eine der führenden Arbeitsgruppen für erklärbare KI, insbesondere im Sicherheits-Kontext, auszubauen. Außerdem wurden Ergebnisse des Projekts auf mehreren internationalen Konferenzen präsentiert. Durch das Projekt wurden zwei Bachelor- und Masterarbeiten unterstützt sowie eine Dissertation erfolgreich abgeschlossen.

2.5. Bekannt gewordener Fortschritt auf dem Gebiet des Vorhabens

Während der Projektlaufzeit hat es weitere Fortschritte in Deep-Learning-Modellen gegeben, z.B. im Bereich der Objekterkennung oder Textanalyse. Auch das Feld der erklärbaren KI für Deep Learning hat sich erweitert. Trotzdem bleibt das Problem der Erklärbarkeit auch bei neueren Modellen bestehen, und es gibt derzeit keine grundsätzlich neuen Lösungen. Die entwickelten Methoden sind weiterhin weitgehend kompatibel mit aktuellen Lösungen. Es sind aber keine direkten vergleichbaren Vorhaben oder Veröffentlichungen im polizeilichen Kontext bekannt.

Als wesentliche Entwicklung im rechtlichen Bereich ist die Gesetzgebung auf EU-Ebene zu nennen, insbesondere der EU AI-Act. Dieser schreibt feste Richtlinien für den Einsatz von KI vor, insbesondere für Hochrisiko-Anwendungen. Diese Risiko-Anwendungen schließen Teile der im Projekt untersuchten Bereiche mit ein. Der Beschluss des AI-Acts erfolgte während der regulären Projektlaufzeit und konnte in den finalen Standardisierungsdokumenten berücksichtigt werden.

2.6. Erfolgte Veröffentlichungen

Im Rahmen des Projekts ergaben sich folgende internationale Veröffentlichungen:

1. A. Theissler, F. Spinnato, U. Schlegel and R. Guidotti (2022). "Explainable AI for Time Series Classification: A Review, Taxonomy and Research Directions" in IEEE Access, vol. 10, pp. 100700-100724, 2022. doi: 10.1109/ACCESS.2022.3207765
2. Schlegel, U., & Keim, D. A. (2023). Interactive dense pixel visualizations for time series and model attribution explanations. Workshop on Machine Learning Methods in Visualisation for Big Data (MLVis) @ EuroVIS. doi: 10.2312/mlvis.20231113
3. Schlegel, U., Keim, D.A. (2023). A Deep Dive into Perturbations as Evaluation Technique for Time Series XAI. In: Longo, L. (eds) Explainable Artificial Intelligence. xAI 2023. Communications in Computer and Information Science, vol 1903. Springer, Cham. doi: 10.1007/978-3-031-44070-0_9
4. Schlegel, U., Keim, D.A. (2025). Introducing the Attribution Stability Indicator: A Measure for Time Series XAI Attributions. In: Meo, R., Silvestri, F. (eds) Machine Learning and Principles and Practice of Knowledge Discovery in Databases. ECML PKDD 2023. Communications in Computer and Information Science, vol 2135. Springer, Cham. doi: 10.1007/978-3-031-74633-8_1
5. Schlegel U., Rauscher J., & Keim D. A. (2024). Interactive Counterfactual Generation for Univariate Time Series. Workshop on eXplainable Knowledge Discovery in Data Mining (XKDD). doi: 10.48550/arXiv.2408.10633
6. Schlegel U., Keim D. A.; & Sutter T. (2024) Finding the DeepDream for Time Series: Activation Maximization for Univariate Time Series. (2024). Workshop on Explainable AI for Time Series and Data Streams (TempXAI) @ ECML-PKDD. doi: 10.48550/arXiv.2408.10628
7. Buchmüller, R., Jäckl, B., Behrisch, M., Keim, D. A., Dennig, F. (2024). cPro: Circular Projections Using Gradient Descent. EuroVA: International Workshop on Visual Analytics. doi: 10.2312/eurova.20241111
8. Buchmüller, R., Zymla, M., Butt, M., Keim, D. A., Sevastjanova, R. (2024). „Exploration of Preference Models using Visual Analytics”. In MLVis: Machine Learning Methods in Visualisation for Big Data (2024). Eindhoven: Eurographics, 2024. doi: 10.2312/mlvis.20241127
9. Lucas Joos, Bastian Jäckl, Daniel A. Keim, Maximilian T. Fischer, Ladislav Peska, and Jakub Lokoč. 2024. Known-Item Search in Video: An Eye Tracking-Based Study. In Proceedings of the 2024 International Conference on Multimedia Retrieval (ICMR '24). Association for Computing Machinery, New York, NY, USA, 311–319. doi: 10.1145/3652583.3658119
10. Buchmüller R., Körte F., Keim D. A. (2024). Seeing the Shift: Keeping an Eye on Semantic Changes in Times of LLMs. 2024 IEEE Visualization in Data Science (VDS). doi: 10.1109/VDS63897.2024.00010

Eine DIN-Normierung:

11. Fischer, M. T., Schlegel, U., Keim, D. A., Altmann, S., Grote, C., Reuter, P., Coleman, G., Geierhos, M., Maoro, F., Kluin, M., Weinbruch, M., Aden, H., Kleemann, S., Tahraoui, M., Louban, A., Arndt, M., Schönrock, S., Brandner, L. T., Hirsbrunner, S. D., et al., Yilmaz, Yusuf (2025). DIN SPEC 91517:2025-05: Anforderungen an vertrauenswürdige KI-Methoden in polizeilichen Anwendungen. DIN Media GmbH. doi: 10.31030/3612025

Des Weiteren wurden Ergebnisse im Rahmen einer Bachelor-, einer Masterarbeit sowie einer Dissertation verarbeitet.