TEXTE
25/2026

Abschlussbericht

Nachhaltigkeitspotentialanalyse für die Zweckmäßigkeit und den Aufwand von Datenannotationen für Machine Learning (ML)-Modelle

Projekt LabelledGreenData4All

von:
Flaminia Catalli, Franziska Hochenegger, Thorsten Reitz
wetransform GmbH, Darmstadt
Eva Klien, Kevin Kocon, Michel Krämer
Fraunhofer-Institut für Graphische Datenverarbeitung IGD, Darmstadt

Herausgeber:
Umweltbundesamt

Für Mensch & Umwelt

Umwelt Bundesamt

TEXTE 25/2026

REFOPLAN des Bundesministeriums Umwelt,
Naturschutz, nukleare Sicherheit und Verbraucherschutz

Forschungskennzahl 3723 11 602 0
FB001826

Abschlussbericht

Nachhaltigkeitspotentialanalyse für die Zweckmäßigkeit und den Aufwand von Datenannotationen für Machine Learning (ML)-Modelle

Projekt LabelledGreenData4All

von

Flaminia Catalli, Franziska Hochenegger, Thorsten Reitz
wetransform GmbH, Darmstadt

Eva Klien, Kevin Kocon, Michel Krämer
Fraunhofer-Institut für Graphische Datenverarbeitung IGD,
Darmstadt

Im Auftrag des Umweltbundesamtes

Impressum

Herausgeber
Umweltbundesamt
Wörlitzer Platz 1
06844 Dessau-Roßlau
Tel: +49 340-2103-0
Fax: +49 340-2103-2285
buergerservice@uba.de
Internet: www.umweltbundesamt.de

Durchführung der Studie:
Wetransform GmbH
Fraunhofer Str. 5
64283 Darmstadt

Abschlussdatum:
Februar 2025

Redaktion:
Referat Z 2.3 Digitale Transformation und Beratungsstelle Green IT
Cathleen Mitzschke

DOI:
https://doi.org/10.60810/openumwelt-7930

ISSN 1862-4804

Dessau-Roßlau, März 2026

Die Verantwortung für den Inhalt dieser Veröffentlichung liegt bei den Autorinnen*Autoren.

Kurzbeschreibung: Nachhaltigkeitspotentialanalyse für die Zweckmäßigkeit und den Aufwand von Datenannotationen für Machine Learning (ML)-Modelle

Die digitale Transformation führt dazu, dass Daten zunehmend als wertvolle Ressource betrachtet werden, insbesondere im Umweltbereich. Das Projekt **LabelledGreenData4All** untersucht die strategische Bedeutung annotierter Umweltdaten für den Einsatz von Machine Learning (ML) und im weiteren Sinne Künstlicher Intelligenz (KI) zur Bewältigung gesellschafts- und umweltpolitischer Herausforderungen. Ziel war es, Anwendungsbereiche mit hohem Potenzial für ML-Modelle zu identifizieren und ein Vorgehensmodell zu entwickeln, welches dazu dienen soll, Datenannotationen auf Basis definierter Entscheidungspfade und unter Berücksichtigung verschiedener Ausgangssituationen hinsichtlich der Verfügbarkeit annotierter Trainingsdaten vorzunehmen. Die Annotationen selbst sollten dabei (ressourcen-)effizient durchgeführt werden. Zusätzlich sollten strategische und politische Empfehlungen für die sektorübergreifende Bereitstellung von Umweltdaten als Grundlage für künftige Fördermaßnahmen entwickelt werden.

Methodisch wurde ein explorativer Forschungsansatz gewählt, bestehend aus einem systematischen Review zum aktuellen Stand von Wissenschaft und Technik, einer Bedarfserhebung durch Umfragen und Stakeholder-Workshops sowie der prototypischen Umsetzung eines Entscheidungsmodells für die Datenannotation.

Die Potential- und Wirkungsanalyse zeigt, dass eine verbesserte Datenverfügbarkeit und Standardisierung essenziell für die effektive Nutzung von KI im Umweltsektor sind. Herausforderungen bestehen insbesondere in der Datenqualität, Interoperabilität und rechtlichen Rahmenbedingungen. Die Potentialanalyse unterstreicht, dass KI erhebliche Vorteile für das Umweltmanagement, politische Entscheidungsprozesse und nachhaltige Wirtschaftsstrategien bietet. Gleichzeitig bestehen ökologische Risiken, darunter ein hoher Energie- und Wasserverbrauch, die zunehmende Erzeugung von Elektroschrott, sowie ethische Risiken, welche die Förderung sozialer Ungleichheiten verstärken können.

Die erarbeiteten Handlungsempfehlungen für das Teilen von annotierten Umweltdaten beinhalten unter anderem die Förderung der FAIR-Prinzipien, die Standardisierung von Datenformaten, die Etablierung von Datenräumen und Datentreuhändern als digitale Ökosysteme und Anreizsysteme für die Datenbereitstellung.

Das Vorgehensmodell adressiert insbesondere die Entwicklung und Nutzung von ML-Anwendungen bei begrenzt verfügbaren annotierten Daten, wobei Pseudo-Labeling und Transfer Learning im Rahmen der Prototypisierung als besonders vielversprechend identifiziert wurden.

Die Ergebnisse tragen insgesamt dazu bei, KI-Anwendungen im Umweltbereich effizient und nachhaltig zu gestalten und umweltpolitische Maßnahmen datenbasiert zu unterstützen.

Abstract: Data annotation for ML applications in the environmental sector

The digital transformation means that data is increasingly seen as a valuable resource, especially in the environmental sector. The **LabelledGreenData4All** project investigates the strategic importance of annotated environmental data for the use of machine learning (ML) and, in a broader sense, artificial intelligence (AI) to tackle social and environmental challenges. The aim was to identify areas of application with high potential for ML models and to develop a process model that is intended to be used to carry out data annotations on the basis of defined decision paths and taking into account different initial situations with regard to the availability of annotated training data. The annotations themselves should be carried out in a (resource)

efficient manner. In addition, strategic and political recommendations for the cross-sectoral provision of environmental data were to be developed as a basis for future funding measures.

Methodologically, an explorative research approach was chosen, consisting of a systematic review of the current state of science and technology, a needs assessment through surveys and stakeholder workshops, and the prototypical implementation of a decision model for data annotation.

The potential and impact analysis shows that improved data availability and standardization are essential for the effective use of AI in the environmental sector. Challenges exist in particular in terms of data quality, interoperability and legal framework conditions. The potential analysis underlines that AI offers considerable advantages for environmental management, political decision-making processes and sustainable economic strategies. At the same time, there are ecological risks, including high energy and water consumption, the increasing generation of electronic waste, as well as ethical risks, which can reinforce the promotion of social inequalities.

The recommendations for the sharing of annotated environmental data include the promotion of FAIR principles, the standardization of data formats, the establishment of data spaces and data trustees as digital ecosystems and incentive systems for data provision.

The process model particularly addresses the development and use of ML applications with limited available annotated data, whereby pseudo-labeling and transfer learning were identified as particularly promising in the context of prototyping.

Overall, the results contribute to the efficient and sustainable design of AI applications in the environmental sector and provide data-based support for environmental policy measures.

Inhaltsverzeichnis

Abbildungsverzeichnis

Tabellenverzeichnis

Abkürzungsverzeichnis

Abkürzung	Erläuterung
AI	Artificial Intelligence
AIA	Automatic Image Annotation
AL	Active learning
AP	Arbeitspaket
API	Application Programming Interface
BMBF	Bundesministerium für Bildung und Forschung
BMUV	Bundesministerium für Umwelt, Naturschutz, nukleare Sicherheit und Verbraucherschutz
CA	Collaborative annotation / Kollaborative Annotation
CNN	Convolutional Neural Network
DL	Deep Learning
DTM	Datentreuhandmodelle
END	Environmental Noise Directive
ESA	Europäischen Raumfahrtbehörde
EU	Europäische Union
FAIR	Findable, Accessible, Interoperable, Reusable
FDG	Forschungsdatengesetz
GAN	Generative Adversarial Network
GAP	Gemeinsame Agrarpolitik
GBF	Kunming-Montreal Global Biodiversity Framework
GDI-DE	Geodateninfrastruktur Deutschland
GFW	Global Forest Watch
GFZ	GeoForschungsZentrum
GPU	Graphics Processing Unit
IDSA	International Data Spaces Association
IEA	Internationalen Energieagentur
IGD	Fraunhofer-Institut für Graphische Datenverarbeitung
IKT	Informations- und Kommunikationstechnik
ISPRS	International Society for Photogrammetry and Remote Sensing
KI	Künstliche Intelligenz
ML	Machine Learning

Abkürzung	Erläuterung
MLOps	Machine Learning Operations
NASA	National Aeronautics and Space Administration
NFDI	Nationale Forschungsdateninfrastruktur
NOAA	National Oceanic and Atmospheric Administration
OGC	Open Geospatial Consortium
SDG	Social Development Goal
SSL	Semi-überwachtes Lernen
UBA	Umweltbundesamt
UFZ	Helmholtz Zentrum für Umweltforschung GmbH
USGS	United States Geological Survey
VAE	Variational Autoencoder
WE	wetransform
WP	Work package
WS	Workshop

Zusammenfassung

Digitalisierung und die digitale Transformation führen dazu, dass Daten zu den größten Vermögenswerten gezählt und als eine wesentliche Ressource in allen Bereichen angesehen werden. Umweltdaten und umweltrelevante Daten sind essenziell für inter- und transdisziplinäre Forschung und haben ein enormes Innovationspotential. Durch die rasche Entwicklung von Methoden der künstlichen Intelligenz (KI) ergeben sich innovative Möglichkeiten, um die dringendsten gesellschafts- und umweltpolitischen Herausforderungen unserer Zeit zu analysieren und adäquate Lösungen zu finden. KI-Technologien haben das Potenzial, die Forschung in den Bereichen Umweltschutz und Nachhaltigkeit maßgebend voranzutreiben und zu verändern (BMUV, 2023).

Projekt LabelledGreenData4All

Das Projekt **LabelledGreenData4All** - „Nachhaltigkeitspotentialanalyse für die Zweckmäßigkeit und den Aufwand von Datenannotationen für ML-Modelle“ – zielte darauf ab, für das Umweltressort strategische Empfehlungen zu erarbeiten, in welchen Anwendungsbereichen und mit welchen Daten die größten Potentiale für den Einsatz von Machine Learning (ML)-Modellen bestehen.

Das Team der wetransform GmbH und des Fraunhofer-Institut für Graphische Datenverarbeitung IGD arbeiteten gemeinsam daran, die Innovationskraft annotierter Umweltdatensätze anhand von Anwendungsfeldern zu erforschen, um den Einsatz von KI im Umweltbereich effizient und nachhaltig zu gestalten. Wetransform leitete das Projekt und fokussierte sich auf den Bedarf, das Potential und die Wirkung annotierter Daten. Ein zentraler Aspekt war, die Verfügbarkeit von annotierten Umweltdaten und umweltrelevanten Daten zu verbessern und den sektorübergreifenden Datenaustausch in Datenräumen zu fördern. Das Fraunhofer IGD konzentrierte sich auf die Evaluierung und Analyse vorhandener Annotationsverfahren sowie ihrer Skalierbarkeit und ihrer Ergebnisqualität. Auf dieser Basis wurde ein Vorgehensmodell für die Datenannotation unter Berücksichtigung verschiedener Anwendungsfälle und am Beispiel von Fernerkundungsdaten entwickelt. Die Forschungsergebnisse von **LabelledGreenData4All** tragen dazu bei, KI im Umweltbereich effizient und nachhaltig zu nutzen und KI-basierten Klima- und Umweltschutz zu beschleunigen.

Methodisches Vorgehen

Das Vorhaben gliedert sich in insgesamt fünf Arbeitspakete. Neben einem Arbeitspaket (AP) zum Projektmanagement und einem für den kreativen Ergebnis- und Wissenstransfer, die Wissenschaftskommunikation sowie das Community Building, beinhaltete das Vorhaben drei fachlich-inhaltlich geprägte AP.

Für die Durchführung der Studie wurde ein explorativer Ansatz gewählt: Vom systematischen Review zum Stand der Wissenschaft und Technik in AP 2, über die Bedarfsermittlung und einer Potential- und Wirkungsanalyse in AP 3 bis hin zur Entwicklung und prototypischen Umsetzung eines Vorgehensmodells in AP 4.

Die Literaturrecherche in AP 2 hatte das Ziel, das derzeitige Verständnis und die Praxis der Datenannotation für ML-Anwendungen im Bereich der Umweltforschung und Nachhaltigkeit zu erheben und zu bewerten. Dabei wurde ein systematischer und standardisierter Ansatz entwickelt, der unter anderem die Definition spezifischer Suchanfragen, die Festlegung von Ein- und Ausschlusskriterien sowie eine systematische Überprüfung umfasste.

Aufbauend auf der Analyse und den Ergebnissen aus AP 2 wurde eine vertiefende Literaturrecherche mit dem Fokus auf die Anforderungen und den Bedarf an Umweltdaten und

umweltrelevanten Daten durchgeführt, wobei der Schwerpunkt auf den Sektoren Forst- und Landwirtschaft, Biodiversität, Mobilität und Klimaanpassung im urbanen Raum lag. Es wurden darüber hinaus die Potentiale und möglichen Wirkungen von ML-Verfahren im Allgemeinen und annotierter Daten im Speziellen betrachtet. Zudem wurden die Stellungnahmen zum Forschungsdatengesetz in der Analyse berücksichtigt.

Zur Erhebung des Bedarfes an annotierten Umweltdaten bzw. umweltrelevanten Daten wurde im ersten Schritt ein Fragebogen zum Thema „ML-Modelle und Datenannotation und deren Datenbedarf für KI-Anwendungen im Umweltsektor“ konzipiert und eine Onlineumfrage durchgeführt. Der Befragungszeitraum war von Juni 2024 bis Oktober 2024. Weiterhin wurden die Stakeholder des Vorhabens im Zuge von fünf virtuellen Stakeholder-Workshops im Zeitraum von Juni 2024 bis Oktober 2024 in die Untersuchung miteinbezogen. Thematisch fokussierten sich die Stakeholder-Workshops auf den Bedarf sowie das Potential annotierter Daten insbesondere in der Landwirtschaft, der Forstwirtschaft und im Sektor Biodiversität.

In AP 4 wurde schließlich ein Vorgehensmodell für die Datenannotation als Entscheidungshilfe für die Entwicklung von ML-Anwendungen mit wenigen annotierten Daten ausgearbeitet. Für die Entwicklung und Evaluierung des Vorgehensmodells wurden die folgenden Schritte bearbeitet:

- Eine Plattform zur automatisierten Modellierung und Ausführung von Datenverarbeitungsprozessen für das effiziente Training von KI-Modellen wurde eingerichtet.
- Verschiedene Annotationsmethoden wurden evaluiert.
- Ein Entscheidungsbaum wurde entwickelt, um ML-Methoden basierend auf der Verfügbarkeit annotierter Daten auszuwählen.
- Eine prototypische Umsetzung und Evaluierung erfolgte durch die Anwendung des entwickelten Vorgehensmodells auf zwei ausgewählte Anwendungsfälle sowie einen Referenzdatensatz.

Stand der Wissenschaft und Technik

Im systematischen Review zum Stand der Wissenschaft und Technik wurden drei Hauptkategorien von Annotationsmethoden identifiziert, die sich auf die am Prozess beteiligten Akteurinnen und Akteure beziehen: nutzer*inbasierte Annotation, Mensch-KI-Kollaboration und KI-basierte Annotation. Während die nutzer*inbasierte Annotation ein traditioneller Ansatz ist, der auf der manuellen Annotation von Daten durch Menschen basiert, ist die Mensch-KI-Kollaboration ein hybrider Ansatz, bei dem Menschen und KI-Systeme zusammenarbeiten, um Daten zu annotieren. Reine KI-Methoden zielen darauf ab, den Mangel an annotierten Daten zu beheben, indem entweder mehr Datenpunkte generiert oder Algorithmen entwickelt werden, die aus begrenzt verfügbaren oder verrauschten Daten lernen können. Da der Mangel an Trainingsdaten nach wie vor eine große Hürde darstellt, wurden insbesondere ML-Methoden und -Ansätze, die mit wenigen oder keinen annotierten Trainingsdaten arbeiten können, für die Entwicklung des Vorgehensmodells in AP 4 berücksichtigt.

Ergebnisse der Bedarfsermittlung

Die thematische Bedarfsermittlung verdeutlicht das Bedürfnis nach qualitätsgesicherten und leicht zugänglichen Daten, um Umweltprobleme wie Biodiversitätsverlust und Klimawandel effizient zu adressieren. Speziell in der Forstwirtschaft sind umfassende, standardisierte Inventurdaten erforderlich, die jedoch oft aus betriebsinternen Gründen nicht geteilt werden. Im

Bereich Biodiversität besteht ein dringender Bedarf an aktuellen und interoperablen Daten sowie Daten mit einer hohen räumlichen Repräsentativität. In der Landwirtschaft wird vor allem eine verbesserte Datenverfügbarkeit für die Erkennung von Pflanzenarten und Schaderregern angestrebt.

Die Bedarfsermittlung zeigt weiterhin, dass insbesondere der Zugang zu annotierten Umweltdaten als herausfordernd eingeschätzt und vielfach durch technische, rechtliche und infrastrukturelle Hindernisse eingeschränkt wird. Die Analyse veranschaulicht des Weiteren die Vielfalt bestehender Plattformen für Umweltdaten und benennt die zentralen Plattformtypen, die teilweise auch Integrationen für Datenwissenschaftler*innen umfassen (z. B. GitHub, Kaggle). Diese Plattformen stellen jedoch häufig keinen vollständigen API-Zugang bereit, was die Datenverfügbarkeit einschränkt. Weitere Schwächen sind das Fehlen von Metadaten und die eingeschränkte Auffindbarkeit von Datensätzen sowie das Nichtvorhandensein von Mechanismen für die gemeinsame Nutzung sensibler Daten.

Zu den wesentlichen Herausforderungen in Bezug auf die sektorübergreifende Zugänglichkeit von Umweltdaten im Allgemeinen und annotierten Umweltdaten im Besonderen zählen Defizite in den Bereichen Datenqualität, Interoperabilität und Datenstandardisierung. Technische Barrieren, die fehlende Verwendung offener Datenformate und unzureichende Standards für Annotationen erschweren die effektive Datennutzung. Ein klar definierter rechtlicher Rahmen, insbesondere durch einheitliche Lizenzen und transparente Nutzungsbedingungen, kann dabei die Nachnutzbarkeit und den rechtssicheren Umgang mit Daten verbessern. Initiativen wie Gaia-X und die Common European Data Spaces sollen diese Aspekte adressieren, befinden sich jedoch noch im Aufbau und sind erst mittelfristig verfügbar.

Die Empfehlungen für die Förderung der Nutzbarkeit, Verfügbarkeit und Sichtbarkeit relevanter Daten mit Umweltbezug aus verschiedenen Sektoren beinhalten eine stärkere Implementierung der FAIR-Prinzipien, die Förderung einer nachhaltigen Dateninfrastruktur und die Schaffung von Anreizsystemen für Datenbereitstellung und -nutzung. Die Einführung von Metadatenstandards, nachhaltigen Datentreuhandmodellen sowie einer verstärkten sektorübergreifenden Datenintegration könnten die Basis für eine leistungsfähige, vertrauenswürdige Dateninfrastruktur im Umweltsektor bilden.

Potentialanalyse

Die Potentialanalyse zeigt auf, dass der Einsatz von KI- und ML-Methoden im Umweltbereich signifikante Vorteile bietet und diese sektorübergreifend genutzt werden können. Politisch gesehen, unterstützt die datenbasierte Analyse politische Entscheidungsträger*innen im Umwelt- und Naturschutz, insbesondere durch die Bereitstellung fundierter Empfehlungen. Ein Beispiel ist die EU-Agrarpolitik, in der KI-gestützt Satellitendaten genutzt werden, um die Einhaltung von Förderkriterien aus der Gemeinsamen Agrarpolitik (GAP) zu kontrollieren. Auch im Bereich der forstwirtschaftlichen Zertifizierung, wie dem Forest Stewardship Council (FSC), wird KI genutzt, um die Effizienz und Transparenz von Nachhaltigkeitsbewertungen zu erhöhen.

In Bezug auf die wissenschaftliche und technologische Perspektive kann KI ein besseres Verständnis komplexer Umweltsysteme fördern und Vorhersagemodelle für Umweltüberwachung und -schutz optimieren. Dabei verbessert die Integration von Big Data und KI die Analyse von Umweltdaten und ermöglicht neue Ansätze im Katastrophenmanagement und der Klimamodellierung.

Wirtschaftlich gesehen, optimieren ML-Verfahren Prozesse in Land- und Forstwirtschaft, z. B. durch eine effizientere Ressourcenverwaltung und Ernteprognosen, was zur Reduktion von Betriebskosten und zum Erhalt der Biodiversität beiträgt.

In sozialer und ökologischer Hinsicht fördert KI die Erreichung der Nachhaltigkeitsziele (SDGs) durch eine effizientere Analyse und strategische Planung von Maßnahmen im Bereich des Umweltschutzes und der Nachhaltigkeit. So kann ML etwa das Wassermanagement und das Monitoring gefährdeter Arten unterstützen und damit einen direkten Beitrag zu Erfüllung der SGDs leisten. Darüber hinaus können KI- und ML-Anwendungen umweltpolitische Prozesse und nachhaltige Marktgestaltung fördern, indem sie Transparenz über Herstellungsprozesse und Produktursprünge schaffen sowie den Handel zu umweltgerechten Strategien anregen. Ein digitaler Produktpass und regulatorische Maßnahmen sind dabei zentrale Instrumente, um informierte Konsumentscheidungen zu ermöglichen und nachhaltigen Konsum für alle zugänglich zu machen.

Wirkungsanalyse

Die Wirkungsanalyse weist auf die komplexen Herausforderungen hin, die positiven und negativen Effekte von KI bzw. ML-Verfahren in Bezug auf umwelt- und gesellschaftspolitische Ziele in geeigneter Weise auszubalancieren. Zu den positiven Wirkungen gehört die Unterstützung datenbasierter Entscheidungen, welche unter anderem eine evidenzbasierte politische Beratung und eine effektive Umsetzung von Umweltrichtlinien ermöglichen. Durch präzise und annotierte Datensätze kann die Durchführung von Risikoanalysen optimiert werden, was etwa im Forst- und Agrarsektor für klimaangepasstes Management und frühzeitige Schädlingsbekämpfung genutzt werden kann. Darüber hinaus fördern KI-gestützte Technologien die Erreichung der Nachhaltigkeitsziele und wirken in Bereichen wie Biodiversitätsschutz und Klimawandel mit hoher Effektivität. Studien zeigen, dass KI etwa 79 % der SDG-Unterziele positiv beeinflussen kann.

Gleichzeitig sind signifikante negative Wirkungen in mehreren Dimensionen zu beobachten. In Bezug auf die Nachhaltigkeitsziele könnte KI in etwa 35 % der Unterziele hinderlich sein, insbesondere durch ihren hohen Energiebedarf, der das Ziel „Bezahlbare und saubere Energie“ (SDG 7) und „Klimaschutzmaßnahmen“ (SDG 13) negativ beeinflusst. Die für den Einsatz von KI erforderlichen Rechenzentren verursachen erhebliche CO_2-Emissionen und steigenden Wasserverbrauch, was sowohl ökologische als auch soziale Konflikte verschärfen kann. Ein weiterer problematischer Aspekt ist der steigende globale Elektroschrott des IT-Sektors, welcher allein 2022 weltweit etwa 62 Millionen Tonnen betrug (Baldé, et al., 2024). Hinzu kommt der Ressourcenverbrauch für die Produktion und Entsorgung von Hardware.

Da Geräte für den Betrieb von KI-Infrastrukturen und die KI-Nutzung oft in Ländern mit niedrigen Arbeitsstandards gefertigt werden, sind darüber hinaus soziale und ethische Implikationen in die Betrachtung einzubeziehen. Der KI-Einsatz könnte globale und lokale Ungleichheiten verschärfen, weil der Zugang zu den erforderlichen Rechen- und Bildungsressourcen weltweit ungleich verteilt ist. Zudem könnte die Ökosystemüberwachung durch KI auch zur Ausbeutung natürlicher Ressourcen beitragen, wenn wirtschaftliche Interessen dominieren, wie etwa bei der Unterstützung der Rohstoffwirtschaft durch KI-Dienste.

Politische Handlungsempfehlungen

Die politischen Handlungsempfehlungen für den Einsatz und das Management von annotierten Daten im Umweltbereich umfassen mehrere strategische Ansätze zur Verbesserung von Datenkompetenz, Datenzugänglichkeit und -sicherheit. Basierend auf den Analyseergebnissen wurden folgende Handlungsempfehlungen formuliert:

- **Förderung der Datenkompetenz und Implementierung von Datenmanagement-Richtlinien**: Um Daten nachhaltig und sicher zu nutzen, ist eine gezielte Förderung der Datenkompetenz in Verwaltung, Forschung und Industrie erforderlich. Ein konsistentes

Datenmanagement, und insbesondere die Etablierung von Data Governance, schützt wertvolle Daten vor Verlust und Missbrauch und ermöglicht die Einhaltung rechtlicher Standards. Verbindliche Datenrichtlinien sollten diesen Prozess unterstützen, etwa durch Schulungen und einheitliche Datenmanagementpraktiken.

- **Etablierung von Datenräumen und Datentreuhändern**: Offene Datenökosysteme nach Standards der International Data Spaces Association (IDSA) und Gaia-X bieten eine Lösung zur Überwindung isolierter Datenplattformen und fördern diskriminierungsfreien Zugang und nachhaltige Nutzung. Datentreuhänder*innen könnten als Vermittlende fungieren, den Zugang zu sensiblen Daten vereinfachen und eine langfristige Nutzung von Daten für Wissenschaft und Gesellschaft sicherstellen.
- **Standardisierung als Grundlage für das Datenteilen**: Die Anwendung offener Datenformate und standardisierter Schnittstellen (APIs) stellt eine zentrale Anforderung dar, um Daten nachhaltig nutzbar zu machen. Ein stufenweises Interoperabilitätsmodell sollte den schrittweisen Ausbau der Datenkompatibilität fördern und langfristig zu einer vollständigen Interoperabilität führen.
- **Rechtlicher Rahmen für Risikobewertung**: Ein evidenzbasierter Ansatz zur objektiven Bewertung der Risiken beim Datenaustausch ist essenziell, um eine verlässliche Grundlage für die Freigabe von Daten zu schaffen und Vorbehalte gegen die Weitergabe abzubauen.
- **Verpflichtendes Teilen von Forschungsdaten**: Daten aus öffentlich geförderten Projekten sollten nach den FAIR-Prinzipien bereitgestellt werden. Dabei sollten Forschende eine exklusive, aber zeitlich begrenzte Auswertungsphase erhalten, um die Daten und ihre Forschungsergebnisse wissenschaftlich zu publizieren. Projekte sollten bereits bei der Antragstellung einen Datenmanagementplan darlegen und könnten im Falle einer Nichteinhaltung Sanktionen, wie das Zurückhalten von Mitteln, erfahren.
- **Etablierung von Anreizstrukturen**: Neben Verpflichtungen sollten Anreize wie steuerliche Vorteile für Datenspenden und immaterielle Bilanzierungsmöglichkeiten für Unternehmen geschaffen werden. Weitere Anreize könnten sein, den Zugang zu spezifischen Funktionen auf Datenbereitstellende zu beschränken oder ihre Sichtbarkeit in Datenportalen zu erhöhen, um den Austausch und die Nutzung von Umweltdaten aktiv zu fördern.

Vorgehensmodell für die Datenannotation

Gegenstand von AP 4 war die Entwicklung eines Vorgehensmodells für die Datenannotation als Entscheidungshilfe für das Arbeiten mit wenigen annotieren Daten bei der Entwicklung von ML-Anwendungen im Umweltbereich sowie die Validierung dieses Modells anhand von zwei konkreten Anwendungsfällen: die Erkennung von Freiflächen-Photovoltaik (PV)-Anlagen und die Baumartenklassifikation in Wäldern. Die Studie fokussierte sich auf Fernerkundungsdaten (Luftbilder, Satellitenbilder, Punktwolkendaten). Maßgebliche Kriterien für das Design des Modells waren zudem Automatisierbarkeit und Skalierbarkeit der Annotation. Wirtschaftliche Aspekte wurden bei der Entwicklung des Vorgehensmodells nicht berücksichtigt.

Der gewählte Ansatz umfasste die Einrichtung einer technischen Experimentierplattform zur Automatisierung von Datenverarbeitungsprozessen sowie die Evaluierung verschiedener Annotationsmethoden im Bereich daten- und modellzentrierter Ansätze. Auf dieser Grundlage wurde zunächst ein Entscheidungsbaum entwickelt, der, abhängig von der Verfügbarkeit gelabelter Daten, geeignete Annotationsansätze vorschlägt. Für die anschließende prototypische Umsetzung wurden die komplementären Ansätze Pseudo-Labeling und Transfer Learning für die beiden Anwendungsfälle ausgewählt und - auch unter Einbeziehung des ISPRS-Potsdam

Datensatzes als Benchmark-Datensatz - evaluiert. Die Ergebnisse der prototypischen Umsetzung zeigen, dass mit Pseudo-Labeling selbst bei geringer Verfügbarkeit von gelabelten Daten qualitativ hochwertige Ergebnisse erzielt werden können. Zudem erweist sich Transfer Learning als ressourcenschonende Methode, die die Trainingszeit erheblich verkürzt.

Der Entscheidungsbaum als zentrales Ergebnis des AP 4 bietet Entwickler*innen von ML-Modellen Handlungsempfehlungen für verschiedene Szenarien, die sich hinsichtlich der Verfügbarkeit gelabelter Daten unterscheiden. Anhand der Ergebnisauswertung für die beiden Anwendungsfälle konnte veranschaulicht werden, dass eine Kombination aus daten- und modellzentrierten Ansätzen eine effektive Strategie zur Bewältigung der Herausforderungen bei der Arbeit mit wenigen annotierten Daten sein kann. Das entwickelte Vorgehensmodell fungiert somit als praktische Orientierungshilfe für die komplexen Aufgaben guter Modellbildung.

Das für die Annotation relevante Entscheidungskriterium der Qualität der Referenzlabels wurde nicht explizit als eigenes Kriterium betrachtet, sondern die Qualität stets als ausreichend angenommen. Grund sind die aktuell bestehenden Limitationen in der Qualitätsbewertung von Referenzlabels. Hier besteht ein dringender Bedarf an standardisierten Verfahren, um diese beurteilen und in die Entscheidungsfindung einbeziehen zu können. Darüber hinaus wird empfohlen, das Vorgehensmodell perspektivisch zu erweitern, um die Anwendungsfälle stärker zu differenzieren, weitere Datentypen zu berücksichtigen und weitere Entscheidungskriterien, wie z. B. Aspekte der Wirtschaftlichkeit und Nachhaltigkeit, aufzunehmen.

Summary

Digitalization and the digital transformation mean that data is one of the greatest assets and is seen as an essential resource in all areas. Environmental and environmentally relevant data are essential for interdisciplinary and transdisciplinary research and have enormous potential for innovation. The rapid development of artificial intelligence (AI) methods is creating innovative opportunities to analyze the most pressing social and environmental challenges of our time and find adequate solutions. AI technologies have the potential to significantly advance and change research in the fields of environmental protection and sustainability (BMUV, 2023).

Projekt „LabelledGreenData4All“

The **LabelledGreenData4All** project - “Sustainability potential analysis for the usefulness and effort of data annotations for ML models” - aimed to develop strategic recommendations for the environmental sector on which application areas and with which data the greatest potential for the use of machine learning (ML) models exists.

The team from wetransform GmbH and the Fraunhofer Institute for Computer Graphics Research IGD worked together to explore the innovative power of annotated environmental data sets based on fields of application to make the use of AI in the environmental sector efficient and sustainable. Wetransform led the project and focused on the need, potential and impact of annotated data. A central aspect was to improve the availability of annotated environmental data and environmentally relevant data and to promote cross-sectoral data exchange in data spaces. Fraunhofer IGD focused on evaluating and analyzing existing annotation methods, their scalability and the quality of their results. On this basis, a process model for data annotation was developed, taking into account various use cases and using the example of remote sensing data. The research results of **LabelledGreenData4All** contribute to the efficient and sustainable use of AI in the environmental sector and to accelerating AI-based climate and environmental protection.

Methodological approach

The project was divided into a total of five work packages (WP). In addition to the WP on project management, three WPs were characterized by technical content and one WP explicitly related to the topics of creative transfer of results and knowledge, science communication and community building.

An explorative approach was chosen to conduct the study: From the systematic review of the state of science and technology in WP 2, to the needs assessment and a potential and impact analysis in WP 3, through to the development and prototypical implementation of a process model in WP 4.

The literature review in WP 2 aimed to assess and evaluate the current understanding and practice of data annotation for ML applications in the field of environmental research and sustainability. A systematic and standardized approach was developed, including the definition of specific search queries, the definition of inclusion and exclusion criteria and a systematic review.

Building on the analysis and results from WP 2, an in-depth literature review was conducted with a focus on the requirements and needs for environmental data and environmentally relevant data, focusing on the sectors of forestry and agriculture, biodiversity, mobility and climate adaptation in urban areas. The potential and possible effects of ML processes in general and annotated data in particular were also considered. The comments on the Research Data Act were also taken into account in the analysis.

To determine the need for annotated environmental data or environmentally relevant data, the first step was to design a questionnaire on the topic of "ML models and data annotation and their data requirements for AI applications in the environmental sector" and conduct an online survey. The survey period was from June 2024 to October 2024. The project's stakeholders were also included in the study in the course of five virtual stakeholder workshops from June 2024 to October 2024. Thematically, the stakeholder workshops focused on the need for and potential of annotated data, particularly in agriculture, forestry and the biodiversity sector.

Finally, in WP 4, a process model for data annotation was developed as a decision-making support for the development of ML applications with scarce annotated data. The following steps were taken to develop and evaluate the process model:

- A platform for the automated modeling and execution of data processing procedures for the efficient training of AI models was set up.
- Various annotation methods were evaluated.
- A decision tree was developed to select ML methods based on the availability of annotated data.
- A prototype implementation and evaluation was carried out by applying the developed process model to two selected use cases and a reference dataset.

State of the Art

In the systematic review of the state of the art, three main categories of annotation strategies were identified that relate to the actors involved in the process: user-based annotation, human-AI collaboration and AI-based annotation. While user-based annotation is a traditional approach based on the manual annotation of data by humans, human-AI collaboration is a hybrid approach in which humans and AI systems work together to annotate data. Pure AI methods aim to address the lack of annotated data by either generating more data points or developing algorithms that can learn from limited or noisy data. As the scarcity of training data remains a major hurdle, the study focussed on ML methods and approaches that can work with scarce or no training data. These methods were considered for the development of the process model in WP 4.

Results of the needs assessment

The thematic needs assessment highlights the need for quality-assured and easily accessible data in order to efficiently address environmental problems such as biodiversity loss and climate change. Forestry in particular requires comprehensive, standardized inventory data, which is often not shared for internal reasons. In the area of biodiversity, there is an urgent need for up-to-date and interoperable data as well as data with a high degree of spatial representativeness. In agriculture, the main aim is to improve data availability for the identification of plant species and pathogens.

The needs assessment also shows that access to annotated environmental data in particular is considered challenging and is often restricted by technical, legal and infrastructural obstacles. The analysis further illustrates the diversity of existing platforms for environmental data and identifies the central platform types, some of which also include integrations for data scientists (e.g. GitHub, Kaggle). However, these platforms often do not provide full API access, which limits data availability. Other weaknesses include the lack of metadata and the limited findability of datasets, as well as the absence of mechanisms for sharing sensitive data.

The main challenges with regard to the cross-sectoral accessibility of environmental data in general and annotated environmental data in particular include deficits in the areas of data quality, interoperability and data standardization. Technical barriers, the lack of use of open data formats and inadequate standards for annotations hamper the effective use of data. A clearly defined legal framework, in particular through uniform licenses and transparent terms of use, can improve the reusability and legally compliant handling of data. Initiatives such as Gaia-X and the Common European Data Spaces are intended to address these aspects, but are still being developed and will only be available in the medium term.

The recommendations for promoting the usability, availability and visibility of relevant environmental data from different sectors include a stronger implementation of the FAIR principles, the promotion of a sustainable data infrastructure and the creation of incentive systems for data provision and use. The introduction of metadata standards, sustainable data trust models and increased cross-sectoral data integration could form the basis for an efficient, trustworthy data infrastructure in the environmental sector.

Potential analysis

The potential analysis shows that the use of AI and ML methods in the environmental sector offers significant advantages and that these can be used across sectors. From a political perspective, data-based analysis supports political decision-makers in environmental protection and nature conservation, in particular by providing well-founded recommendations. One example is EU agricultural policy, where AI-supported satellite data is used to monitor compliance with support criteria from the Common Agricultural Policy (CAP). AI is also used in the area of forestry certification, such as the Forest Stewardship Council (FSC), to increase the efficiency and transparency of sustainability assessments.

Scientifically and technologically, AI can promote a better understanding of complex environmental systems and optimize forecasting models for environmental monitoring and protection. The integration of big data and AI improves the analysis of environmental data and enables new approaches in disaster management and climate modeling.

In economic terms, ML methods optimize processes in agriculture and forestry, e.g. through more efficient resource management and harvest forecasts, which helps to reduce operating costs and preserve biodiversity.

In social and environmental terms, AI promotes the achievement of the Sustainable Development Goals (SDGs) through more efficient analysis and strategic planning of measures in the area of environmental protection and sustainability. For example, ML can support water management and the monitoring of endangered species and thus make a direct contribution to meeting the SGDs. In addition, AI and ML applications can promote environmental policy processes and sustainable market design by creating transparency about manufacturing processes and product origins and encouraging retailers to adopt environmentally friendly strategies. A digital product passport and regulatory measures are key instruments for enabling informed consumer decisions and making sustainable consumption accessible to all.

Impact analysis

The impact analysis points to the complex challenges of appropriately balancing the positive and negative effects of AI and ML processes in relation to environmental and socio-political goals. The positive effects include the support of data-based decisions, which enable evidence-based policy advice and the effective implementation of environmental guidelines, among other things. Precise and annotated data sets can optimize the implementation of risk analyses, which can be used for climate-adapted management and early pest control in the forestry and agricultural

sectors, for example. In addition, AI-supported technologies promote the achievement of sustainability goals and are highly effective in areas such as biodiversity conservation and climate change. Studies show that AI can have a positive impact on around 79 % of the SDG targets.

At the same time, significant negative effects can be observed in several dimensions. With regard to the sustainability goals, AI could be a hindrance in around 35 % of the sub-goals, particularly due to its high energy requirements, which negatively impacts the goal "Affordable and clean energy" (SDG 7) and "Climate action" (SDG 13). The data centers required for the use of AI cause considerable CO_2 emissions and increasing water consumption, which can exacerbate both ecological and social conflicts. Another problematic aspect is the increasing global e-waste from the IT sector, which amounted to around 62 million tons worldwide in 2022 alone (Baldé, et al., 2024). Added to this is the consumption of resources for the production and disposal of hardware.

As devices for the operation of AI infrastructures and the use of AI are often manufactured in countries with low labor standards, social and ethical implications must also be taken into account. The use of AI could exacerbate global and local inequalities because access to the necessary computing and educational resources is unevenly distributed worldwide. In addition, ecosystem monitoring using AI could also contribute to the exploitation of natural resources if economic interests dominate, such as when AI services are used to support the raw materials industry.

Policy recommendations

The policy recommendations for the use and management of annotated data in the environmental sector include several strategic approaches to improve data literacy, data accessibility and data security. Based on the results of the analysis, the following recommendations for action were formulated:

- **Promote data literacy and implement data management guidelines**: To use data sustainably and securely, targeted promotion of data literacy in administration, research and industry is required. Consistent data management, and in particular the establishment of data governance, protects valuable data from loss and misuse and enables compliance with legal standards. Binding data guidelines should support this process, for example through training and standardized data management practices.
- **Establishment of data rooms and data trustees**: Open data ecosystems according to the standards of the International Data Spaces Association (IDSA) and Gaia-X offer a solution to overcome isolated data platforms and promote non-discriminatory access and sustainable use. Data trustees could act as intermediaries, simplify access to sensitive data and ensure the long-term use of data for science and society.
- **Standardization as the basis for data sharing**: The use of open data formats and standardized interfaces (APIs) is a key requirement for making data usable in the long term. A step-by-step interoperability model should promote the gradual expansion of data compatibility and lead to complete interoperability in the long term.
- **Legal framework for risk assessment**: An evidence-based approach to objectively assess the risks of data sharing is essential to create a reliable basis for sharing data and to reduce reservations about sharing.
- **Mandatory sharing of research data**: Data from publicly funded projects should be made available in accordance with the FAIR principles. Researchers should be given an exclusive

but time-limited evaluation phase in order to publish the data and their research results scientifically. Projects should present a data management plan at the application stage and could be subject to sanctions, such as the withholding of funds, in the event of non-compliance.

- **Establishment of incentive structures**: In addition to obligations, incentives such as tax benefits for data donations and immaterial accounting opportunities for companies should be created. Further incentives could be to restrict access to specific functions to data providers or to increase their visibility in data portals in order to actively promote the exchange and use of environmental data.

Procedure Model for data annotation

The subject of WP 4 was the development of a procedure model for data annotation as a decision-making aid for working with a small amount of annotated data when developing ML applications in the environmental field and the validation of this model using two specific use cases: the detection of ground-mounted photovoltaic (PV) systems and the classification of tree species in forests. The study focused on remote sensing data (aerial images, satellite images, point cloud data). Other key criteria for the design of the model were the automation and scalability of the annotation. Economic aspects were not taken into account for the development.

The chosen approach comprised the establishment of a technical experimentation platform for the automation of data processing procedures and the evaluation of various annotation methods in the area of data- and model-centered approaches. On this basis, a decision tree was first developed that suggests suitable annotation approaches depending on the availability of labeled data. For the subsequent prototypical implementation, the complementary approaches of pseudo-labeling and transfer learning were selected for the two use cases and evaluated - also using the ISPRS-Potsdam dataset as a benchmark dataset. The results of the prototype implementation show that high-quality results can be achieved with pseudo-labeling even when labeled data is not readily available. In addition, transfer learning proves to be a resource-saving method that considerably shortens the training time.

The decision tree as the central result of WP 4 offers developers of ML models recommendations for action for various scenarios that differ in terms of the availability of labeled data. The evaluation of the results for the two use cases illustrated that a combination of data-centric and model-centric approaches can be an effective strategy for overcoming the challenges of working with little annotated data. The developed process model thus serves as a practical orientation aid for the complex tasks of good modeling.

The decision criterion of the quality of the reference labels, which is relevant for the annotation, was not explicitly considered as a separate criterion, but the quality was always assumed to be sufficient. This is due to the current limitations in the quality assessment of reference labels. There is an urgent need for standardized procedures in order to be able to assess these and include them in the decision-making process. In addition, it is recommended that the process model be expanded in the future in order to differentiate the use cases more strongly, to consider additional data types and to include further decision criteria, such as aspects of economic efficiency and sustainability.

1 Einleitung

1.1 Bedeutung annotierter Datensätze

Digitalisierung und die digitale Transformation führen dazu, dass Daten zu den größten Vermögenswerten gezählt und als eine wesentliche Ressource in allen Bereichen angesehen werden. Umweltdaten und umweltrelevante Daten sind essenziell für inter- und transdisziplinäre Forschung und haben ein enormes Innovationspotential. So ergeben sich durch die rasche Entwicklung von Methoden der künstlichen Intelligenz (KI) innovative Möglichkeiten, um die dringendsten gesellschafts- und umweltpolitischen Herausforderungen unserer Zeit zu analysieren und Lösungen zu finden. KI-Technologien haben das Potenzial, die Forschung in den Bereichen Umweltschutz und Nachhaltigkeit maßgebend voranzutreiben und zu verändern (BMUV, 2023).

KI-Werkzeuge bieten zwar weitreichende Möglichkeiten, jedoch müssen ihre Ergebnisse zuverlässig sein. Neben erklär- oder interpretierbaren Modellen ist vor allem die Verfügbarkeit von qualitativ hochwertigen, harmonisierten Trainingsdaten (im Folgenden *annotierten oder gelabelten Daten*) essenziell, um zuverlässige, breit anwendbare KI-Modelle aufzubauen (Halevy, Norvig, & Pereira, 2009). Annotierte Daten bilden die Grundlage für eine gute Modellbildung und dienen als treibende Kraft für die Weiterentwicklung KI-gestützter Umweltforschung. Dabei beeinflussen vor allem die Datengenauigkeit und -qualität das finale Ergebnis. Der Aufwand, der für die originäre Datengenerierung und –aufarbeitung notwendig ist, ist grundsätzlich sehr hoch. Daher sind insbesondere das Erkennen von Datenlücken, die Identifizierung von Anwendungsbereichen mit großem Bedarf und die Entwicklung von effizienten Annotationsmethoden essenziell für den erfolgreichen und wirksamen Einsatz von KI im Umweltbereich.

1.2 Projekt LabelledGreenData4All - Motivation und Projektziele

Angesichts des zunehmenden Einsatzes von KI-Technologien in der gesamten Gesellschaft ist es wichtig zu verstehen, auf welche Weise KI den Fortschritt beim Klima- und Umweltschutz beschleunigen oder behindern kann und wie verschiedene Interessensgruppen diese Entwicklungen steuern können. So kann KI die Entwicklung von Klimaschutz- und Anpassungsstrategien in einer Reihe von Sektoren wie Energie, Produktion, Land- und Forstwirtschaft und Katastrophenmanagement unterstützen und erleichtern (BMUV, 2023). Sowohl die Nutzung von KI für die Erreichung der Nachhaltigkeitsziele als auch die nachhaltige Gestaltung und Nutzung von KI sind dabei aber auch mit wirtschaftlichen Aspekten verknüpft (Vinuesa, et al., 2020). Während Investitionen in KI-Anwendungen zunächst Kosten verursachen, können die langfristigen Vorteile erheblich sein und zu Ressourcen- und Kosteneinsparungen führen (Xu, et al., 2021). Andererseits kann KI auch zum Anstieg der Treibhausgasemissionen beitragen, wenn Anwendungen in Sektoren mit hohem Schadstoffausstoß zum Einsatz kommen oder die Verbrauchernachfrage steigern, sowie durch den damit verbundenen Wasser- und Energieverbrauch (IEA, 2024; Ganesh, McCallum, & Strubell, 2019; Lizarraga & Solon , 2023).

Ziel des Projektes **LabelledGreenData4All** („Nachhaltigkeitspotentialanalyse für die Zweckmäßigkeit und den Aufwand von Datenannotationen für ML-Modelle“) war es, für das Umweltressort strategische Empfehlungen zu erarbeiten, in welchen Anwendungsbereichen und mit welchen Daten die größten Potentiale für den Einsatz von Machine Learning (ML)-Modellen bestehen. Darauf aufbauend waren die Entwicklung eines Vorgehensmodells für die Datenannotation unter Berücksichtigung verschiedener Datenarten und Anwendungsfälle sowie

die prototypische Umsetzung annotierter Umweltdatensets (Trainingsdaten als Schlüssel) konkrete Schritte, um die Forschung in die Praxis umzusetzen. Das Projekt hat nicht zuletzt auch die Frage behandelt, wie das Teilen von sogenannten „gelabelten" bzw. annotierten Umweltdaten aus der Ressortforschung des Bundes unterstützt und wie diese in Datenräumen sektorübergreifend geteilt werden können (Datenräume als digitales Ökosystem).

Im Detail sollten folgende Ziele erreicht werden:

1. Erhebung und Analyse von Zweck-Mittel-Relationen, um Bereiche im Umweltsektor und angrenzenden Sektoren zu identifizieren, wo annotierte Daten gut genutzt werden können
2. Förderung der Entwicklung innovativer KI-Anwendungen
3. Aufdecken von Leerstellen, um eine mehrwertstiftende Nutzung von Zukunftstechnologien wie KI innerhalb des Umweltbereiches und über Sektoren hinweg zu fördern
4. Reduzierung der Fragmentierung und Erhöhung der Verfügbarkeit annotierter Datensätze sowie Möglichmachen des sektorübergreifenden Teilens
5. Ressourcen- und energieeffiziente Datenannotationsprozesse

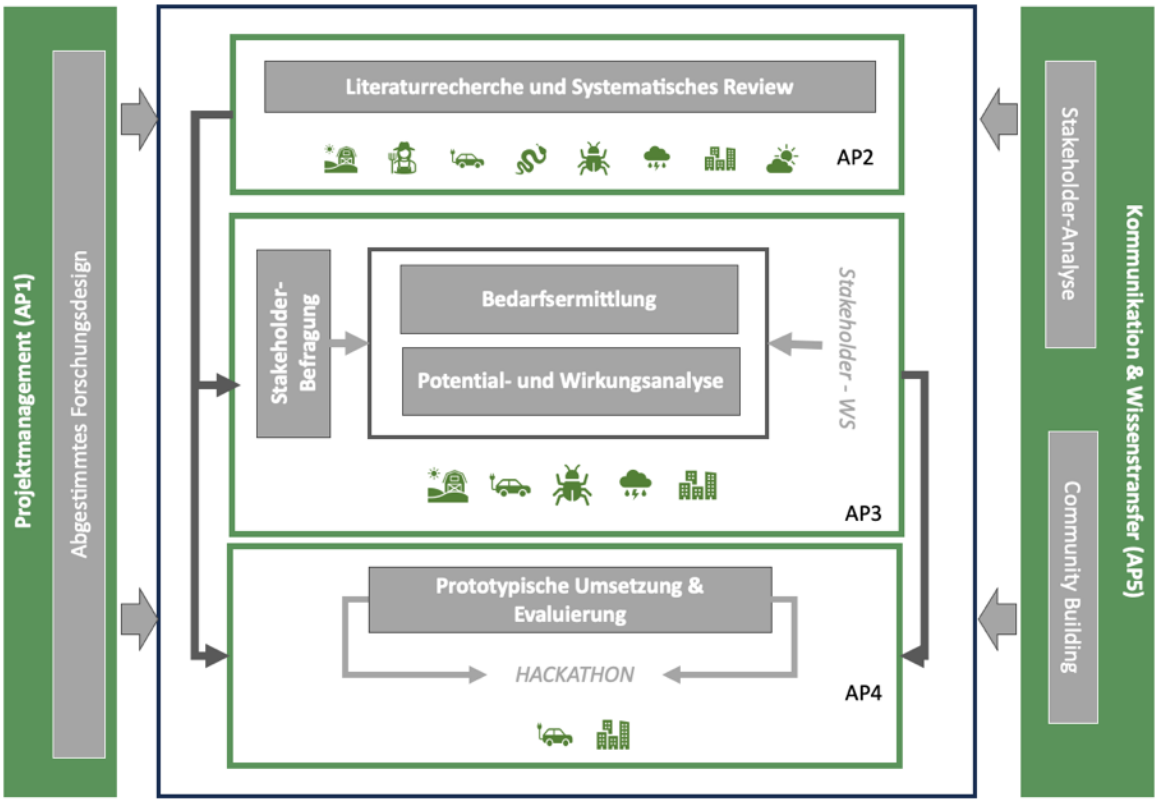
Um anwendbare Ergebnisse zu liefern, konzentrierten sich die Analysen im Forschungsvorhaben auf die folgenden thematischen Bereiche:

- Land- und Forstwirtschaft
- Biodiversität
- Mobilität
- Klimafolgenanpassungen im urbanen Raum

Im Forschungsvorhaben wurde ein datenzentrierter Ansatz mit dem Ziel, die Bereitstellung von qualitativ hochwertigen Umweltdaten bzw. umweltrelevanten Daten nach den „FAIR"-Prinzipien[1] zu verbessern, verfolgt. In Arbeitspaket (AP) 2 wurde dazu der Stand der Wissenschaft zur Annotation von Umweltdaten bzw. umweltrelevanten Daten für ML-basierte Umwelt- und Nachhaltigkeitsanwendungen mit dem Fokus Methoden, Wirtschaftlichkeit und Datenverfügbarkeit mittels systematischem Review ermittelt und zusammengefasst. AP 3 beinhaltete eine Bedarfsermittlung für annotierte Umweltdatensets sowie eine Potential- und Wirkungsanalyse dieser und zielte darauf ab, mittels aktiver Einbindung von Stakeholdern einen tieferen Einblick in die jeweiligen Sektoren zu erlangen. Basierend darauf erfolgte eine prototypische Umsetzung des in AP 4 entwickelten Vorgehensmodells für die Datenannotation, um skalierbare Annotationsmethoden für die ressort- und sektorübergreifende Bereitstellung annotierter Datensets anhand zweier Anwendungsfälle zu demonstrieren. Begleitet wurde die fachliche Arbeit durch ein professionelles Projektmanagement (AP 1) und ein dediziertes Arbeitspaket für Wissenschaftskommunikation, Wissenstransfer und Community-Building (AP 5). Letzteres sollte eine effektive Verwertung der Forschungsergebnisse garantieren.

[1] https://www.go-fair.org/fair-principles/

Abbildung 1: Schematische Darstellung des Projektdesigns



Quelle: eigene Darstellung, wetransform GmbH

1.3 Aufbau des Berichtes

Ziel des vorliegenden Abschlussberichtes ist es, ausgehend von einer systematischen Literaturarbeit zu Annotationsansätzen bei ML-Verfahren mit Fokus auf Methoden, Wirtschaftlichkeitsaspekte und den Umgang mit wenigen verfügbaren Daten einen Überblick zu Bedarf sowie Potential und Wirkung annotierter Daten im Umweltbereich zu geben, die Ergebnisse der prototypischen Umsetzung des entwickelten Vorgehensmodells für die Datenannotation zu veranschaulichen und weiteren Forschungsbedarf zusammenzufassen.

Nach der Beschreibung des methodischen Vorgehens (Kapitel 2) wird in Kapitel 3 der Stand der Wissenschaft und Technik detailliert. Dem folgen die Ergebnisse der Bedarfsermittlung mit Fokus auf Land- und Forstwirtschaft sowie Biodiversität (Kapitel 4). Das Potential annotierter Daten wird in Kapitel 5 dargestellt, wobei eine Unterscheidung zwischen politischem, wirtschaftlichem, wissenschaftlich/technischem, sozialem und ökologischem Potential getroffen wird. Die Wirkungsanalyse betrachtet positive und negative Wirkungen, welche gegeneinander abgewogen werden sollten, bevor ML-Modelle eingesetzt werden (Kapitel 6). Mögliche politische Handlungsempfehlungen zur Förderung und Forcierung der Einrichtung und Etablierung kollaborativer Datenräume als Plattformen für die Nachnutzung insbesondere auch von sensiblen Trainingsdaten werden in Kapitel 7 erörtert. Die Ergebnisse zur Erarbeitung eines Vorgehensmodells zum Umgang mit wenigen annotierten Daten für die Entwicklung von ML-Anwendungen sind in Kapitel 8 dargestellt. Der Bericht schließt mit einem Kapitel zu den Engpässen, den Potentialen und Synergieeffekten sowie möglichen Anknüpfungspunkten für weitere Forschungsvorhaben aus dem Projekt ab (Kapitel 9).

2 Methodisches Vorgehen

2.1 Systematisches Review

Das Hauptziel des Literaturreviews war es, das aktuelle Verständnis und die Praktiken rund um das Thema Datenannotation für ML-Anwendungen im Bereich der Umweltforschung und Nachhaltigkeit zu recherchieren und zu bewerten. Der Rechercheraum ist dabei mehrdimensional, mit einer Vielzahl von Domänen, Datenformaten, Labeling-Strategien und ML-Modellen, die berücksichtigt werden müssen. Aufgrund dieser Komplexität konzentrierte sich die Studie daher auf die Domänen Land- und Forstwirtschaft, Biodiversität, Mobilität und Klimafolgenanpassungen im urbanen Raum. Darüber hinaus war das Review auf die folgenden Datenformate beschränkt:

- Rasterdaten (z.B. Satellitendaten)
- Vektordaten (z.B. Sensordaten)
- Punktwolken (z.B. Laserscanning) und
- andere georeferenzierte Formate.

Tabellarische Daten oder Textdaten wurden im Rahmen dieses Forschungsvorhabens nicht berücksichtigt.

Das methodische Vorgehen für die systematische Literaturarbeit zum Stand von Wissenschaft und Technik in Bezug auf Annotationsansätze bei ML-Verfahren im Allgemeinen und in umweltrelevanten Sektoren im Speziellen mit Fokus auf Wirtschaftlichkeitsaspekte umfasst die in den folgenden Abschnitten beschriebenen Kernelemente.

2.1.1 Identifikation von Forschungsfragen

Zu Beginn des Projekts wurde eine Reihe von Forschungsfragen identifiziert, die die Kernziele der Literaturarbeit adressieren und dazu beitrugen, den Umfang der Recherche fokussiert zu halten. Generell lässt sich das Forschungsthema in zwei übergeordneten Themen zusammenfassen:

1. Daten und ihre Annotation (Forschungsfragen 1-4) und
2. KI-Methoden, die gut mit begrenzt verfügbaren Daten umgehen können (Forschungsfragen 5 und 6).

Forschungsfragen:

1. Welche Daten werden verwendet, um Machine-Learning-Modelle im Umweltsektor zu trainieren?
2. Welche Methoden gibt es zum Annotieren von Daten?
3. Befassen sich die Autorinnen*Autoren mit dem Kostenfaktor (Zeit, Aufwand), der sich auf die in ihren Arbeiten vorgestellten Annotationsmethoden bezieht?
4. Inwiefern sind ausreichend annotierte Daten in Formaten zugänglich, die sich für das Training von Machine Learning-Modellen eignen?
5. Welche Machine Learning-Ansätze/Modelle kommen zum Einsatz?
6. Welche Modelle in der Umweltforschung können gut mit wenigen Trainingsdaten arbeiten?

2.1.2 Ableitung von Kernsuchbegriffen

Basierend auf den zuvor definierten Forschungsfragen wurden zentrale Suchbegriffe identifiziert. Diese Suchbegriffe wurden nach ihrer semantischen Funktion in der Literaturrecherche kategorisiert. Die Suchbegriffe, die zum Filtern relevanter Arbeiten verwendet werden, sind in Tabelle 1 aufgeführt.

Tabelle 1: Angewendete Suchbegriffe

Kategorie	Suchbegriff
Overarching subject matter	Deep Learning Machine Learning
Sub-Interest within the subject matter	Data Annotation Data Labeling
Concrete problem of interest	Scarcity of Data Availability of Data Unavailability of Data
Data domains of interest	Satellite Data Orthophotography Aerial Photography Point Clouds
Application domains of interest	Mobility Agriculture Forest Biodiversity Climate

2.1.3 Auswahl von Literaturdatenbanken und Abfragetools

Die Wahl der Literaturdatenbank ist ein wesentlicher Bestandteil für eine effektive Literaturrecherche. Scopus (Elsevier) wurde als primäre Quellendatenbank für die Suche ausgewählt, da sie eine umfangreiche, multidisziplinäre Sammlung von Peer-Review-Zeitschriften, Konferenzberichten und Patenten bietet und eine umfassende Abdeckung der wissenschaftlichen Ergebnisse gewährleistet. Die fortschrittlichen Such- und Analysetools helfen Forschenden, relevante Literatur effizient zu finden, Zitationstrends zu verfolgen und den Einfluss der Forschung zu bewerten. Scopus bietet eine ausgereifte Suchschnittstelle mit logischen Operatoren zum Gruppieren von Begriffen und Klammern zur Kontrolle der Rangfolge sowie Funktionen wie die Abstandssuche zwischen Begriffen und Platzhaltern. Diese Eigenschaften machen Scopus gut geeignet für eine systematische Übersichtsarbeit.

2.1.4 Definition der Suchanfragen

Im Rahmen der Literaturarbeit wurden zwei Suchansätze verwendet, die von Scopus zur Verfügung gestellt werden. Zunächst wurde mit der erweiterten Dokumentensuche gearbeitet. Dies ist eine Schnittstelle, die es ermöglicht, ausdrucksstarke Abfragen zu erstellen, indem mehrere Suchwörter, Platzhalter usw. mit Hilfe logischer Operatoren kombiniert werden. Zweitens wurde die Scopus AI Search verwendet, die es ermöglicht, die Suchen in ganzen Sätzen zu übergeben.

2.1.4.1 Erweiterte Dokumentensuche

Um sicherzustellen, dass die Suche qualitativ hochwertige Ergebnisse liefert, sind effektive Suchanfragen von grundlegender Bedeutung. Zu diesem Zweck wurden zunächst detaillierte Suchbegriffe aus den Kernsuchbegriffen abgeleitet. Tabelle 2 bietet eine Übersicht über alle identifizierten Abfragebegriffe.

Tabelle 2: Suchbegriffe und Abfragekomponenten

Kategorie	Suchbegriff	Begriff der Abfrage
Overarching subject matter	Deep Learning	Deep Learning
	Machine Learning	Machine Learning
Sub-Interest within the subject matter	Data Annotation	Annotation
	Data Labeling	Label*
Concrete problem of interest	Scarcity of Data	data W/8 scarc* limited data* limited training data*
	(Un)-Availability of Data	data W/8 *avail*
Data domains of interest	Satellite data	satellite satellite data* sentinel earth observation space-born*
	Orthophotography	orthophoto*
	Aerial photography	aerial photo*
	Point clouds	point cloud*
Application domains of interest	Mobility	mobility
	Agriculture	agriculture
	Forest	forest
	Biodiversity	biodiversity
	Climate	climate

Für die Suche wurden die Suchbegriffe innerhalb einer Kategorie mit dem Operator „ODER“ kombiniert und in Klammern gruppiert, um den Vorrang zu gewährleisten. Die Unterabfragen für die Kategorien wurden dann mit dem Operator „UND“ verknüpft. Dies hat zur Folge, dass eine sehr breite Suche innerhalb der einzelnen Kategorien entsteht, die Ergebnisse jedoch auf diejenigen beschränkt werden, die allen Kategorien angehören. Daraus ergeben sich die folgenden zwei Suchanfragen:

Query 1: Umgang mit spärlich/unzureichend annotierten Trainingsdaten

Bei der ersten Suche wird eine logische Kombination aller Abfragebegriffe verwendet, die in Tabelle 2 aufgeführt sind. Die Logik wird im Folgenden dargestellt:

Logische Abfrage 1:

TITLE-ABS-KEY (("machine learning" **OR** "deep learning") **AND** (annotation OR label*) **AND** ((data W/8 *avail*) **OR** (data W/8 scarc*) **OR** ("limited data*") **OR** ("limited training data*")) **AND** (satellite **OR** "satellite data*" **OR** sentinel **OR** "earth observation" **OR** "space-born* data*" **OR** orthophoto* **OR** "aerial photo*" **OR** "point cloud*") **AND** (mobility **OR** agriculture **OR** forest **OR** biodiversity **OR** climate))

Query 2: Kosten und Aufwand für das Annotieren von Daten aus den Datendomänen

Die zweite Suche zielte auf die Kosten und den Aufwand ab, die mit der Annotation verbunden sind. Dafür wurden die Begriffe „Kosten" und „Aufwand" hinzugefügt. Des Weiteren wurden die Anwendungsdomäne und die übergeordnete Thematik in der Abfrage außer Acht gelassen, um einen breiteren Überblick zu erhalten. Die Logik wird im Folgenden dargestellt:

Logische Abfrage 2:

TITLE-ABS-KEY (("data annotation" OR "data labeling") **AND** (**cost OR effort**) **AND** (satellite **OR** "satellite data*" **OR** sentinel **OR** "earth observation" **OR** "space-born* data*" **OR** orthophoto* **OR** "aerial photo*" **OR** "point cloud*"))

2.1.4.2 KI-Suche

Zwei weitere Recherchen wurden mit der von Scopus bereitgestellten KI-Suche durchgeführt. Zu diesem Zweck wurden die oben bereitgestellten Suchanfragen in vollständige Sätze umformuliert.

- **AI-Query 1:** "The effort or cost involved in annotating data for machine learning. Focus on environment, agriculture, forest, biodiversity, climate, etc. Data involving satellite data, aerial photography, orthophotography, point clouds, etc."
- **AI-Query 2:** "Data labeling techniques evaluated for cost and effort. Focus on earth observation images and point clouds."

Da sowohl die erweiterte Dokumentensuche als auch die KI-Suche keine zufriedenstellende Antwort auf Aussagen zur Kosteneffizienz lieferten, wurde eine zweite Iteration als Online-Recherche nach frei verfügbaren Berichten und Studien durchgeführt.

2.1.5 Definition der Ein- und Ausschlusskriterien

Nach der Durchführung der definierten Suchanfragen wurden die Ergebnisse einem ersten Review unterzogen, um die Literatur weiter zu filtern und die Anzahl der Arbeiten zu reduzieren. Um dies systematisch und standardisiert zu erreichen, wurde eine Reihe von Ein- und Ausschlusskriterien festgelegt, um den Review-Prozess zu leiten (siehe Tabelle 3). Hinsichtlich des Betrachtungszeitraums für einzubeziehende Studien wurde keine zeitliche Begrenzung definiert; gleichwohl wurden in der Regel neuere Arbeiten für relevanter als ältere bewertet.

Tabelle 3: Ein- und Ausschlusskriterien

Einschlusskriterien:	Ausschlusskriterien
▶ Anwendung des maschinellen Lernens in der Umweltforschung ▶ Herausforderungen der Datenannotation ▶ Herausforderungen bei begrenzten Trainingsdaten	▶ Artikel in Fremdsprachen (nicht Deutsch oder Englisch) ▶ Keine Verweise oder Diskussion der Herausforderungen, die mit begrenzten Trainingsdaten verbunden sind ▶ Diskutiert hauptsächlich Anwendungen oder Daten aus anderen Domänen

2.1.6 Zusammenfassung von Suchergebnissen und manuelles Screening

Alle Arbeiten, die die in Abschnitt 2.1.5 definierten Kriterien erfüllten, wurden weiter nach der Art der Veröffentlichung sortiert, das heißt, Zeitschriftenartikel, Konferenzbeiträge sowie Datensätze und Benchmarks. Tabelle 4 fasst die Ergebnisse der Suche und des anschließenden Screenings zusammen. Die Recherchen ergaben insgesamt 131 Ergebnisse. Basierend auf den definierten Ein- und Ausschlusskriterien wurden in der ersten Iteration insgesamt 69 dieser Publikationen aussortiert.

Tabelle 4: Suchergebnisse nach Publikationskategorie

Anfrage	Gesamtergebnisse	Konferenzartikel	Datensätze/Benchmarks	Journalartikel	Aussortiert
Query 1	78	7	12	18	41
Query 2	40	6	0	11	23
AI-Query 1	5	2	0	0	3
AI-Query 2	8	1	0	5	2

2.1.7 Systematische Überprüfung

Die verbliebenen 62 Publikationen wurden im Detail analysiert. Weitere 30 Publikationen wurden daraufhin aussortiert, da ein klarer Fokus auf den Umgang mit wenigen Trainingsdaten fehlte. Tabelle 5 zeigt die Struktur an, mit der die restlichen 32 Veröffentlichungen einem systematischen Review unterzogen wurden. Die Struktur hilft dabei, die wichtigsten Aussagen und Ergebnisse der einzelnen Artikel miteinander zu vergleichen und kann zum weiteren Nachschlagen verwendet werden. Tabelle 5 zeigt ein Beispiel für Mirpulatov et al. (2023). Die Übersichten zu den wichtigsten Publikationen dieser Studie sind im Anhang von Zwischenbericht 1 zu diesem Vorhaben dokumentiert (Klien, et al., unveröffentlicht)[2].

[2] auf Anfrage erhältlich

Tabelle 5: Tabelle für das systematische Review

Inhalt	Ergebnis
Titel	Pseudo-Labeling Approach for Land Cover Classification Through Remote Sensing Observations with Noisy Labels
Authors	Islombek Mirpulatov, Svetlana Illarionova, Dmitrii Shadrin, and Evgeny Burnaev.
Journal and year	IEEE Access (2023)
Abstract and motivation	The limitation of ML and DL applications with Remote Sensing data is known: the quality and amount of annotated data. In this study they propose a framework for pseudo-labeling inaccurate and low-resolution data for solving land-cover and land-usage segmentation tasks. Pseudo-labeling helps improve and enlarge training datasets by using the predictions made by an already trained model on unlabelled data to generate labels. Generally, a Random Forest model, which shows itself to be robust to label noise, is a good choice for this task. The improved training set go then through a sampling strategy to reduce data size; and is finally fed into a DL (CNN) model for classification. The study improves the recognition task on the test set (F1 score).
Main aims	Improve the recognition task.
Main methods	Pseudo-labeling, semi-supervised learning (also teacher-student framework)
Data or model-oriented solution	Both, with focus on data.
KI-model applied	Random Forest as semi-supervised model for pseudo-labeling; sampling techniques as clustering for data reduction; and CNN for the final classification.
Datasets used	National Land Cover Database (NLCD) 2019 for segmentation masks (markup resolution per pixel is 30 m).
Weakness/ Challenge	If a class is too small, the proposed pipeline may struggle with data imbalance and could potentially neglect rare classes. Computational cost is a limitation when estimating clustering centers (for reducing dataset size)
Economic Considerations	N/A
Main conclusion	Pseudo-labeling can help increase labels for training and improve final model performance.

2.2 Einbindung der Stakeholder zur Bedarfsermittlung, Potential- und Wirkungsanalyse

Die Grundlage für die Bedarfsermittlung sowie die Analyse von Potential und Wirkung von annotierten Umweltdaten und umweltrelevanten Daten stellte eine umfassende Literaturrecherche dar. Der Fokus der Analyse lag auf den Sektoren Forst- und Landwirtschaft, Biodiversität, Mobilität und Klimafolgenanpassung im urbanen Raum.

Um die Ergebnisse der Literaturrecherche zu untermauern, wurden zudem verschiedenste Stakeholder (siehe Kapitel 2.2.1) mittels speziell konzipierten Stakeholder-Workshops (siehe Kapitel 2.2.2) aktiv in die Studie eingebunden.

2.2.1 Beschreibung der Stakeholder

Eine multidisziplinäre Interessengruppe aus Wissenschaft, Technologie, Wirtschaft, Industrie und staatlichen Einrichtungen sowie ziviler Gesellschaft formten die Stakeholder des Forschungsvorhabens. Dazu zählen:

- Umweltbundesamt mit Anwendungslabor für Künstliche Intelligenz und Big Data (KI-Lab) als nachgeordnete Bundesbehörde und Ressortforschungseinrichtung
- Wissenschaft und Forschung
- Wirtschaft
- Politik und Verwaltung
- Zivilgesellschaft

Die Stakeholdergruppen sind in Zwischenbericht 2 detailliert beschrieben (Catalli, et al., unveröffentlicht).

2.2.2 Stakeholder-Workshops

Insgesamt wurden fünf virtuelle Stakeholder-Workshops im Zeitraum von Juni 2024 bis Oktober 2024 durchgeführt, um den Bedarf sowie das Potential und die Wirkung annotierter Umweltdaten bzw. umweltrelevanter Daten über die verschiedenen Sektoren hinaus zu eruieren. Zudem ermöglichten die Workshops einen intensiven Austausch zwischen den Teilnehmenden und boten die Möglichkeit, diese in den weiteren Projektverlauf, wie Community Building Events, miteinzubinden. Einen Überblick über die durchgeführten Workshops gibt Tabelle 6.

Tabelle 6: Überblick der durchgeführten Stakeholder-Workshops[3]

Datum	Titel	Anzahl der Teilnehmenden
13.06.2024	ML-Modelle und Datenannotation und deren Datenbedarf im Umweltsektor	26
08.07.2024	UBA-interner Workshop zum Thema Datenbedarf	9
27.09.2024	Potential und Wirkung annotierter Daten in der Landwirtschaft	7
30.09.2024	Potential und Wirkung annotierter Daten in der Forstwirtschaft	7
02.10.2024	Potential und Wirkung annotierter Daten im Sektor Biodiversität	8

2.2.3 Onlineumfrage zum Datenbedarf

Zur Erhebung des Bedarfes an annotierten Umweltdaten bzw. umweltrelevanten Daten wurde auf Basis eines anwendungszentrierten Ansatzes ein Fragebogen zum Thema „ML-Modelle und Datenannotation und deren Datenbedarf für KI-Anwendungen im Umweltsektor“ durch wetransform konzipiert. Der Fragebogen wurde in deutscher und englischer Sprache zur Verfügung gestellt. Als Befragungswerkzeug wurde Google Formulare (GoogleForms) als Umfrageverwaltungssoftware verwendet (Abbildung 2).

[3] Anmerkung: Die Inhalte der Stakeholder-Workshops sind in Zwischenbericht 2 detailliert.

Abbildung 2: Einstieg in den Onlinefragebogen erstellt mittels GoogleForms

Quelle: eigene Darstellung, wetransform GmbH

Der Befragungszeitraum war von Juni 2024 bis Oktober 2024. Der Auftraggeber UBA sowie der Auftragnehmer wetransform und dessen Subunternehmer Fraunhofer IGD nutzten ihre bekannten Kanäle (z.B. LinkedIn, E-Mail-Kampagne, Website) und Partner in ihrem jeweiligen Netzwerk, um eine breite Verteilung der Onlineumfrage zu erreichen[4]. Insbesondere richtete sich der Fragebogen an Akteure der öffentlichen Verwaltung, der Wissenschaft und Forschung, der Wirtschaft und Industrie sowie der Zivilgesellschaft.

2.2.3.1 Beschreibung des Fragebogens

Der Fragebogen bestand aus 6 Sektionen mit insgesamt 32 Fragen (Single und Multiple Choice, offene Fragen sowie Fragen der Likert-Skala). Aufgrund der Länge wurde auf redundante Fragen verzichtet. Eine Beschreibung der einzelnen Sektionen enthält Tabelle 7.

Tabelle 7: Beschreibung des Onlinefragebogens zum Datenbedarf

Sektion	Beschreibung
Sektion 1: Kontext des KI-Anwendungsfalls	Definition der Organisation, für die der jeweilige KI-Anwendungsfall entwickelt wurde

[4] Eine ausführliche Liste der Beiträge ist in der Kommunikationsstrategie zum Forschungsvorhaben zu finden.

Sektion	Beschreibung
Sektion 2: Beschreibung des KI-Anwendungsfalls	Detaillierte Informationen zum jeweiligen KI-Anwendungsfall, insbesondere zu dessen Kategorien (z.B. Biodiversität schützen, nachhaltige Stadtentwicklung), den erreichten Reifegrad des entwickelten KI-Modells sowie den Erwartungen an das KI-Modell.
Sektion 3: Benötigte Daten für den KI-Anwendungsfalls	Genaue Beschreibung der für die Entwicklung des KI-Anwendungsfalls benötigten Daten in Bezug auf Datentyp (z.B. Fernerkundungsdaten, Sensordaten, Fotos), der jeweiligen thematische Zuordnung (z.B. Klima- und Wetterdaten, Geodaten) und der Qualität. Zudem wurde nach dem Vorhandensein von Annotationen und deren Qualität sowie optional der verwendeten Methode für die Annotation der Daten gefragt. Ebenfalls wurden auch mögliche Probleme (z.B. Daten nicht gefunden oder zu teuer), die Angabe der Lizenz (z.B. Open Data) sowie die Beschreibung möglicher Konsequenzen abgefragt.
Sektion 4: Aufwände in Bezug auf die Daten	In dieser Sektion wurden mögliche Aufwände abgefragt, die im Zusammenhang mit den Daten standen. Dazu zählen u.a. Angabe zu den Aufwänden für das Daten finden, das Abklären der Datennutzung oder Lizenzieren der Daten, die Prüfung der Daten, die Harmonisierung der Daten sowie das Datenannotieren.
Sektion 5: Limitierungen	Fragen zu den begrenzenden Faktoren wie beispielsweise den Kosten, der Kapazität, der Datenverfügbarkeit oder der Dateninfrastruktur.
Sektion 6: Angaben zu möglichen Maßnahmen	In dieser Sektion erfolgte eine Bewertung verschiedener Maßnahmen, wie beispielsweise Bereitstellung von annotierten Daten, Datenräumen, Verbesserung der Qualität und Standardisierung, anhand einer Likert-Skala.

2.2.3.2 Limitierungen

Da im Fragebogen keine Daten zu den Ausfüllenden erhoben wurden, ist es nicht möglich, die Repräsentativität der Stichprobe über verschiedene Stakeholder-Gruppen oder entlang anderer, beispielsweise soziodemographischer Merkmale zu beurteilen. Der Fragebogen wurde so entworfen, dass auch eine rein qualitative Auswertung möglich ist.

2.3 Entwicklung des Vorgehensmodells

Die folgenden Schritte bilden das methodische Vorgehen für die Entwicklung eines Vorgehensmodells für den Umgang mit wenigen annotierten Daten in ML-Anwendungen im Umweltbereich ab:

a) Einrichtung einer Plattform zur automatisierten Modellierung und Ausführung von Datenverarbeitungsprozessen, Nutzung von OpenStack-Cloud, GPU-Cluster, geteiltem Dateisystem und einem Workflowmanagementsystem (Kapitel 8.1)

b) Entwicklung eines Entscheidungsbaums zur Auswahl von ML-Methoden basierend auf der Verfügbarkeit annotierter Daten, Optionen für Szenarien ohne gelabelte Daten und Empfehlungen für verschiedene Datenmengen (Kapitel 8.2)

c) Anwendung des Modells auf zwei Anwendungsfälle (Erkennung von PV-Anlagen und Baumarten) sowie einer Land Cover Classification, Evaluation der Methoden Pseudo-Labeling und Transfer Learning hinsichtlich Effizienz, Genauigkeit und Skalierbarkeit (Kapitel 8.3).

3 Stand der Wissenschaft und Technik

Dieses Kapitel gibt einen Überblick über die wichtigsten Erkenntnisse aus der systematischen Literaturarbeit und zusätzlicher Internetrecherchen (z. B. Plattformen auf GitHub). Zunächst wird der Schwerpunkt auf die Datenannotation gelegt, wobei deren entscheidende Rolle und die anhaltenden Herausforderungen im Zusammenhang mit unzureichend annotierten Daten hervorgehoben werden. In Anbetracht der Tatsache, dass der Mangel an Trainingsdaten nach wie vor eine erhebliche Hürde darstellt, werden im zweiten Abschnitt ML-Methoden und -Ansätze untersucht, die so konzipiert sind, dass sie mit begrenzt verfügbaren oder gar keinen annotierten Trainingsdaten effektiv arbeiten können. In den folgenden Abschnitten werden verschiedene Daten- und Annotationsplattformen detailliert beschrieben und die wirtschaftlichen Aspekte dieser Tools diskutiert.

3.1 Klassifizierung von Annotationsmethoden

Annotationsmethoden können entlang der am Prozess beteiligten Akteure in drei Hauptkategorien eingeordnet werden.:

- „Users“
- „Users and AI“ und
- „AI“

Die erste Kategorie „User“ bezieht sich auf alle Methoden, bei denen Nutzende die Daten manuell annotieren und entweder als Expertinnen*Experten (z. B. Forschende, Data Scientists) oder als Nicht-Expertinnen*Experten (z. B. Citizen Science) eingebunden werden.

Innerhalb der zweiten Kategorie „Users and AI“ wird die Datenannotation nur teilweise allein durch einen Menschen oder mithilfe einer KI-gestützte Methode ausgeführt; sie ist vielmehr eine Synergie aus beiden Ansätzen. Wenn sowohl Nutzende als auch eine KI zur Datenannotation beitragen, wird von einem „Human in the Loop“-Ansatz gesprochen, bei dem Expertinnen*Experten die Annotation einiger Datenpunkte verbessern, um die Genauigkeit des ML-Modells zu erhöhen. Der Mensch fungiert in diesem Fall als „Orakel“ für das ML-System und validiert und korrigiert die maschinelle Annotation, um die Leistung oder Entscheidungsfindung des Systems zu verbessern. Durch die Integration menschlicher Aufsicht und Eingriffe in ML-Systeme können die Zuverlässigkeit, Verantwortlichkeit und das Vertrauen in KI-Anwendungen in verschiedenen Bereichen verbessert werden.

Die letzte Kategorie „KI“ umfasst alle Annotationsmethoden, die einen KI-basierten Algorithmus als Lösung für das Problem unzureichend verfügbarer Annotationen verwenden. Im Fokus der Literaturrecherche stehen (teil-)automatisierte Methoden, um den hohen Aufwand der manuellen Datenannotation zu reduzieren. Aus diesem Grund werden im Folgenden die Kategorien „Users + AI“ sowie „AI“ genauer betrachtet. Hierzu lassen sich zunächst die beiden Kategorien in weitere Unterkategorien aufteilen:

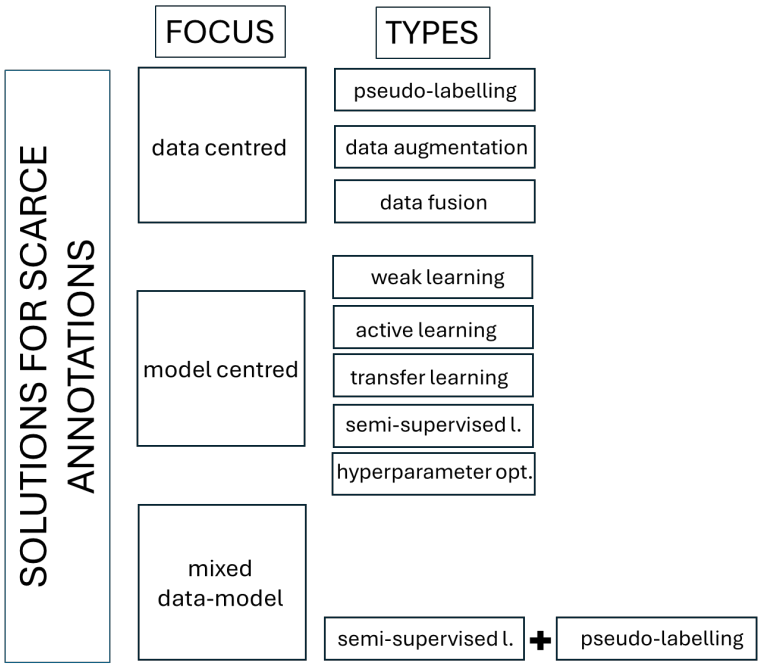
- **Datenzentrierte Methoden,** d. h. Methoden, um die Anzahl der mit Labeln versehenen Datenpunkte zu erhöhen.
- **Modellzentrierte Methoden,** bei denen es sich um KI-Algorithmen handelt, die in der Lage sind, aus unzureichend gelabelten Daten zu lernen.

- **Kombinierte Daten-Modell-Methoden,** die eine datenzentrierte Lösung mit einer modellzentrierten Lösung verbinden.

3.1.1 Datenzentrierte Methoden

Unter dem Oberbegriff der datenzentrierten Methoden wurden folgende Techniken identifiziert: (i) „**Pseudo-Labeling“** [7/22 Artikel], (ii) „**Data Augmentation**“ [1/22 Artikel] und (iii) „**Data Fusion**“ [1/22 Artikel], wie in Abbildung 3 dargestellt.

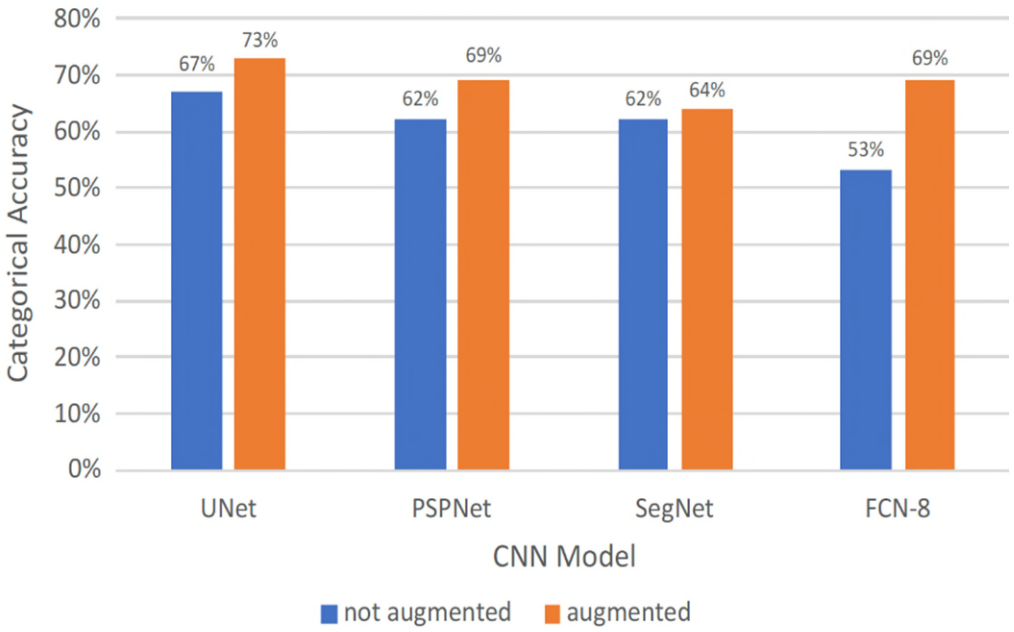
Abbildung 3: Lösungen für wenige Annotationen



Quelle: eigene Darstellung, Fraunhofer IGD

Ein sehr weit verbreiteter datenzentrierter Ansatz ist das **Datenaugmentieren**. Dabei wird die Größe und Vielfalt eines gelabelten Datensatzes künstlich erhöht, indem verschiedene Transformationen auf die vorhandenen Daten angewendet werden. Besonders bei schlecht annotierten Datensätzen kann Datenaugmentation helfen, die Robustheit und Generalisierbarkeit von Modellen zu verbessern. So haben die Studien von (Yang, et al., 2022), (Kocon, Krämer, & Würz, 2022) und (Wang & Perez, 2017) gezeigt, dass durch das Augmentieren der Trainingsdaten bessere Genauigkeiten beim Validieren erzielt werden können als ohne. Kocon et al. (2022) haben im Falle der Waldtyperkennung demonstriert, dass für alle verwendeten CNN-Modelle eine signifikante Verbesserung der Genauigkeit erreicht werden konnte, im Falle des FCN-8-Netzes sogar um 16 % (Abbildung 4). Augmentiert werden kann durch Methoden wie Hinzufügen von Rauschen, geometrische Transformationen oder Farbtransformationen. Aufgrund dieser Eigenschaften ist das Augmentieren beim Trainieren von KI-Modellen im Computer Vision Bereich mittlerweile State-of-the-Art für nahezu alle Anwendungsfälle. Daher wird es im Weiteren nicht als ein separater Ansatz betrachtet, sondern sollte immer als fester Bestandteil des Trainingsprozesses angewendet werden. Auch ist eine unbegrenzte Skalierbarkeit gegeben, da sich jedes Datum unabhängig augmentieren lässt, weshalb eine starke Parallelisierung möglich ist. Die Augmentation bietet somit eine gute Möglichkeit, die Trainingsdaten innerhalb eines bestimmten Bereichs zu erweitern und dadurch höhere Genauigkeiten zu erreichen und gleichzeitig einer Überanpassung entgegenzuwirken.

Abbildung 4: Vergleich verschiedener CNN-Modelle zur Waldtyperkennung mit wenigen Trainingsdaten und der Einfluss von Augmentation



Quelle: Kocon et al. (2022)

Pseudo-Labeling (Lee, 2013) ist eine Technik, bei der ursprünglich ungelabelte Daten genutzt werden, um die Genauigkeit des KI-Modells zu verbessern. Hierbei ist zu berücksichtigen, dass es im Laufe des Prozesses zu systematischen Fehlinterpretationen kommen kann. Beim Pseudo-Labeling-Ansatz wird zunächst ein Modell mit den vorhandenen gelabelten Daten trainiert und dieses Modell wird dann zum Labeln der ungelabelten Daten verwendet. Diese Predicts werden als Pseudo-Labels bezeichnet und zur Erweiterung des Trainingsdatensatzes für weitere Trainingsiterationen verwendet.

Wenn die vorhandenen annotierten Trainingsdaten von mittelmäßiger Qualität oder unzureichender Quantität sind, kann Pseudo-Labeling besonders wertvoll sein. Beim Pseudo-Labeling beispielsweise in Szenarien, in denen die verfügbaren visuellen Darstellungen der beobachteten Objekte in der Fernerkundung detaillierter sind als die bereitgestellten Annotationen, besteht jedoch die Notwendigkeit einer Anpassung. Dieser Anpassungsprozess kann ebenfalls als Pseudo-Labeling-Aufgabe bezeichnet werden, wie von Desai, S. &. Ghose, D. (2022) beschrieben. Der Anpassungsprozess umfasst das Aktualisieren oder Hinzufügen von Annotationen, um den Inhalt der visuellen Daten besser zu beschreiben. Während beim traditionellen Labeling Daten von Grund auf neu annotiert werden, geht es bei dieser Pseudo-Labeling-Aufgabe darum, vorhandene Annotationen auf der Grundlage zusätzlicher Informationen oder des Kontexts zu verfeinern, welche durch die visuelle Darstellung bereitgestellt werden.

Chen X. M. (2022) stellen fest, dass die Verwendung von Pseudo-Labels während des Trainings die Modellgenauigkeit verbessert, wie auch in anderen Studien gezeigt (Mirpulatov, Illarionova, Shadrin, & Burnaev, 2023; Chen, et al., 2021; Li, Zou, Lu, & Zhang, 2022). Sie kommen jedoch auch zu dem Schluss, dass die Erhöhung der Anzahl manuell gelabelter Trainingsdatenpunkte nach wie vor der effektivste Weg ist, um die Leistung eines Modells zu optimieren.

Auch in puncto Skalierbarkeit ist das Pseudo-Labeling kein Flaschenhals, da es ein iterativer Prozess ist, welcher wie jedes andere Training auch, stark parallelisiert in der GPU ausgeführt

wird. Offen bleibt allerdings die Frage, ab welcher Datenmenge dieser Algorithmus gegen ein gutes Ergebnis konvergiert. Diese Frage ist ein Kern der Evaluierung in Kapitel 8.3.3 und wird dort aufgelöst.

Data Fusion ist eine Methodik, bei der Informationen aus mehreren Quellen kombiniert werden, um die Genauigkeit der Daten zu verbessern. Es kann ein wertvoller Ansatz sein, um die Herausforderung weniger annotierter Trainingsdaten zu bewältigen, indem genauere, konsistentere und nützlichere Informationen produziert werden, als die von jeder einzelnen Quelle bereitgestellten. Dieser Ansatz kann dazu beitragen, die Inkonsistenzen und Lücken in der Annotation von Daten sowohl in zeitlicher als auch räumlicher Hinsicht zu verringern. Zum Beispiel verwenden Schiller et al. (2024) zwei verschiedene Referenzdatensätze mit unterschiedlichen Genauigkeiten, um ihr Modell zur Vorhersage des Waldsterbens in Deutschland vorzutrainieren und zu verfeinern. Beim Kombinieren von Annotationen aus mehreren Quellen können Votingverfahren verwendet werden, um die mit hoher Wahrscheinlichkeit korrekte Annotation zu bestimmen. Wenn beispielsweise in einer Klassifizierungsaufgabe drei von fünf Annotatoren einen Datenpunkt als „A“ und zwei ihn als „B“ bezeichnen, kann die Mehrheitsentscheidung als die richtige Annotation angesehen werden (Kumari, Kumar, & Mittal, 2021).

Sind keine gelabelten Daten vorhanden, kann das **Generieren von synthetischen Trainingsdaten** eine Lösung sein. Synthetische Trainingsdaten sind künstlich erzeugte Datensätze. Sie werden durch Simulationen, Datenaugmentation oder generative Modelle erstellt. Synthetische Trainingsdaten lassen sich unabhängig vom Anwendungsfall generieren. Die Literaturrecherche hat allerdings aufgezeigt, dass dieser Ansatz einige Nachteile mit sich bringt und in der wissenschaftlichen Literaturrecherche wenig vertreten war. Ein sehr großer Nachteil ist die Tatsache, dass es sich um synthetische und keine realen Daten handelt. Synthetische Daten spiegeln nicht alle komplexen Variationen und Muster realer Daten wider, was zu einer geringeren Generalisierbarkeit des Modells führen kann. Es gibt keine Garantie, dass ein Modell, das auf synthetischen Daten trainiert wurde, auch mit realen Daten gut funktioniert. Weitere problematische Aspekte sind Fehler bei der Generierung der Daten, Mangel an Diversität, aber auch die Komplexität der Generierung. Da das Generieren von synthetischen Daten ein Werkzeug zum Umgang mit nicht vorhandenen gelabelten Daten ist, wird es zwar bei der Entwicklung des Entscheidungsbaums miteinbezogen (siehe Kapitel 8.2), allerdings aufgrund der genannten Nachteile nicht mehr bei der prototypischen Umsetzung (Kapitel 8.3).

3.1.2 Modellzentrierte Methoden und kombinierte Daten-Modell-Methoden

Neben den genannten datenzentrierten Ansätzen spielen auch modellzentrierte Ansätze eine wichtige Rolle. Das Ziel der modellzentrierten Ansätze ist es nicht, die schlechte Datenlage zu verbessern, sondern durch Anpassung der KI-Modelle mit dieser Situation möglichst gut umzugehen. Dies ist ein wichtiger Aspekt, da der künstlichen Anreicherung von Daten Grenzen gesetzt sind und in einigen Fällen überhaupt keine gelabelten Daten zur Verfügung stehen. Da aber auch die Möglichkeiten modellzentrierter Ansätze begrenzt sind, kann es im Einzelfall sinnvoll sein, beide Ansätze zu kombinieren.

Sind keine gelabelten Daten vorhanden, bieten sich **Unsupervised Learning**-Verfahren (unüberwachtes Lernen) an. Unüberwachtes Lernen benötigt keine gelabelten Daten, sondern zielt darauf ab, Muster oder Strukturen in den Daten zu finden. Zu den häufigsten Methoden gehören Clustering, Dimensionalitätsreduktion und generative Modelle wie Autoencoder, Generative Adversarial Networks (GANs) und Variational Autoencoders (VAEs). (Abid, et al., 2021) verwendete einen Clustering-Ansatz, um Pseudo-Labels aus unbeschrifteten Bildern zu

erstellen und ein neuronales Netzwerk zu trainieren, um mit Sentinel-2-Bildern verbrannte Wälder ohne anfängliche annotierte Daten zu erkennen. Da unüberwachtes Lernen eine Möglichkeit ist, Muster in Daten zu erkennen, wenn keine Labels zur Verfügung stehen, wird dies beim Entscheidungsbaum berücksichtigt. Aufgrund der begrenzten Anzahl von Anwendungsfällen im Rahmen des Projekts ist das unüberwachte Lernen jedoch nicht Teil der prototypischen Umsetzung.

Beim **Semi-Supervised Learning** wird während des Trainings eine kleine Menge an gelabelten Daten mit vielen ungelabelten Daten kombiniert. Wie im vorherigen Abschnitt erwähnt, umfasst diese Methode das Trainieren eines Modells auf Basis des bereits gelabelten Teils der Daten und die anschließende Verwendung dieses Modells, um Labels für den nicht annotierten Teil vorherzusagen. Das Modell wird sowohl mit den ursprünglich gelabelten Daten als auch mit den neu gelabelten Daten aus seinen Vorhersagen (d. h. den Pseudo-Labels) neu trainiert. Dies kann die Leistung des Modells iterativ verbessern. Diese Methode verknüpft modell- und datenzentrierte Ansätze, da sie nicht nur eine Möglichkeit darstellt, die gelabelten Trainingsdaten zu erweitern, sondern sich auch auf Algorithmen konzentriert, die gut mit verrauschten oder knappen Datensätzen umgehen können, wie z.B. die Random-Forest-Modelle, wie Mirpulatov et al. (2023) hervorhebt. Eine weitere Kategorie von semi-überwachten Ansätzen stellen die graphenbasierten Methoden dar. Diese Methoden stellen Datenpunkte als Knoten in einem Graphen dar, wobei Kanten Ähnlichkeiten zwischen Datenpunkten symbolisieren. Semi-überwachte Lernalgorithmen, wie z. B. die Label-Propagation, verteilen Label-Informationen über den Graphen von beschrifteten zu unbeschrifteten Knoten und nutzen so effektiv die Struktur der Daten (Rajadell & García-Sevilla, 2013).

Active Learning ist eine Methode, bei der der Lernalgorithmus eine*n menschliche*n Annotator*in (oder ein anderes „Orakel“) abfragen kann, um neue Datenpunkte zu kennzeichnen, von denen erwartet wird, dass sie für die Verbesserung des Modells am informativsten sind. Diese Modelle sind in die Kategorien „Users and AI“ und „modellzentriert“ einzuordnen und beinhalten verschiedene Vorgehensweisen, um die aussagekräftigsten Datenpunkte zu identifizieren, wie z. B. Unsicherheitsstichproben und Abfragen durch ein sogenanntes Komitee. Im ersten Fall wählt das Modell die Datenpunkte aus, bei denen es sich in seinen Vorhersagen am wenigsten sicher ist, wie z. B. in Kölle (2023) gezeigt, wo die Autorinnen*Autoren eine Entropie-Sampling-Methode verwenden, um diese unsicheren Punkte zu identifizieren. Diese werden dann von der Annotatorin oder vom Annotator gelabelt. Beim zweiten Ansatz gibt es ein „Komitee von Modellen“, das Datenpunkte auswählt, über die sich die Modelle am meisten uneinig sind. Die Uneinigkeit zeigt an, dass der Datenpunkt informativ ist und eine Kennzeichnung wert ist (Cui, 2023).

Ein verbreiteter Ansatz beim Umgang mit wenig gelabelten Daten ist **Transfer Learning.** Dieses Verfahren bietet sich insbesondere dann an, wenn für einen ähnlichen Use Case viele annotierte Daten vorhanden sind oder sogar schon ein trainiertes Basismodell existiert. So setzt sich Transfer Learning aus zwei Schritten zusammen. Zunächst wird ein Basismodell mit gelabelten Daten eines ähnlichen Use Cases trainiert. Im zweiten Schritt erfolgt ein Fine-Tuning dieses Modells mit den (wenigen) gelabelten Daten für den gegebenen Anwendungsfall. Dieser Ansatz skaliert sehr gut, da mit einem guten Basismodell das Training nie bei null beginnt, sondern mit geringerem Aufwand für den aktuellen Use Case nachtrainiert wird und hohe Genauigkeiten erzielt werden. Aufgrund dieser positiven Eigenschaften ist Transfer Learning ein Teil der prototypischen Umsetzung, auf die im weiteren Verlauf gesondert eingegangen wird (siehe Kapitel 8.3.1.2).

Bei der **Hyperparameteroptimierung** geht es darum, die besten Hyperparameter für ein Modell zu finden, die seine Leistung auch bei begrenzt verfügbaren Daten erheblich verbessern

können und mitunter eine Verbesserung erreichen, die größer ist als bei der Anwendung semi-überwachter Lernansätze (Golyadkin, Gambashidze, Nurgaliev, & Makarov, 2024). In Kombination mit Transfer Learning wird diese Methode besonders leistungsfähig (Perera, Witharana, Manos, & Liljedahl, 2024). Dies bedeutet, dass ein vortrainiertes Modell für eine verwandte Aufgabe verwendet und für die Zielaufgabe feinabgestimmt wird. Bei diesem Ansatz kann das Wissen aus der Ausgangsaufgabe genutzt werden, wodurch die Menge der für die Zielaufgabe benötigten gelabelten Daten reduziert wird. Techniken wie die Feinabstimmung eines vortrainierten neuronalen Netzes sind gängige Beispiele. Es sollte jedoch beachtet werden, dass die Optimierung von Hyperparametern sehr zeitintensiv ist und daher parallelisiert, im besten Fall verteilt in der Cloud, erfolgen sollte.

Weak Learning beinhaltet die Verwendung unsicherer oder verrauschter Labels, die einfacher oder kostengünstiger zu erzeugen oder zu erhalten sind als genaue Labels. Methoden in dieser Kategorie zielen darauf ab, robuste Modelle zu erstellen, die effektiv aus solchen Daten lernen können. Zu ihnen gehört auch das Pseudo-Labeling, das im vorigen Abschnitt behandelt wurde.

3.2 Überblick bestehender Daten-Plattformen

Annotierte Datensätze, die entweder direkt als Grundlage für die Entwicklung neuer Modelle oder als Input für die in Abschnitt 3.1.1 vorgestellten Verfahren zur Erhöhung der Anzahl der mit Annotationen versehenen Datenpunkte dienen, stehen auf zahlreichen Daten-Plattformen zur Verfügung. Diese Daten-Plattformen bieten Zugang zu einer Vielzahl von Daten, die bereits manuell oder halbautomatisch gelabelt wurden, und ermöglichen die weitere Integration und Verfeinerung der Annotationen durch den Einsatz der in Abschnitt 3.1.1 vorgestellten datenzentrierten Verfahren (*Pseudo-Labeling, Data Augmentation, Data Fusion*). Dies umfasst zumindest Daten ohne Schutzbedarf. Für Daten mit Schutzbedarf gibt es kaum zentrale Plattformen; hier wären Datenintermediäre und Datentreuhänderstrukturen denkbare Optionen.

Zu diesen Plattformen gehören dedizierte Plattformen für Datenwissenschaftler*innen sowie diverse Plattformen für Open Data, Geodaten oder Umweltdaten. Horizontal betrachtet lassen sich die Plattformen in die folgenden Kategorien aufteilen:

- **Integrierte Plattformen**, die Daten und Modelle sowie manchmal Hosting- und Mehrwertdienste kombinieren
- **Wettbewerbsorientierte Plattformen**, in der Regel eine Variante integrierter Plattformen
- **Datenplattformen**, die sich auf das Hosting von Datensätzen konzentrieren
- **Metadaten-Plattformen**, die darauf abzielen, Daten, die an anderer Stelle gehostet werden, auffindbar zu machen
- **Verteilte Datenplattformen** wie z.B. Datenräume

Die meisten dieser Datenplattformen bieten keinen Zugriff auf die Daten über APIs (mit wenigen Ausnahmen wie z. B. replicate.com). Die Daten werden entweder online an der Quelle der Daten verarbeitet oder heruntergeladen oder an einem Ort der Wahl der Nutzenden verarbeitet.

3.2.1 Plattformen speziell für Datenwissenschaftler*innen

Tabelle 8 gibt einen Überblick über die bekanntesten dedizierten Plattformen für Datenwissenschaftler*innen.

Tabelle 8: Spezielle Plattformen für Datenwissenschaftler*innen

Name	Beschreibung
github.com	Die bekannteste integrierte Plattform für die Verwaltung von Code und zugehörigen Assets wird auch zum Hosten von Daten und Modellen sowie zugehörigen Artefakten verwendet. Github unterstützt das Bearbeiten und Ausführen von Notebooks.
huggingface.co	Eine bekannte integrierte Plattform für den Austausch von Modellen und Datensätzen, die eine große Anzahl von annotierten Datensätzen enthält. Bei den Umweltdatensätzen besteht das Hauptproblem darin, dass sie oft sehr klein sind.
paperswithcode.com	Eine Alternative zu huggingface
Kaggle.com	Eine sehr bekannte, auf Wettbewerbe ausgerichtete integrierte Plattform. Es gibt einige Umweltdatensätze, die Themen wie Klimawandel, Luft- und Wasserqualität, Energieverbrauch und Biodiversität abdecken. Ein Beispiel bildet der ISPRS Potsdam-Datensatz, der für die Modellentwicklung zur Klassifizierung der Landbedeckung verwendet werden kann.
Colab.google	Eine Modell- und Datenplattform, die Cloud-Dienste sowie einige leistungsstarke Umweltdatensätze über die Google Earth Engine integriert, aber weniger eine offene Community-Plattform ist.
Zenodo.org	Eine bekannte Daten- und Metadatenplattform, die über eine relativ große Anzahl von Umwelt- und Geodatensätzen verfügt, z.B. 126 Datensätze für „Landbedeckung“. Die meisten Assets sind gut beschrieben und verfügen über nützliche Metadaten.

3.2.2 Plattformen für Umweltdaten

Neben Plattformen speziell für Datenwissenschaftler*innen gibt es auch einige Plattformen, die darauf abzielen, den Zugang zu Open Data und Umweltdaten zu ermöglichen (Tabelle 9).

Tabelle 9: Plattformen für Umweltdaten

Name	Beschreibung
Umwelt.info	Eine Metadatenplattform, die Informationen über zugängliche Umweltdatenressourcen sammelt. Die meisten dieser verknüpften Elemente enthalten keine Anmerkungen zu Datensätzen.
INSPIRE Geoportal	Eine europäische Metadatenplattform mit umfangreichen Metadaten zu jedem Datensatz, aber ohne explizite Anmerkungen. Viele dieser Datensätze können für die Informationsfusion oder für anonymisierte Annotationen verwendet werden. https://inspire-geoportal.ec.europa.eu
Mobilithek des BMDV	Die Mobilithek ist eine zentrale Plattform für den Austausch digitaler Mobilitätsdaten, die von Anbietern, Infrastrukturbetreibern und Verkehrsbehörden bereitgestellt werden. Sie ermöglicht den Zugriff auf verschiedene verkehrspolitisch relevante Informationen, wie Fahrplandaten oder Verkehrsinformationen in Echtzeit, die zur Planung und Durchführung von Reisen genutzt werden können. Die Mobilithek bietet Start-Ups und Unternehmen die Möglichkeit, Daten auszutauschen und neue Geschäftsmodelle zu erproben, wobei viele der Daten als Open Data frei verfügbar sind. https://mobilithek.info
GovData	GovData, das zentrale Datenportal für Deutschland, bietet einen gebündelten Zugang zu Verwaltungsdaten von Bund, Ländern und Kommunen sowie zu Daten von

Name	Beschreibung
	öffentlichen Versorgungsunternehmen, Hochschulen, Forschungseinrichtungen und Forschungsförderern. Ziel ist es, diese Daten zentral verfügbar zu machen und ihre Nutzung zu erleichtern. Im Sinne von „Open Data“ setzt GovData darauf, offene Lizenzen zu fördern und das Angebot an maschinenlesbaren Rohdaten auszubauen. Zudem beinhaltet GovData auch Daten, die eingeschränkt nutzbar sind. https://www.govdata.de
Produktkatalog Open Data des Bundesamts für Kartographie und Geodäsie	Dieser Katalog bietet u.a. Geobasisdaten und 3D-Datensätze (DTM/DGM/LoD2) und enthält zusätzlich Informationen über Datenquellen, Aktualität, Genauigkeit und Anwendungsmöglichkeiten von Geodaten.
Geoportal.de	Zentrales Metadatenportal, das auch Geodaten aus den Bereichen Landwirtschaft- und Forstwirtschaft, Klima und Wetter sowie Energie und Umwelt in der GDI-DE bündelt. https://geoportal.de
Informationssystem zur agrarmeteorologischen Beratung für die Länder (ISABEL)	Kostenfreies Angebot agrarmeteorologischer Informationen des Deutschen Wetterdienstes (DWD) für die Bundesländer. Grundlage der Zusammenarbeit ist das DWD-Gesetz. In § 4 Absatz 4 ist festgelegt, dass der DWD, eine nachgeordnete Behörde des Bundesministeriums für Verkehr und digitale Infrastruktur (BMVI), die Länder bei der Durchführung ihrer Aufgaben im Bereich des Katastrophen-, Bevölkerungs- und Umweltschutzes unterstützt. Die bereitgestellten agrarmeteorologischen Informationen dienen der Förderung einer umwelt- und ressourcenschonenden Landbewirtschaftung. https://isabel.dwd.de
Global Biodiversity Information Facility (GBIF)	GBIF—das Globale Informationssystem zur biologischen Vielfalt—ist ein internationales Netzwerk und eine Dateninfrastruktur, die von den Regierungen weltweit finanziert wird. Ziel ist es, jedem und überall freien Zugang zu Daten über alle Lebensformen auf der Erde zu ermöglichen. https://www.gbif.org

3.2.3 Limitierungen

Im Ergebnis ist festzuhalten, dass ausreichend und leistungsfähige Plattformen für das Hosting von Datensätzen und Annotationen zur Verfügung stehen, die jedoch Defizite aufweisen. Zu den allgemeinen Schwächen gehören das Fehlen von Metadaten (zur Verarbeitung und Qualität der annotierten Daten) und die eingeschränkte Auffindbarkeit von Datensätzen. Darüber hinaus sind Mechanismen für die gemeinsame Nutzung sensibler Daten noch nicht installiert. Es gibt Initiativen, die darauf abzielen, sensible Daten für ML-Projekte zugänglich zu machen. Dazu gehören Gaia-X[5] und die Common European Data Spaces[6].

Die meisten dieser Initiativen, insbesondere im Zusammenhang mit Umweltdaten (z. B. der „Green Deal Data Space“), befinden sich noch in einem frühen Stadium der Einführung und werden erst zwischen 2026 und 2028 als produktionsreife Systeme verfügbar sein. Diese Initiativen befassen sich mit Themen wie Datensouveränität, Datensicherheit, Einhaltung gesetzlicher und behördlicher Anforderungen, Datenqualität, Transparenz und Vertrauen, die heute die Haupthindernisse für den Datenaustausch darstellen.

[5] https://gaia-x.eu

[6] https://digital-strategy.ec.europa.eu/en/policies/data-spaces

3.3 Annotationswerkzeuge und Annotationsplattformen

Tabelle 10 gibt einen Überblick über die wichtigsten Annotationswerkzeuge und Annotationsplattformen, die im Rahmen der Recherche identifiziert wurden.

Tabelle 10: Übersicht über relevante Annotationswerkzeuge und -plattformen

Werkzeug/Plattform	Beschreibung
Zenml	• Datenzentriertes Open-Source-Tool, das eine sehr breite Bibliothek auf Github für die Datenannotation von KI-Anwendungen bereitstellt • bietet manuelles Labeling und ein teilweise autonomes Labeling an (wenn möglich) • stellt Werkzeuge für den gesamten Workflow zur Annotation und für das Training zur Verfügung • bündelt viele einzelne Annotationsbibliotheken • Lizenz: Open-Source-Lizenz • Status: aktiv gepflegt und zweckmäßig • berücksichtigte Datentypen: Multi Modal / Multi Domain, Text, Bilder, Audio, Video, Zeitreihen, Sonstige • Link: https://github.com/zenml-io/awesome-open-data-annotation
BasicAI	• BasicAI Cloud ist eine kostenlose KI-gestützte Trainingsdatenplattform mit KI-gestützten Annotationstools und einem starken Teammanagement für schnellere, skalierbare und qualitativ hochwertige Ergebnisse. • Plattform für die manuelle Annotation von LiDAR-Daten (u.a.), aber auch für die automatische Beschriftung • Das Annotationstool verwendet KI-Unterstützung, um das Labeln zu erleichtern (z. B. Erkennen von Objektgrenzen) --> https://docs.basic.ai/docs/object-detection-tracking • Link: https://www.basic.ai/basicai-cloud-data-annotation-platform
Segments.ai	• Segments.ai bietet Software für das Labeling speziell für Bilder an. Es verwendet KI, die eine Automatisierung ermöglicht, nachdem die ersten Bilder beschriftet wurden. Es kann in verschiedenen Branchen eingesetzt werden, z. B. in der Automobilindustrie oder in der Medizin. Die Annotationsplattform von Segments.ai ermöglicht das gleichzeitige Labeling über mehrere Sensoren hinweg. Es verfügt unter anderem über alle Funktionen, die zum Labeln von Bildern und Punktwolken erforderlich sind, wie z.B. das Überlagern von 3D-Punktwolken mit 2D-Bildern • Link: https://segments.ai/
Dataloop	• Dataloop bietet eine Plattform für End-to-End-Datenmanagement, Automatisierungspipelines und ein qualitätsorientiertes Datenlabeling. • Link: https://dataloop.ai
Amazon SageMaker	• Amazon SageMaker ist eine Cloud-basierte ML-Plattform, die Entwicklerinnen*Entwicklern die Erstellung, das Training und die Bereitstellung von ML-Modellen in der Cloud ermöglicht • Link: https://docs.aws.amazon.com/sagemaker/latest/dg/data-label.html

3.4 Analysen zur Wirtschaftlichkeit

In diesem Kapitel werden die Ergebnisse der Forschung zu den Kosten der Datenannotation und des Trainings von KI-Modellen vorgestellt. Im Fokus standen dabei die volkswirtschaftlichen Kosten, aber auch die Kosten für die Umwelt. Insgesamt war ein großer Rechercheaufwand

notwendig, da dieses Thema weder in der Literatur noch bei einer anschließenden Online-Recherche stark vertreten war.

Werden zunächst die Ergebnisse von Query 1 der Literaturrecherche betrachtet, so fällt auf, dass das Thema Kosten in keiner einzigen Publikation (von 78 Treffern) vertreten war. Aus diesem Grund wurde die Query 2 definiert, die sich explizit auf das Thema Kosten (in welcher Form auch immer) bezog. Klare Aussagen waren aber auch hier nicht zu finden. Die häufigsten allgemeinen Aussagen waren, dass das Labeln und Trainieren einer KI sehr kosten- und zeitintensiv sind. Auch die anschließende Online-Recherche bestätigte, dass keine klaren Aussagen getroffen werden, insbesondere was die Kosten für das Labeln von Daten angeht. Eine Möglichkeit, klare Fakten zu erhalten, bestünde darin, Kostenanfragen bei Unternehmen zu stellen, die Datenlabeling betreiben. Das Labeln von Trainingsdaten ist jedoch sehr individuell, weshalb es schwierig wäre, diese Werte zu verallgemeinern.

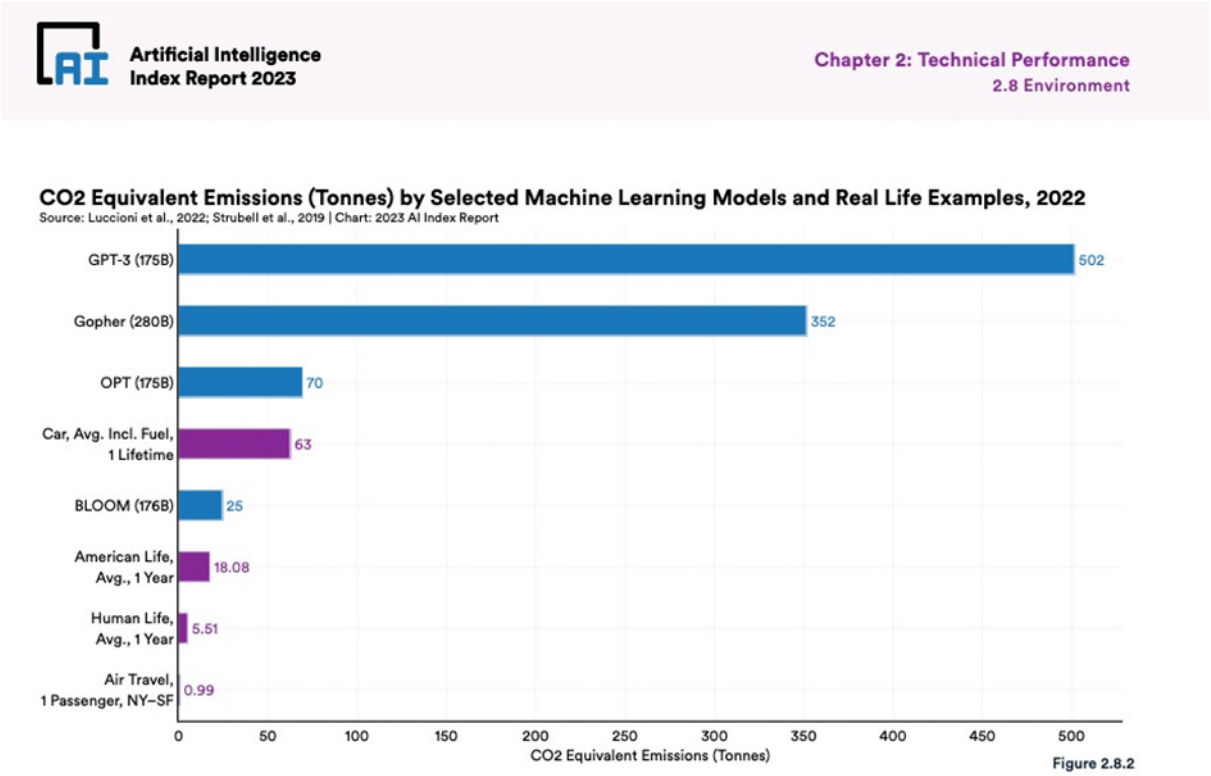
Im Hinblick auf die Kosten für das Training und den Einsatz einer KI fiel auch auf, dass das Thema Kosten auch in diesem Bereich spärlich behandelt wurde. Die wenigen Publikationen, die in dieser Richtung zu finden sind, beschäftigen sich mit dem Stromverbrauch von KI-Anwendungen. (Heguerte et al., 2023) befassten sich mit dem Vergleich verschiedener Tools zur Messung des Stromverbrauchs. Bei (LeCun, Cortes, & Burgers, 2024) wird der Stromverbrauch für das Training eines einfachen 2-Layer-Netzwerks mit dem MNIST[7] Datensatz (50MB) zur Buchstabenerkennung gemessen. Je nach Batchgröße liegt der Stromverbrauch bei bis zu 252 Wh. Entsprechend der Studie von (Luers, et al., 2024) und (de Vries, 2023) wird erwartet, dass KI-Prozessoren, die im Jahr 2023 installiert werden, jährlich 7-11 Terawattstunden (TWh) Strom verbrauchen werden, was etwa 0,04 % des weltweiten Stromverbrauchs entspricht. Allein das Training von Large Language Models (LLMs), einschließlich GPT-3, Gopher und Open Pre-trained Transformer (OPT), verbrauchte Berichten zufolge 1.287, 1.066 bzw. 324 MWh für das Training und hatte einen Energiebedarf von 564 MWh pro Tag (de Vries, 2023).

Der Forschungsbericht CO:DINA[8] verweist auf den großen CO_2-Fußabdruck des Trainings von KI-Anwendungen und stellt fest, dass der CO_2-Fußabdruck des Trainings eines einzelnen KI-Modells 284 Tonnen CO_2 entspricht, was dem Fünffachen der Lebenszeitemissionen eines durchschnittlichen Autos gleichkommt (Wurm, et al., 2022). Diese Aussage wird jedoch nicht näher erläutert, weshalb unklar ist, um welches Modell es sich beispielsweise handelt. Diese pauschale Aussage ist beispielhaft für die Art von Ergebnissen, die im Zusammenhang mit den Umweltkosten im KI-Sektor zu finden sind. Weitere Vergleiche finden sich in Abbildung 5, die dem AI-Index-Report 2023 entnommen ist.

[7] Modified National Institute of Standards and Technology

[8] https://codina-transformation.de

Abbildung 5: CO_2-äquivalente Emissionen von ausgewählten ML-Modellen und Beispielen aus der Praxis



Quelle: Masleij et al. (2023)

Abschließend sei darauf hingewiesen, dass die zuvor vorgestellten Ergebnisse bezüglich der unzureichenden Kostendarstellung von Chadli (2024, S. 16) bestätigt werden, der eine Studie über „The Environmental Cost of Engineering Machine Learning-Enabled Systems“ veröffentlichte und auch feststellte, dass „es einen bemerkenswerten Mangel an Forschung gibt, die sich auf die Umweltkosten im Zusammenhang mit Datenmanagement, Tests, Bereitstellung und Schlussfolgerung beziehen. [...] Von 52 ausgewählten Studien haben 5 Arbeiten [...] Metriken eingeführt oder erweitert, die darauf abzielen, den CO_2-Fußabdruck über eine oder mehrere ML-Pipelines hinweg zu verfolgen oder zu berechnen.“

Mit Blick auf die ökonomische Kosten-Nutzen-Rechnung für den Einsatz von ML im Umweltbereich zeigt sich, dass ein dringender Bedarf an gezielter Forschung besteht, um die verschiedenen Aspekte der Nachhaltigkeit und Wirtschaftlichkeit weiter zu beleuchten. Es bedarf Studien, die Benchmarks entwickeln und so umfassende Kosten-Nutzen-Analysen ermöglichen. Darüber hinaus besteht Forschungs- und Studienbedarf zu wirtschaftlichen Entscheidungsmodellen für ML-Anwendungen, insbesondere im Hinblick auf den Aufwand für die Datenannotation und die Nachhaltigkeit dieser Praktiken. Diese Entscheidungsmodelle sollten den Nutzenden helfen, sich für den optimalen Ansatz zum Sammeln, Annotieren und Verwalten der Daten zu entscheiden, die zum Trainieren von ML-Modellen verwendet werden, einschließlich verschiedener Beschaffungsstrategien.

Beispiele für Annotations-/Beschaffungsstrategien:

- Intern: Eigene Mitarbeitende annotieren Daten manuell.

- Outsourcing: Für die Durchführung der Datenannotation werden externe Dienstleister oder Plattformen beauftragt.
- Crowdsourcing: Crowdsourcing-Plattformen kommen zum Einsatz, bei denen eine große Anzahl von Personen Labeling-Aufgaben übernimmt.
- Automatisierte Annotation: Es werden Modelle und Algorithmen des maschinellen Lernens zur automatischen Annotation der Daten verwendet.

Aus diesen Erkenntnissen lässt sich der Bedarf an weiterer Forschung in den folgenden Richtungen ableiten:

- Was sind die Benchmarks für Kosten-Nutzen-Analysen von ML-Anwendungen im Umweltbereich und wie können diese Benchmarks zur Bewertung von Wirtschaftlichkeit und Nachhaltigkeit herangezogen werden?
- Wie können ökonomische Entscheidungsmodelle den Prozess der Datenannotation optimieren, wobei sowohl die wirtschaftlichen Kosten als auch die Nachhaltigkeitsauswirkungen während des Annotations- und Trainingsprozesses berücksichtigt werden?
- Was sind die effektivsten wirtschaftlichen Entscheidungsmodelle für die Datenannotation in ML-Anwendungen, und wie können diese Modelle die Nachhaltigkeit der Datenannotationspraktiken verbessern?

3.5 Schlussfolgerung

Abschließend werden an dieser Stelle die zuvor beschriebenen Ergebnisse im Bezug auf die definierten Forschungsfragen kurz zusammengefasst.

1. *Welche Daten werden verwendet, um Machine-Learning-Modelle im Umweltsektor zu trainieren?*

Die Datenvielfalt ist generell sehr groß und die verwendeten Daten stark abhängig vom Anwendungsfall. Im Umweltsektor sind Fernerkundungsdaten sehr dominant, weshalb die Recherche auf diesen Bereich eingegrenzt wurde. Häufig in der Recherche anzutreffende Daten waren Satellitenbilder, Orthophotos und/oder 3D-Punktwolken. Diese Daten stehen in großer Menge zur Verfügung. Für das Training von (überwachten) Machine-Learning-Modellen bedarf es allerdings zusätzlich gelabelter Daten. Die Recherche bestätigte, dass die Verfügbarkeit von gelabelten Daten sehr limitiert ist. Zwar stehen für „einfache“ Aufgaben gelabelte Daten frei zur Verfügung. Bei komplexeren Anwendungen wurden die Daten von den Entwicklern*Entwicklerinnen selbst gelabelt und häufig nicht öffentlich bereitgestellt.

2. *Welche Methoden gibt es zum Annotieren von Daten?*

Daten werden in der Regel manuell annotiert. Automatisierte Verfahren, die das manuelle Annotieren direkt ersetzen, wurden nicht gefunden. Eine Möglichkeit für den Fall, dass keine annotierten Daten für das Training zur Verfügung stehen, ist das Generieren von synthetischen Daten. Ein zu beachtender Aspekt dabei ist, dass dies ebenfalls rechenintensiv und zusätzlich fehleranfällig ist. Deshalb ist dieser Ansatz mit Vorsicht zu betrachten. Da die Verfügbarkeit und das Erzeugen von gelabelten Daten als ein großer Flaschenhals bei der Entwicklung von KI-Anwendungen identifiziert wurde, wurde der Schwerpunkt der weiteren Arbeit auf den Umgang mit einer limitierten Menge an Trainingsdaten gelegt.

*3. Befassen sich die Autorinnen*Autoren mit dem Kostenfaktor (Zeit, Aufwand), der sich auf die in ihren Arbeiten vorgestellten Annotationsmethoden bezieht?*

Auf das Labeln der Daten wird wenig im Detail eingegangen. Die Limitierung an gelabelten Daten und die hohen Kosten für das manuelle Labeln von neuen Daten ist hingegen ein allgegenwärtiges Problem und wird in vielen Publikationen angesprochen. Ebenfalls wird häufig darauf hingewiesen, dass das Training und die Ausführung von KI-Anwendungen sehr kostenintensiv sind. Detaillierte Studien zu diesem Thema waren nicht verfügbar.

4. Inwiefern sind ausreichend annotierte Daten in Formaten zugänglich, die sich für das Training von Machine-Learning-Modellen eignen?

Wie bereits zuvor erläutert, sind nur wenige annotierte Daten für das Training von KI-Anwendungen frei zugänglich. Für das Trainieren und Evaluieren von „einfachen" Aufgaben sind geeignete Daten vorhanden, für komplexere Anwendungen allerdings nicht.

5. Welche Machine-Learning-Ansätze/Modelle kommen zum Einsatz?

Diese Frage lässt sich schwierig verallgemeinern. Die Auswahl eines Machine Learning-Ansatzes ist stark abhängig von verschiedenen Dingen, wie beispielsweise dem Anwendungsfall oder der Menge der zur Verfügung stehenden annotierten Daten. Häufig verwendet werden klassische Deep Learning-Ansätze, wie Neuronale Netze. Für den Umgang mit wenig Trainingsdaten wurden klassische Trainingsverfahren weiterentwickelt. Hierzu zählen beispielsweise Verfahren, wie das Active Learning oder Pseudo Labeling. Um die Auswahl eines passenden Verfahrens zu vereinfachen, wurde im Nachfolgenden ein Entscheidungsbaum entwickelt (siehe Kapitel 8.2).

6. Welche Modelle in der Umweltforschung können gut mit wenigen Trainingsdaten arbeiten?

Unsupervised Learning-Verfahren sind am besten geeignet für den Umgang mit wenigen annotierten Daten, da diese Verfahren gar keine annotierten Daten brauchen. Allerdings ist es mit diesem Verfahren nicht möglich, komplexe Muster und Sachverhalte zu erlernen. Im Deep Learning-Bereich wurden mit UNets gute Ergebnisse mit weniger Daten erzielt. Ansätze wie das Hyperparametertuning und Datenaugmentation verbessern zusätzlich die Ergebnisse. Aufgrund des Mangels an annotierten Daten wurde das klassische Training weiterentwickelt, speziell für den Umgang mit wenigen Daten. Verfahren wie Active Learning, Pseudo Labeling oder Transfer Learning zählen zu den bekanntesten.

4 Bedarfsermittlung

Ziel der Bedarfsermittlung war es, den aktuellen Stand des fachlichen Bedarfes an Umweltdaten und umweltrelevanten Daten mit Fokus auf annotierte Datensätze zu analysieren. Dabei wurden insbesondere die Bedarfe aus den Bereichen öffentliche Verwaltung, Forschung, Wirtschaft/Industrie sowie Zivilgesellschaft berücksichtigt. Der Fokus lag auf den Sektoren Biodiversität, Land- und Forstwirtschaft, Mobilität und Klimaanpassung im urbanen Raum.

4.1 Anwendung von Machine Learning im Umweltsektor

Das folgende Kapitel soll einen Überblick über die zahlreichen und vielfältigen Anwendungen von ML-Ansätzen in den Sektoren Biodiversität, Forst- und Landwirtschaft, Mobilität und Klimafolgenanpassung im urbanen Raum geben. Der Schwerpunkt liegt auf der Erhebung der Bereitschaft sowie des aktuellen Stands der Einbindung von ML-Ansätzen im jeweiligen Sektor. Details zu den jeweiligen Sektoren sind in den folgenden Kapiteln aufgeführt.

4.1.1 Biodiversität

Durch den Globalen Biodiversitätsrahmen von Kunming-Montreal[9] (Kunming-Montreal Global Biodiversity Framework (GBF)), welcher von den Vertragsstaaten auf der 15. Conference of the Parties (COP) im Dezember 2022 geschlossen wurde, rückte einmal mehr das Thema Biodiversität in den internationalen Fokus. Der Rahmenvertrag umfasst vier übergeordnete Zielsetzungen (sogenannte Statusziele) und 23 untergeordnete Zielvorgaben (Handlungsziele). Entscheidend für das GBF ist ein Monitoringsystem, das sicherstellen soll, dass der Zustand der Natur überwacht und über die Fortschritte bei der Verwirklichung der Ziele des GBF berichtet wird. Für dieses Monitoring können KI-Methoden und die Anwendung von ML-Algorithmen einen erheblichen Beitrag leisten.

Seit 2004 ist ein starker Anstieg der Veröffentlichungen im Bereich „KI und Biodiversität“ und seit 2016/17 sogar ein exponentieller Anstieg mit mehr als 5700 Veröffentlichungen, die sich mit KI für die biologische Vielfalt befassen (AI for Good, 2023), zu verzeichnen. Dabei sind die Anwendungsbereiche von ML-Ansätzen in der Biodiversität sehr facettenreich. Sie reichen von der Artenerkennung und -klassifikation (Chen, et al., 2024), über die Erkennung von Ursachen für den Verlust von Biodiversität (Manca, Benedetti-Cecchi, Bradshaw, & et al., 2024; Branco, Correia, & Cardoso, 2023), der Ausweisung von Gebieten für die Wiederherstellung von Ökosystemen (Sahana, Areendran, & Sajjad, 2022), zum Monitoring von invasiven Arten und Krankheitsüberwachung (Seidl, Klonner, Rammer, & et al., 2018) sowie den Wildtierschutz (Tuia, Kellenberger, Beery, & et al., 2022). Insbesondere Real-Time-Methoden, wie die Analyse von Filmmaterial, können dazu beitragen, gefährdete Arten schnell vor akuten Bedrohungen (z.B. Wilderei und Bränden) zu schützen (Thompson, 2023).

Einen direkten Bezug zu geografischen Daten hat dabei die Habitatüberwachung und -modellierung (Ramírez-Delgado, Di Marco, Watson, & et al., 2022). Diese Anwendungen ermöglichen eine schnellere und präzisere Datenerfassung, den Schutz gefährdeter Arten und die proaktive Bewältigung von Umweltauswirkungen. Beispiele ausgewählter Anwendungen sind in Tabelle 11 dargestellt.

[9] https://www.cbd.int/doc/decisions/cop-15/cop-15-dec-04-en.pdf

Tabelle 11: Überblick ausgewählter Beispiele von ML-Ansätzen im Sektor Biodiversität

Anwendungsbereich	Name	Beschreibung
Populationsdynamik und Artenverbreitung	eBird	eBird ist ein weltweit agierendes Citizens Science Projekt. Die Daten werden von Freiwilligen über die eBird Website gesammelt. Das eBird-Wissenschaftsteam nutzt statistische Modelle und ML, um beispielsweise Migration, Bestandsmuster und Verbreitungsgrenzen zu identifizieren. Rohdaten zu Vogelbeobachtungen werden mit hochauflösenden Satellitenbildern von National Aeronautics and Space Administration (NASA), National Oceanic and Atmospheric Administration (NOAA) und United States Geological Survey (USGS) statistisch ausgewertet, um Populationstrends vorherzusagen und zu ermitteln. Weitere Infos: https://ebird.org/home
Arterkennung und -klassifikation	Flora Incognita	Flora Incognita ermöglicht es den Nutzenden, Pflanzenarten interaktiv zu identifizieren und ihre Beobachtungen zu erfassen. Speziell entwickelte Deep-Learning-Algorithmen, die auf einem umfangreichen Fundus von Pflanzenbeobachtungen trainiert wurden, klassifizieren Pflanzenbilder mit hoher Genauigkeit. Durch den Einsatz dieser Technologie in einem kontextadaptiven und interaktiven Identifizierungsprozess sind die Nutzenden nun in der Lage, Pflanzen unabhängig von ihrem botanischen Wissensstand zuverlässig zu identifizieren. Weitere Infos: https://floraincognita.de
Arterkennung und -klassifikation	KInsecta	Anwendung von KI-Methoden und ML für die automatisierte Artbestimmung von Insekten. Ein Ziel ist, Expertinnen*Experten beim Annotieren von Insektenbildern zu unterstützen, indem ML-Verfahren Bilder automatisiert annotieren und diese Annotationen den Expertinnen*Experten erklärt werden – so können etwa falsche Annotationen schneller als solche erkannt werden. Weitere Infos: https://kinsecta.org
Arterkennung und -klassifikation	BUND Insekten Kosmos	Die App verwendet ML-Algorithmen zur Analyse von akustischen Daten und Bildern, die von automatischen Aufzeichnungssystemen gesammelt wurden, um Insektenarten zu identifizieren und ihre Häufigkeit zu überwachen. Weitere Infos: https://www.bund-niedersachsen.de/mitmachen/bund-insekten-kosmos/

4.1.2 Land- und Forstwirtschaft

Die Analyse von Umweltdaten mit ML im Agrar- und Forstsektor bietet die Möglichkeit, die Effizienz, Nachhaltigkeit und Widerstandsfähigkeit der land- und forstwirtschaftlichen Praktiken zu verbessern und gleichzeitig die Erträge zu steigern. Dies kann u.a. dazu beitragen, die Herausforderungen in Bezug auf die Nahrungsmittelproduktion im Kontext des Klimawandels und der steigenden Weltbevölkerung besser zu bewältigen, aber auch zu einer klimaangepassten nachhaltigen Waldbewirtschaftung beisteuern.

4.1.2.1 ML in der Landwirtschaft

Der landwirtschaftliche Sektor befindet sich aktuell in der vierten Revolution, verursacht durch die zunehmende Verwendung von Informations- und Kommunikationstechnologie (IKT) (Walter, Finger, Huber, & Buchmann, 2017). Landwirtschaft 4.0 oder intelligente Landwirtschaft (Smart Agriculture) profitiert enorm von Big Data und der raschen Entwicklung digitaler Technologien. Dabei kommt ML-Anwendungen ein besonderer Stellenwert zu. ML-Algorithmen

werden in allen Bereichen der landwirtschaftlichen Produktion verwendet und haben in Bezug auf Effizienz, Nachhaltigkeit und Produktivität zu signifikanten Fortschritten geführt. Insbesondere können ML-Ansätze dabei helfen, den Einsatz von Chemikalien wie Pestiziden und Düngemitteln zu reduzieren, und führen damit zur Reduzierung des ökologischen Fußabdruckes landwirtschaftlicher Anbaupraktiken. Sie tragen somit zur Förderung der nachhaltigen Landwirtschaft bei und haben einen direkten Einfluss auf SDG 2: Zero Hunger (Bachmann, Tripathi, Brunner, & Jodlbauer, 2022). Tabelle 12 gibt einen Überblick der vielfältigen Anwendungsbereiche von ML im landwirtschaftlichen Sektor.

Tabelle 12: Zentrale Anwendungsbereiche von ML-Ansätzen im Sektor Landwirtschaft

Anwendungsbereich	Beschreibung	Literatur
Präzisions-landwirtschaft (Precision Farming)	Unter Präzisionslandwirtschaft versteht man die Anwendung von Technologien zur Schaffung eines datenzentrierten Ansatzes in der Landwirtschaft. Dabei werden verschiedene Methoden und Technologien wie Sensoren, Drohnen und Satellitenfernerkundung eingesetzt, um Echtzeitdaten über den Gesundheitszustand der Pflanzen, die Bodenbedingungen, Wettermuster und vieles mehr zu sammeln. ML-Methoden werden genutzt, um diese Daten zu verarbeiten und fundierte datenbasierte Entscheidungen zum Management zu treffen. Deutschland und die Tschechische Republik sind in Mitteleuropa führend beim Einsatz von Technologien für die Präzisionslandwirtschaft (Petrović, et al., 2024).	(Ahman, 2024)
Ertragsprognose und -optimierung	Die Vorhersage von Ernteerträgen ist eine wichtige Aufgabe für die Entscheidungsträger*innen auf nationaler und regionaler Ebene. Ein genaues Modell zur Vorhersage der Ernteerträge kann Landwirtinnen*Landwirten helfen, ihre Anbaustrategien anzupassen und somit Ressourcen effizienter einzusetzen.	(Klompenburg, Kassahun, & Catal, 2020)
Wassermanagement	Basierend auf Daten zu beispielsweise Bodentopografie, Bodenfeuchtigkeit und Wettervorhersagen, werden ML-Algorithmen eingesetzt, um die Bewässerung präzise zu steuern sowie Strategien zum effizienten Wassermanagement zu entwickeln.	(Umutoni & Samadi, 2024)
Schädlings- und Krankheits-management	ML-Algorithmen werden für die Früherkennung von Schädlingen und Krankheiten sowie die Prognose von möglichem Schädlingsbefall eingesetzt. Dadurch können ML-Methoden Landwirtinnen*Landwirte dabei unterstützen, den Einsatz von Pestiziden und Chemikalien zu verringern und gleichzeitig die Qualität und die Produktion ihrer Pflanzen zu erhalten und zu verbessern.	(Domingues, Brandão, & Ferreira, 2022)
Bodenanalyse und Fruchtbarkeits-management	ML-Ansätze können dabei helfen, spezifische Empfehlungen für den Einsatz von Düngemitteln zu geben. Dadurch kann die Anwendung effizienter und nachhaltiger gestaltet werden.	(Sujatha & Jaidhar, 2024)

4.1.2.2 ML in der Forstwirtschaft

Auch in der Forst- und Holzwirtschaft sind ML-Ansätze verbreitet. Deep Learning-Methoden werden beispielweise zur Baumartenklassifizierung basierend auf Fernerkundungsdaten

(Sylvain, Drolet, Thiffault, & Anctil, 2024) oder für die Waldinventur und Biomasseschätzung mit Satelliten- und Drohnenbildern sowie Lidar angewandt (Mo, Zohner, Reich, & et al., 2023). Im Waldschutz können ML-Ansätze bei der Krankheitserkennung und Schädlingsüberwachung mittels automatisierter Bildverarbeitung sowie prädiktive Analysen anhand historischer Daten unterstützen (Forzieri, Girardello, Ceccherini, & et al., 2021; Rammer & Seidel, 2019).

Die Optimierung des Ernteprozesses und Entscheidungsunterstützung für nachhaltige Bewirtschaftungsstrategien können ebenfalls von ML-Ansätzen profitieren, indem diese eine Entscheidungsgrundlage liefern. Auch für die Bestimmung von Holzart und Holzqualität können ML-Methoden verwendet werden (Nieradzik, et al., 2024). ML-Algorithmen werden zudem für die Verbesserung der Vorhersage des Waldbrandrisikos angewandt. Durch Verknüpfung unterschiedlicher Datenquellen (z.B. Messdaten zu Wetterbedingungen sowie Satelliten- und Drohnenbilder) können Merkmale wie trockene Vegetation in hoher Auflösung erkannt werden. Diese identifizierten Muster dienen als Trainingsdaten für ML-Algorithmen, die auf umfangreiche und aktuelle Datensätze angewendet werden, um Waldbrandprognosen zu erstellen und das Ausmaß von Waldbränden zu simulieren (BAM, 2022). Im Folgenden sind einige ausgewählte Beispiele gelistet, welche die Anwendung der wissenschaftlichen Methoden mit der Praxis verbinden (Tabelle 13).

Tabelle 13: Ausgewählte Beispiele von ML-Ansätzen im Sektor Forstwirtschaft

Anwendungs-bereich	Name	Beschreibung
Monitoring von Entwaldung	Global Forest Watch (GFW)	Mit Hilfe von überwachten ML-Methoden hat GFW den ersten Prototyp zur automatischen Erkennung von Palmenplantagen anhand von Satellitenbildern entwickelt. Dafür würden mehr als 3.000 Satellitenbilder manuell annotiert. Weitere Infos: https://www.globalforestwatch.org
Krankheits-erkennung und Schädlingsüber-wachung	Projekte BeechSAT und IpsSAT	Im Projekt BeechSAT und IpsSAT wurden unterschiedliche ML-Methoden (Random Forest und U-Net) zur automatisierten Detektion geschädigter Bäume unter Einbeziehung satellitengestützter Fernerkundung und Standortfaktoren im Bayrischen Wald getestet, um Absterbeerscheinungen an Buchen (BeechSAT)[10] bzw. Erfassung und Beobachtung von Borkenkäferschäden an Fichten (IpsSAT)[11] zu untersuchen.
Bewässerung	Quantified Trees (QTrees)	QTrees ist ein vom BMUV gefördertes Forschungsprojekt, das von der Technologiestiftung Berlin koordiniert und gemeinsam mit der Birds on Mars GmbH und dem Grünflächenamt Berlin Mitte durchgeführt wird. Es zielt darauf ab, die Auswirkungen des Klimawandels auf Stadtbäume abzumildern, indem es ein KI-basiertes Prognosemodell dafür bereitstellt, wann genau ein Stadtbaum bewässert werden muss. Mithilfe von auf ML basierenden Vorhersagealgorithmen berücksichtigt QTrees eine Reihe von Indikatoren für den Gesundheitszustand von Bäumen, darunter Baumart und -alter, Bewässerungsdaten, Wetterdaten, Schattenindexdaten, Crowdsourced-Meldungen über Baumschäden und Daten von Bodenspannungssensoren, um vorherzusagen, welche Bäume am meisten gefährdet sind und zusätzliche Pflege benötigen, um gesund zu bleiben. Alle Informationen rund um einen einzelnen Baum werden schließlich in der App

[10] https://www.lwf.bayern.de/informationstechnologie/fernerkundung/269436/index.php

[11] https://www.lwf.bayern.de/informationstechnologie/fernerkundung/270077/index.php

Anwendungs-bereich	Name	Beschreibung
		„Baumblick“ angezeigt. Anhand der Vorhersagen, die die Modelle liefern, können öffentliche Verwaltungen und zivilgesellschaftliche Gruppen ihre Maßnahmen gezielt auf die Bäume ausrichten, die sie am dringendsten benötigen. Weitere Infos: https://www.qtrees.ai

4.1.3 Mobilität

Mobilität ist eines der meistdiskutierten Themen. Zunehmender Verkehr, hohe Verstädterung, starker verkehrsbedingter Lärm und enorme Umweltbelastung in Transitbereichen müssen nachweislich reduziert werden. KI und ML-Algorithmen stellen dabei eine entscheidende Technologie dar, um die Art und Weise, wie wir uns fortbewegen nachhaltiger, sicherer und effizienter zu gestalten. Insbesondere kann die ML-basierte Auswertung von Umweltdaten diesen Prozess unterstützen und einen wesentlichen Beitrag zu Aspekten der Effizienzsteigerung bis zur Verringerung der Umweltauswirkungen leisten (Boll-Westermann, 2020). Die Anwendungsbereiche sind dabei facettenreich und reichen von der Optimierung des Verkehrsflusses (z.B. Stauvorhersage und Umleitung) zur Reduzierung von Emissionen (Köller, 2022), über nachhaltige Stadtplanung und Verkehrsmanagement (umweltbewusste Routenplanung, Integration von ÖPNV und E-Mobilität, Energieeffizienz und Emissionsreduktion bei Elektrofahrzeugen (Batteriemanagement, Ladeinfrastruktur-Optimierung) (Jacobs, 2020), Luftqualitätsüberwachung und -prognose (MVKU, 2023), Adaptive Geschwindigkeitsbegrenzungen (Al-Dweik, Mayhew, Muresan, Shirook, & Shami, 2017; Kumar, Tiwari, Tewatia, Jain, & Popli, 2023) bis hin zum Risikomanagement von Umweltkatastrophen.

4.1.4 Klimafolgenanpassungen um urbanen Raum

Die Analyse von Umweltdaten mithilfe von ML bietet Städten die Möglichkeit, präzise Vorhersagen über zukünftige Klimaveränderungen zu treffen. Zu den wichtigsten Anwendungsbereichen gehören dabei:

- Klimamodellierung und -vorhersage
- Klimawandelanalysen und Anpassung
- Vorhersage von Extremwetter und Anomalieerkennung
- CO_2-Emissionsüberwachung und -vorhersage
- Urbanes Klima und Smart Cities (z.B. Urban Heat Island Effects**)**

ML-Ansätze ermöglichen eine bessere Vorbereitung auf Extremwetterereignisse und die Entwicklung von Frühwarnsystemen. Gleichzeitig kann ML dazu beitragen, Ressourceneinsatz und Infrastrukturplanung zu optimieren, indem es basierend auf Umweltdaten Maßnahmen zur Luftqualitätsverbesserung und Wasserversorgungsoptimierung vorschlägt. Zusätzlich unterstützt es die Stadtplanung und -entwicklung durch die vereinfachte Integration von Klimaanpassungsmaßnahmen in bestehende Infrastrukturen und Gebäude. Diese ganzheitliche Herangehensweise verbessert die Widerstandsfähigkeit städtischer Gebiete gegenüber den Auswirkungen des Klimawandels.

Tabelle 14: Ausgewählte Beispiele von ML-Ansätzen für Klimafolgenanpassung im urbanen Raum

Anwendungsbereich	Beschreibung	Beispielstudie
Urban Heat Island Effect	Urbane Gebiete wachsen weltweit und werden immer dichter. Mit diesem Wachstum geht eine verstärkte lokale Erwärmung einher, die als städtischer Wärmeinseleffekt (Urban Heat Island Effect) bekannt ist. ML-Algorithmen können dabei helfen, Wärme- und Kohlenstoffprozesse in der städtischen Umwelt zu simulieren und somit eine Entscheidungsgrundlage für Stadtplaner bereitstellen, um der raschen Erwärmung in den Städten entgegenzuwirken.	(Peiyuan, Tianfang, Shiqi, & Zhi-Hua, 2022)
Luftqualitäts-verbesserung	ML-Ansätze bieten innovative Lösungen zur Überwachung, Analyse und Vorhersage der Luftverschmutzung. Insbesondere sind sie in der Lage, das breite Spektrum an verfügbaren Daten (z.B. Messungen durch Sensoren, Fernerkundungsdaten, Messungen via Drohne) effektiv zu analysieren und neue Erkenntnisse hervorzubringen, welche Entscheidungsprozesse beeinflussen. Beispiele für die Anwendung sind Echtzeit-Vorhersagen von Verschmutzungsmustern, die Implementierung von Frühwarnsystemen und die Detektion von Verschmutzungs-Hotspots.	(Natarajan, Shanmurthy, Arockiam, & et al., 2024)
Überwachung der Lärmbelastung	Mit dem Reporting für die europäische Umgebungslärm-Direktive (Environmental Noise Directive[12], END) liegen umfangreiche Daten zur Lärmbelastung, deren Quellen und Auswirkungen vor. Diese können verwendet werden, um zukünftige Entwicklungen mittels ML-Ansätzen flächendeckend und schnell zu berechnen. Auf diese Weise kann die Lärmaktionsplanung aktiver gestaltet werden, anstelle primär reaktiv geplant zu werden.	(Carrasco, et al., 2023)

4.2 Thematischer Bedarf

4.2.1 Allgemeine Betrachtung

Die Verfügbarkeit und Nutzung qualitativ hochwertiger Daten spielt eine entscheidende Rolle bei der Bewältigung globaler Umweltprobleme, wie dem Klimawandel, dem Artenverlust und der Umweltverschmutzung. Im Umweltsektor wächst der Bedarf an präzisen, annotierten Daten, um fundierte Entscheidungen treffen zu können, die sowohl auf lokaler als auch auf globaler Ebene wirksam sind. Jedoch ist der Zugang zu diesen meist begrenzt.

In diesem Zusammenhang verweist das Helmholtz Zentrum für Umweltforschung GmbH (UFZ) in seiner Stellungnahme zum Forschungsdatengesetz (FDG) explizit auf „bessere Zugangsmöglichkeiten zu Daten von Behörden und Ämtern (z.B. Forstämtern, Katasterämtern, Meldeämter) sowie Ressortforschungseinrichtungen“ (UFZ, 2023). Auch das GeoForschungsZentrum (GFZ) verweist in seiner Stellungnahme auf „Daten von Behörden, Ämtern und Verbänden“, ohne diese jedoch weiter zu spezifizieren (GFZ & FID GEO, 2023). Basierend auf einer umfangreichen Analyse der Stellungnahmen zum FDG besteht für die Daten aus Tabelle 15 der Bedarf für einen besseren Zugang.

[12] https://environment.ec.europa.eu/topics/noise/environmental-noise-directive_en

Tabelle 15: Bedarf für bessere Zugangsmöglichkeiten

Sektor	Daten
Forst- und Landwirtschaft	Ökologische Daten der Forst- und Agrarämter und Behörden topographische Informationen, aber auch Bodenbeiwerte im landwirtschaftlichen Bereich Einsatz von Stoffeinträgen, wie Düngemitteln und Pestiziden in der Landwirtschaft, zur Landnutzung und Landnutzungsänderung (z.B. INVEKOS) sowie Daten von privaten Landeigentümerinnen*Landeigentümern (z.B. zur Bodennutzung)
Mobilität	Georeferenzierte Daten zu Mobilität und im Umfeld der Infrastruktur (z.B. Umweltdaten) Mobilitätsdaten zur Verkehrszählung und Verkehrsplanung Großflächige Verkehrsdaten mit hoher räumlicher und zeitlicher Auflösung Gelabelte Test- und Trainingsdaten für Autonomes Fahren Erhebungen aus Mobilitätsstudien (gelabelte) Daten aus Fahrzeugen (Sensorrohdaten, vorverarbeitete Daten, Telemetriedaten, Lastprofile) (gelabelte) Daten aus der Verkehrsinfrastruktur/öffentlichem Raum (Bilddaten von Verkehrskameras, Verkehrszähler, V2X Nutzung) Flottendaten (Touren, Auslastung, Fahrt- und Ruhezeiten) ÖPNV-Daten (Netzpläne, aktuelle Fahrt- und Standzeiten, Auslastung, Personenströme und Abhängigkeiten) Daten zu Sharingdiensten Daten zu Parkräumen und Parkraumbewirtschaftung Daten zu Mobilitätsbedarfen umfangreiche, gelabelte Daten aus echten Anwendungsfällen
Biodiversität	Daten zum Umweltschutz Kartierungsprojekte und Programme zur Erfassung der Biodiversität (z. B. aus Planungsverfahren zu Windkraftanlagen, o.ä.)
Klima	Klima- und Wetterdaten Emissionsdaten (z.B. CO_2-Emissionen) Klimarelevante Daten
Sektorübergreifend	Qualitativ hochwertige Metadaten Annotierte Daten Referenzdatensätze, um Vergleichbarkeit verschiedener Ansätze zu ermöglichen (bspw. für das Training von KI-Modellen); im Idealfall sind diese Daten gelabelt Besser auffindbarer Zugang zu digitalen Geländemodellen der Geoportale

4.2.2 Sektorspezifische Betrachtung

4.2.2.1 Landwirtschaft

Die Recherchen und die Rückmeldungen aus den Stakeholder-Workshops zeigten, dass sich im landwirtschaftlichen Sektor vor allem die Wiederverwendbarkeit annotierter Daten oft sehr schwierig gestaltet, da die Anwendungsfälle meist heterogen und spezifisch sind und es nur wenige Überlappungen in Bezug auf Daten und/oder die Methodik gibt. Eine Ausnahme bildet der Pflanzenbau, bei welchem die Methodik auf Objekterkennung von 2D-Bildern (Arterkennung, Erkennung von Krankheiten) ausgerichtet ist und es dabei größere Überschneidungen gibt. Im Gegensatz dazu sind die Ansätze im Bereich Tierwohl und Tiergesundheit sehr individuell.

Bedarf an annotierten Daten besteht beispielweise im Bereich Ökologie und Arterkennung, wo noch ein großes Potential für ML-Anwendungen erschlossen werden kann. Dieses kann jedoch nur dann genutzt werden, wenn die entsprechenden Daten zur Verfügung stehen. In diesem Zusammenhang verweist das IGD im Zuge des Stakeholder-Workshop 5 darauf, dass es aktuell nur wenige Daten von Drohnen für die Arterkennung gibt. Daher werden auch synthetische Daten generiert, um eine Vielfalt künstlich zu erzeugen, die so in der Natur nicht vorhanden ist. Zudem wird die zeitliche Mehrfachdatensammlung in Bereichen mit Trainingsdaten als sehr wertvoll angesehen.

Auch Kontextdaten und Daten zur Bewirtschaftung (z.B. wie oft mäht die*der Landwirt*in, Teilnahme an bestimmten Programmen, welche Richtlinien werden eingehalten) sind essenziell für die Entwicklung solider ML-Algorithmen. Diese Daten werden direkt von den Landwirtinnen*Landwirten gesammelt, jedoch mit Blick auf die Möglichkeiten zur Bemessung des Wertes ihres Betriebes häufig zurückgehalten (ähnlich wie in der Forstwirtschaft), und sind daher selten zugänglich.

4.2.2.2 Forstwirtschaft

Der größte Bedarf besteht in der Bereitstellung von Waldinventurdaten. Nach wie vor werden Forstinventurdaten nur begrenzt zur Verfügung gestellt, wodurch es nicht möglich ist, Algorithmen effektiv zu trainieren. Die Gründe dafür sind vielschichtig. Aus den Interviews und aus den Veranstaltungen, die im Zuge dieses Projektes, aber auch anderer Projekte (z.B. FutureForest[13]) durchgeführt wurden, können folgende Hauptgründe abgeleitet werden:

- Die Daten sind Betriebsgeheimnisse (am häufigsten genannt).
- Die Daten könnten schädlich für Förderungen sein.
- Die Daten könnten Misswirtschaft oder andere Fehler, möglicherweise mit Rechtsfolgen, sichtbar machen.
- Es handelt sich um Daten, die Dritte betreffen, und die vertraglichen Vereinbarungen lassen eine Weitergabe an Dritte nicht zu (das betrifft Verbände und Staatsforste).

Ähnlich sieht es bei der Bereitstellung von Forschungsdaten aus, welche ebenfalls nur bedingt und meist in keiner standardisierten Form zur Verfügung stehen.

4.2.2.3 Biodiversität

Um KI effektiv im Bereich des Biodiversitätsschutzes einzusetzen, muss zunächst die bestehende „Biodiversity Data Gap“ geschlossen werden. Derzeit sind Daten oft fehlend, lückenhaft und veraltet („missing, sparse and outdated“), was eine zuverlässige Analyse und Anwendung von KI erschwert (Ferraglioni, 2024). Zudem sind die existierenden Biodiversitätsdaten äußerst heterogen: Sie umfassen eine Vielzahl von Formaten und Quellen, wie Audiodateien von Vogelgesängen, DNA-Analysen, Bilder von Wildtieren, Drohnen- und Satellitendaten sowie Punktwolken (3D-Daten). Diese Vielfalt stellt hohe Anforderungen an die Interoperabilität und Integration der Daten, um sie für KI-Anwendungen nutzbar zu machen.

Darüber hinaus mangelt es an grundlegender Kenntnis über viele weltweit existierende Arten, insbesondere in schwer zugänglichen oder wenig erforschten Regionen. Dies führt zu einem erheblichen Wissensdefizit, das vor allem die gezielte Erforschung und den Schutz dieser Arten behindert. In diesem Zusammenhang machen (Hughes, et al., 2021) auf die Verzerrungsmuster

[13] https://future-forest.eu

bei der Probenahme und Beobachtung von Tieren aufmerksam. Laut den Autoren sind nur 6,74% der Erde beprobt, wobei die Probenahme in den Tropen verhältnismäßig schlecht sei. Zudem seien Hochebenen und Tiefseegebiete nahezu unerschlossen.

Die Autoren schließen ihre Studie mit folgender Forderung:

„Additional data will be needed to overcome many of these biases, but we must increasingly value data publication to bridge this gap and better represent species' distributions from more distant and inaccessible areas, and provide the necessary basis for conservation and management." (Hughes, et al., 2021, S. 1259)

Ein weiteres Problem besteht darin, dass Biodiversität lokal geprägt ist, was bedeutet, dass besonders sensible Biodiversitätshotspots besser überwacht und mit aktuelleren und genaueren Daten versorgt werden müssen.

Um ML sinnvoll für den Biodiversitätsschutz einzusetzen, ist es entscheidend, die Datenlage zu verbessern. Es müssen standardisierte Protokolle zur Dateninteroperabilität entwickelt werden, um die Vielfalt der Datenquellen effizient zu integrieren. Zusätzlich ist eine regelmäßige und umfassende Aktualisierung der Daten notwendig, um auf den schnellen Verlust von Arten und Lebensräumen reagieren zu können. ML-Verfahren können schließlich nur dann ihre volle Wirkung entfalten, wenn die zugrunde liegenden Daten repräsentativ, aktuell und vollständig sind.

4.3 Technische Anforderungen an die Bereitstellung der Daten

Basierend auf den Auswertungen des Literaturreviews, den Stellungnahmen zum FDG, den Stakeholder-Workshops sowie den Rückmeldungen zur Onlinebefragung wurden die folgenden Anforderungen an Umweltdaten und umweltrelevante Daten definiert. Wo möglich, wurde versucht, auch die Anforderungen an annotierte Daten zu berücksichtigen.

4.3.1 Bereitstellung der Daten nach den FAIR-Prinzipien

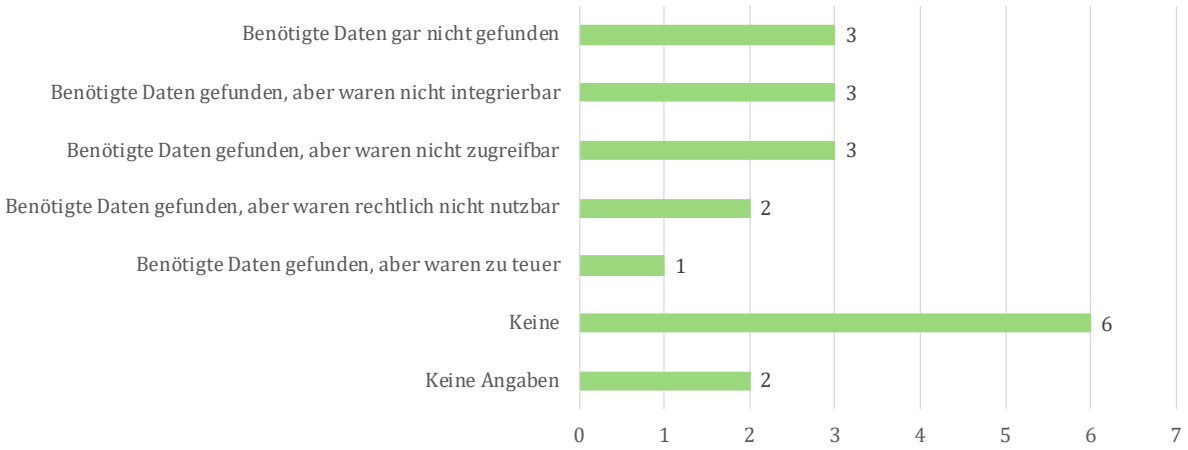
Als grundlegende und zentrale Anforderung an Umweltdaten bzw. umweltrelevante Daten wurde die Bereitstellung dieser nach den FAIR-Prinzipien genannt.

4.3.1.1 Leichtere Auffindbarkeit von Daten (Findability)

Abgesehen von den Zugangsmöglichkeiten, sollten die Daten auch auffindbar sein. Aktuell führt die Verteilung der Daten auf verschiedene Repositorien zu einer Unübersichtlichkeit des Datenangebotes. Diese Situation ist nicht zuletzt eine Folge des Föderalismus und der unterschiedlichen Zuständigkeiten von Bund, Ländern und Kommunen. In der im Rahmen des Vorhabens durchgeführten Onlinebefragung wurde die Auffindbarkeit der Datensätze als ein Problem benannt (Abbildung 6). Rückmeldungen aus den Stakeholder-Workshops unterstützen diesen Sachverhalt und auch innerhalb der Stellungnahmen zum FDG wurde dieser Punkt angesprochen (FZI, 2023; GFZ & FID GEO, 2023).

Abbildung 6: Probleme bei der Auffindbarkeit der Daten

Antworten zu der Frage „Welche Probleme traten beim Auffinden der Datensätze und Annotationen auf?“ Doppelnennungen waren erlaubt. Insgesamt gab es 18 Rückmeldungen.



Quelle: eigene Darstellung, wetransform GmbH

In Bezug auf Geodaten wurde dieses Problem durch die Schaffung der Geodateninfrastruktur Deutschland[14] (GDI-DE) adressiert. Seit 2005 kooperieren Bund, Länder und Kommunen, um auf diese Weise öffentliche Geodaten in einer webbasierten, vernetzten und auf Standards beruhenden Geodateninfrastruktur bereitzustellen.

4.3.1.2 Qualitativ hochwertige Metadaten (Findability)

Das Vorhandensein von qualitativen hochwertigen Metadaten ist essenziell für die Auffindbarkeit und Nachnutzung von Umweltdaten und umweltrelevanten Daten und sollte in einer standardisierten Metadatenbeschreibung zur Verfügung gestellt werden. Dies wurde insbesondere in den Stellungnahmen zum FDG als wichtiger Punkt mehrfach genannt (DFG, 2023; Fraunhofer Gesellschaft, 2023; DLR, 2023; GFZ & FID GEO, 2023).

Für INSPIRE-relevante Geodaten und Dienste wurde im Zuge der Verordnung (EG) Nr. 1205/2008 (Implementing Rules on Metadata) dies bereits standardmäßig umgesetzt. Insbesondere definiert INSPIRE ein spezifisches Metadaten-Profil, das sicherstellt, dass die Metadaten standardisiert und interoperabel sind. Dieses Profil basiert auf internationalen Standards wie ISO 19115 und ISO 19119.

4.3.1.3 Verbesserung des Zugangs (Accessibility)

Bessere Zugangsmöglichkeiten zu Umweltdaten und umweltrelevanten Daten inklusive der Metadaten und Annotationen können dazu beitragen, die gegenwärtigen und zukünftigen Herausforderungen besser zu verstehen und innovative Lösungen zu entwickeln. Um die Zugänglichkeit zu erhöhen, ist es vor allem wichtig, dass beispielsweise technische Barrieren abgebaut werden. Solche Barrieren können komplexe APIs, proprietäre Dateiformate, der Bedarf, komplexe Desktop-Anwendungen zu nutzen oder auch fehlende Dokumentationen sein. Idealerweise sind Daten in (mehreren) offenen, dokumentierten Formaten und Schemas über allgemein standardisierte Schnittstellen zugänglich und lassen sich leicht in verschiedene Systeme einbinden, bis hin zu allgemein verfügbaren Tools wie Tabellenkalkulationen. Auch einfache Online-Werkzeuge zur Visualisierung der Daten können die Zugänglichkeit erhöhen.

[14] https://www.gdi-de.org

Dennoch stellt der zum Teil fehlende diskriminierungsfreie und barrierefreie Zugang eines der Hauptprobleme bei der Nutzung von Umweltdaten und umweltrelevanten Daten dar. So treten Probleme beim Zugriff auf komplexe INSPIRE-Daten auf, aber auch bei der Nutzung proprietärer Formate wie Filegeodatabases.

4.3.1.4 Standardisierte Formate (Interoperability)

Daten sollen in offenen, standardisierten Formaten bereitgestellt werden, welche eine automatisierte Weiterverarbeitung erlauben. Dies betrifft die Daten, die Metadaten sowie die Annotationen. Um Interoperabilität sicherzustellen, muss neben der Nutzung von offenen, standardisierten Formaten besonderes Augenmerk auf die Dokumentation der Daten gelegt werden. Weiterhin ist die Verwendung einer standardisierten Semantik, sofern vorhanden und anwendbar, ein Schlüsselfaktor zur Verbesserung der Interoperabilität. Im Geo- und Umweltbereich gibt es zahlreiche konzeptionelle und logische Schemata (z.B. von der EU, siehe INSPIRE), Codelisten und Konzepte, die wohldefiniert sind, und welche wann immer möglich auch für annotierte Daten verwendet werden sollten.

Eine ausführliche Analyse dazu, welche Aspekte der Interoperabilität besonders wichtig sind, um die Wiederverwendbarkeit von Daten zu ermöglichen, zeigt weiterhin, dass der Zugriff über offene APIs sowie klare Lebenszyklus-Regeln für Daten (also wann und unter welchen Bedingungen sie ihre Gültigkeit verlieren) nicht nur relativ einfach umzusetzen sind, sondern auch eine hohe Wirkung erzielen können (Reitz, Escriu, Minghini, & Kotsev, In review).

4.3.1.5 Nachnutzbarkeit, Nachvollziehbarkeit und Reproduzierbarkeit (Reuse)

Um eine effiziente Nachnutzung von Umweltdaten und umweltrelevanten Daten zu gewährleisten, ist es unumgänglich, die Details zur Datenverarbeitung zu kennen (DFG, 2023; GFZ & FID GEO, 2023). In diesem Zusammenhang sagt der NFDI, dass *„grundsätzlich […] nur Forschungsdaten für eine Forschung herangezogen werden können, die gut kuratiert sind und die über zertifizierte Systeme mit eindeutigen, langzeitverfügbaren PIDs (Persistent Identifier) versehen sind“* (NFDI, 2023, S. 4).

Darüber hinaus ist es essenziell, die Verarbeitungshistorie der Daten zu kennen. In Bezug auf Trainingsdaten betrifft dies insbesondere die Methodik, welche für das Annotieren der Daten verwendet worden ist, wie beispielweise manuelle, automatisierte oder synthetische Verfahren (siehe auch Kapitel 4.4.4.5). Ebenfalls zentral ist eine klare Dokumentation der Bedingungen für den Datenzugriff und die Datennutzung, wobei bevorzugt Standard-Datenlizenzen zum Einsatz kommen sollten.

Als Standard für die disziplinübergreifende Nachnutzbarkeit gelten die FAIR-Prinzipien, welche diese mittels guter Beschreibung, offenen Formaten, klaren Lizenzen und Gewährleistung der Maschinenlesbarkeit ermöglichen. Diese wurden bereits in die DFG-Leitlinie zur Sicherung guter wissenschaftlicher Praxis übernommen (DFG, 2019). Es sollte jedoch zum Standard werden, dass die FAIR-Prinzipien nicht nur innerhalb der Wissenschaft, sondern auch in anderen Sektoren zur Gewährleistung der Nachnutzbarkeit angewandt werden.

Im Open Geospatial Consortium (OGC)-Standard „TrainingDML-AI” sind entsprechende Standards zur Provenienz von Daten definiert, um grundlegende Herkunftsinformationen darüber, wie der Trainingsdatensatz oder die Trainingsmuster erstellt wurden, zu vermerken (Yue & Shangguan, 2023).

4.3.2 Robuste Infrastruktur

Ein wichtiger Aspekt für die Nachnutzung von Daten und Datenannotationen ist, dass diese über eine robuste Infrastruktur mit hoher Bandbreite und geringen Ausfallzeiten bereitgestellt werden und damit jederzeit auffindbar und abrufbar sind. Der Datenaustausch sollte automatisiert erfolgen (z.B. mittels APIs). Solche stabilen Infrastrukturen können selten in einzelnen Projekten etabliert werden, sondern erfordern eine langfristige Bereitschaft zentraler Stakeholder. Aus diesem Grund sind viele der längerfristig stabilen Plattformen kommerzielle Angebote.

4.3.3 Gewährleistung der Nachhaltigkeit

Ein wichtiger Aspekt in Bezug auf die Anforderung an Umweltdaten ist die Sicherung der Nachhaltigkeit dieser Daten. Dies bedeutet, dass insbesondere die Datenbereitstellung und -pflege langfristig finanziert sein sollte, um Daten- und Planungssicherheit auf Seiten der Datennutzenden, aber auch der Datenbereitstellenden zu garantieren. Dies umfasst unter anderem die Ermöglichung und Sicherstellung regelmäßiger Datenerhebungen in festgelegten Intervallen, um konsistente und verlässliche Zeitreihen zu gewährleisten. Diese Forderung wurde auch von verschiedensten Institutionen im Zuge der Stellungnahmen zum FDG genannt (DFG, 2023; NFDI, 2023; DLR, 2023; GFZ & FID GEO, 2023).

Neben der langfristigen Finanzierung muss bei der Datenerhebung und Datenverarbeitung auch darauf geachtet werden, dass die gesammelten Daten genau und verlässlich sind. Hierfür sollten robuste Qualitätskontroll- und Sicherungsprozesse implementiert werden (siehe auch Abschnitt 4.3.5).

Ein gutes Beispiel für die Innovationskraft langfristig und verlässlich verfügbarer Datensätze sind die Erdbeobachtungsdaten aus dem Copernicus-Programm. Die frei verfügbaren Copernicus-Daten sind in den letzten Jahren ein wichtiger Treiber für KI-gestützte Vorhaben im Umweltsektor gewesen.

Weiterhin sind Investitionen in den Aufbau von Infrastrukturen notwendig, d.h. in die Bereitstellung und Wartung moderner Technologien und Systeme zur Datensammlung, -speicherung und -analyse (DLR, 2023; NFDI, 2023; GFZ & FID GEO, 2023). Dabei ist eine kontinuierliche Schulung und Weiterbildung des Personals wichtig, um sicherzustellen, dass die neuesten Techniken und Best Practices auch angewendet werden. Die Ausprägung von Datenkompetenz wurde ebenfalls im Zuge der Stellungnahmen zum FDG (DFG, 2023) sowie in den Stakeholder-WS zum Vorhaben als Erfolgsfaktor genannt.

Für Daten, die in einem konkreten Projektkontext aufgenommen werden, werden weitere Maßnahmen empfohlen, damit die Daten nicht nur im Rahmen des spezifischen Projekts, sondern auch langfristig und projektübergreifend nutzbar bleiben. Diese Maßnahmen gelten insbesondere für Trainingsdaten, die häufig im Kontext einer bestimmten Anwendung oder eines befristeten Projekts aufgenommen bzw. erzeugt werden.

- Datenzugänglichkeit: Sicherstellung, dass die erhobenen Daten langfristig in offenen und für verschiedene Nutzergruppen zugänglichen Datenbanken gespeichert werden.
- Interoperabilität der Daten: Entwicklung und Einhaltung von Standards, die die Interoperabilität der Daten gewährleisten, sodass unterschiedliche Systeme und Nutzer*innen die Daten effizient nutzen können (siehe auch Abschnitt 4.3.4).

- Nachnutzung und Weiterverwendung: Förderung der Nachnutzung und Weiterverwendung der Daten durch andere Forschungsprojekte, Institutionen und Organisationen, um den größtmöglichen Nutzen aus den erhobenen Daten zu ziehen.

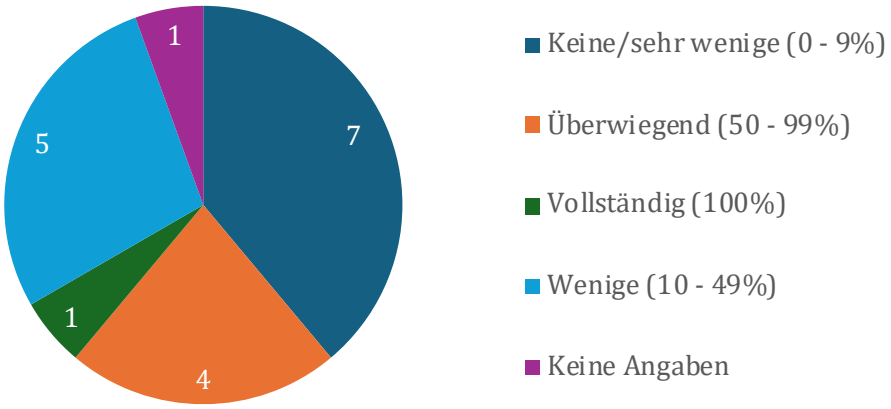
Eine Initiative, die diese Aspekte adressiert, ist die Nationale Forschungsdateninfrastruktur (NFDI).

4.3.4 Vorhandensein standardisierter Annotationen

Vermehrt wurde der Bedarf an annotierten Daten sowie einer standardisierten Beschreibung der Annotationen in den Stellungnahmen zum FDG geäußert (FZI, 2023; DLR, 2023; DZSF, 2023). Beispielsweise verwies das Forschungszentrum Informatik (FZI) auf den Bedarf von gelabelten „Referenzdatensätze, um Vergleichbarkeit verschiedener Ansätze zu ermöglichen (bspw. für das Training von KI-Modellen)" (FZI, 2023, S. 1). Diese Limitierung wurde auch im Zuge der Stakeholder-WS mehrfach angesprochen. Die Onlinebefragung im Forschungsvorhaben zeigte, dass nur für wenige Datensätze Annotationen zur Verfügung stehen (Abbildung 7), jedoch ein großer Bedarf besteht.

Abbildung 7: Vorhandensein von Annotationen

Antworten zu der Frage „Gab es zu diesen Daten bereits Annotationen/Labels?" Doppelnennungen waren nicht erlaubt. Insgesamt gab es 18 Rückmeldungen.



Quelle: eigene Darstellung, wetransform GmbH

Für Erdbeobachtungsdaten wurde ein interner Standard für Trainingsdaten durch das OGC entwickelt, welcher die Prozesse in Bezug auf Trainingsdaten beschreibt[15]. Der Standard enthält folgende Aspekte:

- Der TrainingDML-AI Conceptual Model Standard definiert, wie ML-Trainingsdaten dargestellt und ausgetauscht werden können mit dem Ziel, die Interoperabilität und Verwendbarkeit von EO-Bilder-Trainingsdaten zu maximieren.
- Er definiert verschiedene KI/ML-Aufgaben und Labels in der Erdbeobachtung im Sinne des überwachten Lernens, einschließlich Aufgaben auf Szenen-, Objekt- und Pixelebene.

[15] https://docs.ogc.org/is/23-008r3/23-008r3.html

- Beschreibung der permanenten Kennung, der Version, der Lizenz, der Größe der Trainingsdaten, der für die Beschriftung verwendeten Messungen oder Bilder usw.
- Beschreibung der Qualität (z. B. Fehler in den Trainingsdaten, mangelnde Repräsentativität der Trainingsdaten) und der Herkunft (Kennzeichnung, Kennzeichnende und Kennzeichnungsverfahren).

In den Recherchen zu diesem Vorhaben konnte jedoch kein Leitfaden erschlossen werden, der sich explizit auf Annotationen für Umweltdaten und umweltrelevante Daten bezieht.

4.3.5 Datenqualität

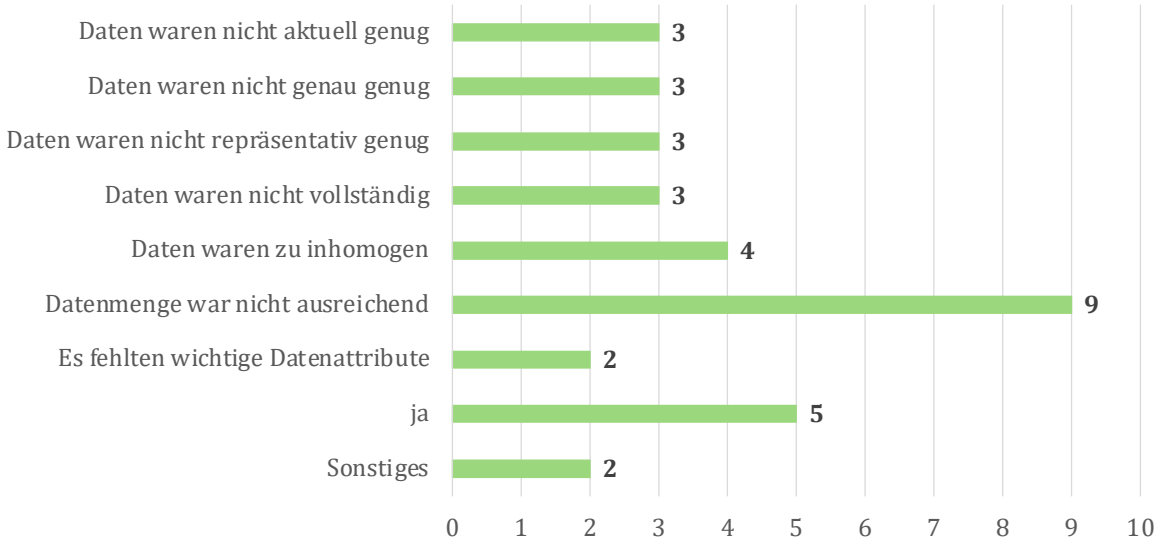
Die Datenqualität spielt eine zentrale Rolle und ist die Basis für die Entwicklung solider ML-Verfahren. Zu Beurteilung der Datenqualität sind verschiedenen Kriterien zu beachten. Dazu zählen:

- Vollständigkeit
- Eindeutigkeit
- Korrektheit
- Aktualität
- Genauigkeit
- Konsistenz
- Redundanzfreiheit
- Relevanz

Die Onlinebefragung zeigte, dass die Datenqualität im Rahmen der Modellentwicklung eine große Rolle spielt (Abbildung 8). Nur in fünf von 18 Antworten wurde die Datenqualität als ausreichend bewertet. Als Hauptproblem nannten die Befragten die Datenmenge, wobei eine Unterscheidung zwischen der Vollständigkeit („Daten waren nicht vollständig“) sowie der Anzahl der zur Verfügung stehenden Daten („Datenmenge war nicht ausreichend“) getroffen werden konnte. Weitere Problemfelder wurden in der unzureichenden Aktualität und der Inhomogenität der verfügbaren Daten, wie beispielsweise Luftbildern, gesehen.

Abbildung 8: Datenqualität

Antworten zu der Frage „War die Datenqualität für ihren Anwendungsfall ausreichend?" Doppelnennungen waren erlaubt. Insgesamt gab es 18 Rückmeldungen.



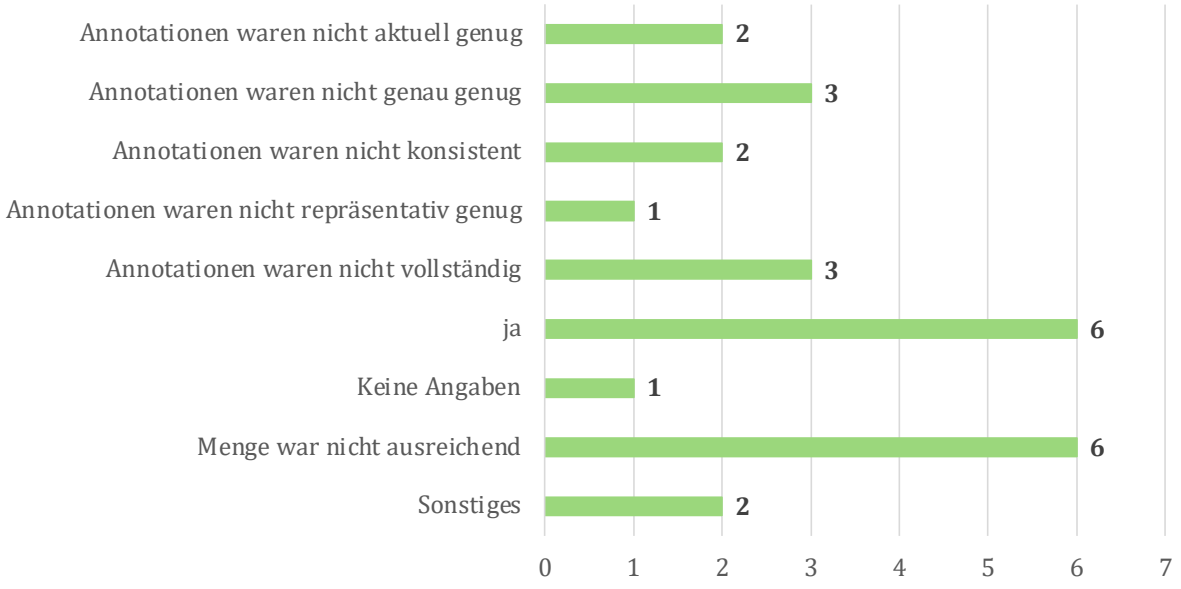
Quelle: eigene Darstellung, wetransform GmbH

Erforderlich sind daher qualitätsgesicherte Datensätze mit guten Metadatenschemata. Etwaige Ungenauigkeiten sollten zwingend in den Metadaten dokumentiert werden. Zudem sollte es Datennutzenden möglich sein, die Qualität der Daten bzw. der Datensatz ausweisenden Stelle eindeutig einschätzen zu können. Für geografische Daten sollte zwingend der ISO 19157-Standard angewandt werden. Denkbar wäre in diesem Zusammenhang beispielsweise die Etablierung eines „Qualitätssiegels", welches basierend auf Standards oder der Bewertung durch Nutzende einer Community auf einer entsprechenden Plattform eine eindeutige Bewertung des Datensatzes erlaubt.

Bezogen auf die Qualität der annotierten Daten zeigte die Onlinebefragung, dass bei vorhandenen Datenannotationen, insbesondere die Datenmenge der Annotationen kritisch ist (Abbildung 9). Dieser Sachverhalt wurde auch mehrfach im Zuge der Stellungnahmen zum FDG genannt (FZI, 2023) und war auch ein wesentlicher Diskussionspunkt in den Workshops. Als weitere Limitierungen wurden die Datengenauigkeit (3 Nennungen) sowie die Vollständigkeit (3 Nennungen) genannt.

Abbildung 9: Qualität der Annotationen

Antworten zu der Frage „War die Qualität der bestehenden Annotationen/Labels für ihren Anwendungsfall ausreichend?" Doppelnennungen waren erlaubt. Insgesamt gab es 18 Rückmeldungen.



Quelle: eigene Darstellung, wetransform GmbH

Im OGC-Standard „TrainingDML-AI" sind Anforderungen an die Labels und den Labeling-Prozess definiert (Yue & Shangguan, 2023, *Kapitel 7.6 AI_Label*):

„Labels for each individual training sample can be represented using a feature, coverage, or a semantic class from ontologies or shared vocabularies (external classification schemes and classes). A training sample typically relates pixels, objects, and scenes to semantic labels, where labels are encoded as coverage, geometric polygon, and text respectively." (Chapter 7.6 AI_Label)

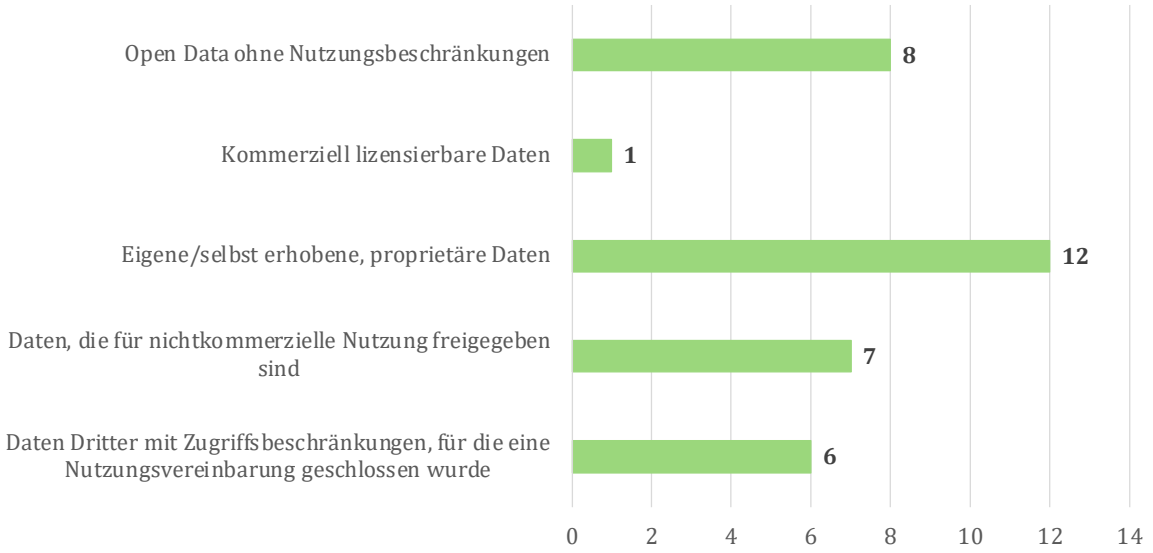
4.3.6 Einheitliches Lizenzsystem und Klärung rechtlicher Fragestellungen

Einen wesentlichen rechtlichen Aspekt für die Bereitstellung und Nutzung von Daten stellt das Lizenzsystem dar. Die Recherchen zeigten, dass Nutzende dazu tendieren, Daten neu aufzunehmen, anstatt bestehende Datensätze wiederzuverwenden, wenn Zugangs- oder Nutzungsbedingungen unklar sind.

Diese Situation spiegeln auch die Ergebnisse der Onlinebefragung wider. Der Großteil der Befragten gibt an, eigene selbst erhobene bzw. proprietäre Daten zu verwenden (Abbildung 10). Auch Open Data, mit klaren Nutzungsbedingungen, wurden vielfach verwendet.

Abbildung 10: Datenlizenzen

Antworten zu der Frage „Unter welche Datenlizenzen fallen die Daten, die Sie genutzt haben?" Doppelnennungen waren erlaubt. Insgesamt gab es 18 Rückmeldungen.



Quelle: eigene Darstellung, wetransform GmbH

Auch in den Workshops war die Unsicherheit bezüglich der rechtlichen Situation ein wesentlicher limitierender Faktor. Als Beispiel führte das KI-Lab die Auswertung von Bildern aus Kleinanzeigen zur Detektion des illegalen Online-Artenhandels an. Aktuell wurden knapp 100.000 Tierbilder eingesammelt und mit relativ verlässlichem Label versehen. Als problematisch eingeschätzt werden jedoch der Datenschutz und das Urheberrecht, da diese Bilder durch Privatpersonen erstellt worden sind, um die Tiere auf den jeweiligen Onlineplattformen zu verkaufen. Die rechtliche Lage zur Wiederverwendung dieser Daten ist daher nicht trivial. Ähnlich komplex gestaltet sich die rechtliche Lage, wenn es sich um Produkte handelt, die aus proprietären Daten abgeleitet worden sind. Grundsätzlich ist die Veröffentlichung von abgeleiteten Produkten zwar möglich, allerdings ist es regelmäßig schwierig zu beurteilen, was ein abgeleitetes Produkt ist, so dass es häufig einer Einzelfallbetrachtung und -entscheidung bedarf.

Die Komplexität der rechtlichen Situation von Daten wurde ebenfalls als limitierender Faktor erwähnt (siehe Kapitel 4.4.1). Zudem wurde die fehlende Klärung rechtlicher Fragestellungen in Bezug auf Datenschutz, Datenurheberschaft, Nutzungs- und Verwertungsrechte, Datensicherheit, Geheimhaltungspflichten sowie IP-Schutz im Zuge der Stellungnahmen zum FDG genannt (FZI, 2023; UFZ, 2023) und verdeutlicht nochmals die Schwierigkeit der Thematik.

4.4 Limitierungen und Hindernisse

Basierend auf den Ergebnissen der Literaturrecherche, den Stakeholder-Workshops, den Experteninterviews sowie der Onlinebefragung werden im folgenden Kapitel die bestehenden rechtlichen, wirtschaftlichen, infrastrukturellen und technischen Limitierungen in Bezug auf die Bereitstellung und Nutzung annotierter Datensets für ML-Modelle beschrieben. Dabei werden drei Stufen verwendet, um die Häufigkeit des Auftretens der Limitierung zu bewerten:

- Gering: Die Limitierung wurde von einzelnen (unter 10%) Stakeholdern genannt.
- Mittel: Die Limitierung wurde von vielen (10 bis 50%) Stakeholdern genannt.

- Hoch: Die Limitierung wurde von den meisten Stakeholdern (50%) genannt.

Die Auswirkungen der Limitierungen wurden nach dem gleichen Schema bewertet, wobei zusätzlich die qualitativen Aussagen berücksichtigt wurden:

- Gering: Einzelne der Stakeholder (unter 10%) schätzen diese Limitierung als Blocker ein.
- Mittel: Viele der Stakeholder (10 bis 50%) schätzen diese Limitierung als Blocker ein.
- Hoch: Die meisten Stakeholder (50%) schätzen diese Limitierung als Blocker ein.

Tabelle 16 gibt einen Überblick; Details werden in den folgenden Kapiteln beschrieben.

Tabelle 16: Bestehende rechtliche, wirtschaftliche, infrastrukturelle und technische Limitierungen und deren Bewertung

Kategorie der Limitierung	Kurz-Beschreibung	Häufigkeit der Limitierung	Auswirkung der Limitierung
Rechtlich	Unsicherheit bzgl. der rechtlichen Situation	Hoch	Hoch
	Fehlende Lizenzen für Datennachnutzung	Hoch	Hoch
	Keine rechtlichen Angaben zur Aufbewahrungsdauer	Gering	Gering
Wirtschaftlich	Hohe Kosten für die Datenaufbereitung	Hoch	Hoch
	Fehlende Anreizsysteme	Gering	Mittel
Infrastruktur	Fehlende nachhaltige und langfristige Datenbereitstellung	Hoch	Hoch
	Fehlende Datenmittlerstruktur	Mittel	Hoch
Technisch	Fehlende Interoperabilität	Hoch	Hoch
	Fehlende Standards für die Datenaufbereitung	Hoch	Hoch
	Fehlender standardisierter Metadatenkatalog	Hoch	Hoch
	Geringe Verfügbarkeit von Datenannotationen	Hoch	Hoch
	Lücke in der Provenienz	Mittel	Hoch
	Heterogenität der bestehenden Datenbanken	Mittel	Mittel

4.4.1 Rechtliche Limitierungen und Hindernisse

4.4.1.1 Unsicherheit bezüglich der rechtlichen Situation

Die hohe Komplexität der rechtlichen Situation von Daten, beispielsweise in Bezug auf Datenurheberschaft, Nutzungs- und Verwertungsrechte, wurde vermehrt im Zuge der Stellungnahmen zum FDG angeführt (DLR, 2023; Fraunhofer Gesellschaft, 2023; FZI, 2023; UFZ, 2023; NFDI, 2023; Thünen Institut, 2023; Hochschule Bochum, 2023; Bitkom, 2023) und erschwert u.a. den kompetenten Umgang mit den Daten. Als Beispiele wurden in diesem Zusammenhang die eingeschränkte wissenschaftliche und kommerzielle Nutzung der Daten der nationalen Satelliten-Erdbeobachtungsmissionen TerraSAR-X und TanDEM-X aufgrund des restriktiven Satellitendatensicherheitsgesetzes und der dazugehörigen Verordnung (SatDSiG, SatDSiV) genannt (DLR, 2023).

Aber auch die Datenschutzanforderungen erschweren das Teilen von Daten. Viele Datensätze müssen vorher anonymisiert werden, wobei die eigentlichen Rohdaten für Forschungszwecke wesentlich interessanter wären (Bitkom, 2023). Als Beispiel wurden Mobilitätsdaten angeführt, welche meist nur mit Verweis auf Geheimnisschutz und auch nur für einzelne Projekte freigegeben werden und nicht für Forschung allgemein. Auch seitens des UBA wurde die komplexe rechtliche Situation betont (siehe Kapitel 4.3.6), welche meist im Einzelfall beurteilt werden muss.

4.4.1.2 Fehlende Lizenzen für Datennachnutzung

Eine Herausforderung in der Nachnutzung von Daten stellen auch die unterschiedlichen bzw. teilweise fehlenden Lizenzen, unter welchen die Daten nachgenutzt werden können, dar (GFZ & FID GEO, 2023). Dies betrifft auch öffentlich verfügbare und behördlichen Daten, bei welchen aufgrund der unterschiedlichen Lizenzen die überregionale Nachnutzung und vor allem die Nachnutzung und Veröffentlichung auch abgeleiteter Datenprodukte oft sehr zeitaufwändig bis unmöglich ist (UFZ, 2023).

4.4.1.3 Keine rechtlichen Angaben zur Aufbewahrungsdauer

Zwar wurde der Zugang zu Umweltdaten rechtlich durch u.a. das Umweltinformationsgesetz, Geodatenzugangsgesetz und §12a EGovG geregelt; jedoch sind Vorgaben zur Aufbewahrungsdauer nur unzureichend definiert (DLR, 2023). Dies gilt insbesondere für Daten, die keiner entsprechenden gesetzlichen Regelung oder Verpflichtung unterliegen (bspw. Geodatenzugangsgesetz).

4.4.2 Wirtschaftliche Limitierungen und Hindernisse

4.4.2.1 Hohe Kosten für die Datenaufbereitung[16]

Einer der wirtschaftlichen Hauptgründe dafür, dass Daten nicht nachhaltig – nach den FAIR-Prinzipien - aufgearbeitet und bereitgestellt werden, sind die zusätzlichen Aufwände für die Datenaufbereitung für Daten von Datenproduzenten bzw. Anbieter für die Bereitstellung der Daten außerhalb der eigenen Organisation. Ein Beispiel hierfür ist das Teilen von Umweltdaten konform zu den Umsetzungsanforderungen der INSPIRE-Direktive: Diese sind auch drei Jahre nach Ablauf der entsprechenden Fristen (2020 bzw. 2021) nur zu 20 bis 25 % vollständig umgesetzt, wobei die datenhaltenden Stellen immer wieder auf hohe Aufwände für die Datenharmonisierung und einen geringen Nutzen für die bestehenden Prozesse verweisen.

Um Daten außerhalb der eigenen Organisation wiederverwendbar zu machen, fallen typischerweise zusätzliche Aufwände in verschiedenen Bereichen an:

- **Rechtliche Einschätzung und Lizenzierung:** Organisationen müssen bewerten, ob Rechte Dritter berührt werden, ob durch das Teilen gesetzliche Vorgaben verletzt werden und unter welchen Zugangs- und Nutzungsbedingungen eine Datenressource zugänglich gemacht werden soll.
- **Dokumentation:** Daten müssen so dokumentiert werden, dass sie für Außenstehende verständlich und nachvollziehbar sind. Dies beinhaltet auch die Erstellung von Metadaten, beispielsweise zur Verarbeitungsgeschichte („Lineage“).

[16] Die vorliegende Studie thematisiert lediglich die Aufwände für die Kosten der externen Datenaufbereitung. Das Eigeninteresse von Organisationen, Daten so aufzubereiten, dass sie für weitere Anwendungsfälle zumindest innerhalb der Organisation und ihres unmittelbaren Umfeldes nachgenutzt werden können, wurde nicht betrachtet.

- **Harmonisierung:** In vielen Fällen müssen Daten, um FAIR – und somit effektiv – geteilt werden zu können, in Bezug auf Formate, Struktur und Semantik harmonisiert werden. Dies kann beispielweise die Nutzung zusätzlicher Labels nach einem standardisierten Vokabular bedeuten und somit Mehraufwände erzeugen.
- **Anonymisierung**: Unternehmen müssen beispielsweise vorab prüfen, ob Daten ggf. personenbeziehbar sind und sie anschließend bereinigen und anonymisieren, bevor die Daten geteilt werden können (Bitkom, 2023).
- **Dauerhafter Betrieb und Pflege**: Auch Datensätze veralten und müssen gewartet werden. Stellt eine Organisation einen solchen Datensatz bereit, welcher dann aktiv genutzt wird, entsteht eine langfristige Verpflichtung zur Pflege und Bereitstellung. Aus dem letzten Punkt ergeben sich in der Regel zusätzliche Kosten für Dateninfrastrukturen.

Auch auf Seiten der Wissenschaft wird bei begrenzten und befristeten Forschungsprojektmitteln oft gezögert, die oben genannten Aufwände für die Datenaufbereitung zu erbringen (European Commission, 2018).

4.4.2.2 Fehlende Anreizsysteme

Neben den hohen Kosten stellen auch fehlende Anreizsysteme für Wissenschaft und Wirtschaft einen der Gründe dar, weshalb Daten nicht geteilt bzw. entsprechend aufbereitet werden (European Commission, 2018). Der Aufbau von Infrastruktur und Informationstechnologie (z.B. Forschungsdatenzentren) ist folglich teuer und wenig lukrativ (Fraunhofer Gesellschaft, 2023). Daher ist es einerseits wichtig zu vermitteln, wie technische, soziale und rechtliche Herausforderungen überwunden werden können, und andererseits klar den Mehrwert abzubilden, um Anreize zu schaffen (NFDI, 2023). Der Fokus sollte dabei auch darauf gerichtet sein, bestehenden Ängste abzubauen, um eine gemeinsame Nutzung der Daten zu fördern.

In seiner Stellungnahme zum Forschungsdatengesetz schlägt der NFDI folgende Maßnahmen vor (Tabelle 17).

Tabelle 17: Anreize, um die Bereitschaft zum Datenteilen zu erhöhen (NFDI, 2023)

Sektor	Daten
Unternehmen und Wirtschaft	• klare rechtliche Rahmenbedingungen, etwa zum Haftungsausschluss; • finanzielle Wertschöpfungsmodelle für Nachnutzung unternehmerischer Daten; • Standards für die Qualität von Daten: viele Unternehmen werden die Nachnutzung von frei verfügbaren Daten von eigenen Kriterien an die Qualität abhängig machen; • Datenteilen als Voraussetzung für Fördermittel;
Wissenschaft	• die frühe Sensibilisierung der Forschenden im Rahmen ihrer akademischen Laufbahn (z.B. Datenmanagement als Pflichtteil von Bachelor-/Master-Studiengängen) sowie deutschlandweite Einführung von Datenmanagement-Plänen in Lehrplänen; • angemessene Ressourcenausstattung; • gleichwertige Anerkennung von zitierbaren FAIRen Datenpublikationen (z.B. in den References der Journale) und zugehöriger Software-Tools; • Schaffung unbefristeter Mittelbaustellen für die Forschungsunterstützung sowie • gesicherte Qualitätsmanagementverfahren und -zertifizierungen, die Vertrauen in bereitgestellte Daten geben, um wirklich den Nutzungswert der Daten einschätzen zu können und damit die Nachnutzung zu erhöhen;

Sektor	Daten
Zivilgesellschaft	• Wertschätzung, Nennung des Beitrages zum Erkenntnisgewinn für Wissenschaft und Gesellschaft (vgl. Motivation für Citizen Science); • größere Sinnstiftung für Citizen-Science-Projekte; • erhöhte Transparenz und Unterstützung;
Öffentliche Stellen	• Aufklärung, Erhöhung von Servicementalität; • Forschungsergebnisse für die eigene Einrichtung gewinnbringend nutzen; • (insbesondere Museen und kuratierend tätige Einrichtungen); • Vereinheitlichung der Zugangswege und einheitliche Interpretationen der Gesetzgebung zu Zugangswegen über die Disziplinen hinweg im Rahmen des Archivrechts.

4.4.3 Infrastrukturelle Limitierungen und Hindernisse

4.4.3.1 Fehlende nachhaltige und langfristige Datenbereitstellung

Die fehlende kontinuierliche und dauerhafte technische Betreuung von Dateninfrastrukturen (Repositorien) stellt ein besonders großes Hindernis dar (GFZ & FID GEO, 2023) und bedingt, dass lediglich der Erhalt des Ist-Zustands möglich ist und kein Raum für Neuentwicklung und Innovation besteht (GFZ & FID GEO, 2023). Insbesondere die Unklarheit über langfristige Sicherung, Kuratierung und Bereitstellung der Datenbestände erschweren eine nachhaltige Datenversorgung.

4.4.3.2 Fehlende Datenmittlerstrukturen

Datenmittler*innen, beispielsweise Datentreuhänder*innen, können bei der Nachnutzung schutzbedürftiger Daten eine zentrale Rolle spielen. Datentreuhänder*innen fungieren als vertrauenswürdige Vermittlende, die sicherstellen, dass sensible Daten gemäß den festgelegten Datenschutz- und Nutzungsvorschriften gehandhabt werden. Sie bieten eine strukturierte und rechtlich abgesicherte Plattform für den vertrauensvollen Datenaustausch. Datentreuhänder*innen und andere Datenmittlerstrukturen können außerdem helfen, die heterogene Datenlandschaft zu aggregieren (UFZ, 2023; NFDI, 2023) und somit die Nachnutzung von Daten sehr stark vereinfachen. Nach Einschätzung der Europäischen Kommission fehlen zurzeit solche Strukturen, während gleichzeitig bestimmte Organisationen und Plattformen als „Gatekeeper“ agieren, also die Nutzung von Daten gemäß ihren Interessen steuern. Im Ergebnis ist der „Single Market for Data“ noch dysfunktional.

Der identifizierte Bedarf für Datenmittlerstrukturen wurde u.a. im Zuge der Stellungnahmen zum FDG deutlich. Die Wichtigkeit wird durch folgendes Zitat des NDFI untermauert:

„Im Kern sind Datenmittler und Datentreuhänder essentiell für den Aufbau von Vertrauen in die Rechtmäßigkeit der Datennutzung sowie für die Vermittlung von kritischen Daten.“ (NFDI, 2023, S. 10).

Aktuell werden Datenmittlerstrukturen auf unterschiedlichen Ebenen adressiert:

Datenmittlerstrukturen stehen im Fokus der europäischen Initiative Gaia-X[17]. Mit Gaia-X wird der Aufbau einer leistungsfähigen, sicheren, offenen, vertrauenswürdigen und föderierten Dateninfrastruktur auf Basis europäischer Werte und Normen angestrebt (siehe auch Abschnitt 3.2.3). Ähnliche Ziele verfolgt der von der EU-Kommission vorgeschlagene Data Governance Act (DGA). Dieser sieht die Möglichkeit vor, sektorenspezifisch neutrale Datentreuhänder*innen

[17] https://gaia-x.eu

(sogenannte „neue Intermediäre“) zu schaffen, welche den Datenaustausch zwischen Unternehmen, Behörden und Forschungseinrichtungen erleichtern und fördern sollen. Aus dem Zusammenspiel zwischen Gaia-X und DGA kann ein neuartiges, komplexes Regulierungsregime entstehen, das diversen Stakeholdern vielfältige rechtliche, wirtschaftliche, politische sowie gesellschaftliche Möglichkeitsräume eröffnet.

Als ein Beitrag zur Datenstrategie der Bundesregierung fördert das Bundesministerium für Bildung und Forschung (BMBF) seit 2022 Vorhaben zur Entwicklung und Erprobung von Datentreuhandmodellen. Zweck der Förderung ist es, Datentreuhandmodelle (DTM) in unterschiedlichen Anwendungsbereichen der Wissenschaft und Wirtschaft zu konzipieren, zu entwickeln und in einem Pilotbetrieb unter realen Praxisbedingungen zu testen[18]. Für den Umweltbereich relevante DTM werden u.a. in den Förderprojekten InGeoDTM[19] und DTMForst[20] entwickelt.

4.4.4 Technische Hindernisse

4.4.4.1 Fehlende Interoperabilität[21]

Eines der größten Hindernisse für die aktive Nachnutzung von Daten stellt die fehlende bzw. limitierte Interoperabilität dar (NFDI, 2023). Dies betrifft die technische, informationelle und organisatorische (Workflows) Interoperabilität.

- Technische Interoperabilität (Konnektivität, Netzwerke): Problematisch gestalten sich dabei vor allem das Fehlen von maschinellen Schnittstellen bzw. standardisierte Zugriffsmöglichkeiten auf Daten, wie beispielsweise APIs (UFZ, 2023; DLR, 2023).
- Informationelle Interoperabilität (Semantik, Kontext): Unterschiede in den Datenformaten und die Verwendung von nicht standardisierten Datenformaten werden als hinderlich angesehen und bedingen zeitaufwändige Datenaufbereitung und Datenumformatierungen (Bitkom, 2023). Als Beispiel wurde im Zuge der Stellungnahmen zum FDG auf INSPIRE relevante Daten verwiesen, welche in ihrer Auflösung sehr unterschiedlich (z.B. täglich, monatlich, jährlich) sind und aufgrund dieser Heterogenität oft nicht im Rahmen von Forschungsprojekten genutzt werden (GFZ & FID GEO, 2023).

4.4.4.2 Fehlende Standards für Annotation und Annotationen

Ein weiterer technischer Hindernisgrund, der zu einer geringen Nachnutzbarkeit von Daten führt, ist die fehlende oder lückenhafte Nutzung von Standards in Bezug auf Metadaten und das Format sowie die Semantik der Annotationen selbst (NFDI, 2023). Teils, wie in der Fortwirtschaft, fehlen auch Standards für die Struktur und Semantik der Daten selbst. Die Fraunhofer-Gesellschaft verweist in diesem Zusammenhang darauf, dass fehlende Standards für Datenformate ein wesentliches Problem darstellen und „*Daten mit uneinheitlichen, inkonsistenten und lückenhaften Formaten [..] nicht direkt ausgelesen werden“* können, wodurch ein wesentlicher Aufwand in der Aufbereitung der Daten entsteht (Fraunhofer Gesellschaft, 2023, S. 4).

Diese Situation wurde auch durch die Teilnehmenden der Stakeholder-Workshops bestätigt. Im landwirtschaftlichen Bereich fehlen u.a. die Standards für die Annotation von Drohnendaten

[18] Siehe https://www.bildung-forschung.digital/digitalezukunft/de/technologie/daten/daten_node.html

[19] https://www.igd.fraunhofer.de/de/forschung/oeffentliche-projekte/software/ingeodtm.html)

[20] https://www.kwh40.de/dtmforst/

[21] Als Interoperabilität wird „die Fähigkeit von zwei oder mehr Systemen oder Komponenten, Informationen auszutauschen und die ausgetauschten Informationen zu nutzen” verstanden.

(Punktwolken). Auch im forstwirtschaftlichen Bereich existiert nur bedingt eine Kultur des Standardisierens und des Datenteilens. Die Festlegung von Standards und Tools ist jedoch für die sektorübergreifende Datennutzung essenziell. Aktuell werden diese meist nur Community-intern definiert und umgesetzt.

In Bezug auf die Entwicklung von Datenstandards verweist das GFZ explizit auf den Bereich des „long tails" der Forschungsdaten, welcher besonders herausfordernd ist. Definiert wird dieser als *„kleine, individuelle und hochvariable Daten, die von einzelnen Wissenschaftlern oder kleinen Teams erhoben werden"* (GFZ & FID GEO, 2023, S. 4). Solche Daten finden sich im Umweltbereich, vor allem im Kontext Biodiversität, besonders häufig. Weiterhin führt das GFZ an, dass für diese Art von Daten selten IT-Systeme zur Archivierung benötigt werden, weshalb Standards, wenn überhaupt, nur innerhalb kleiner Gruppen existieren. (GFZ & FID GEO, 2023)

4.4.4.3 Fehlen standardisierter Metadatenkataloge

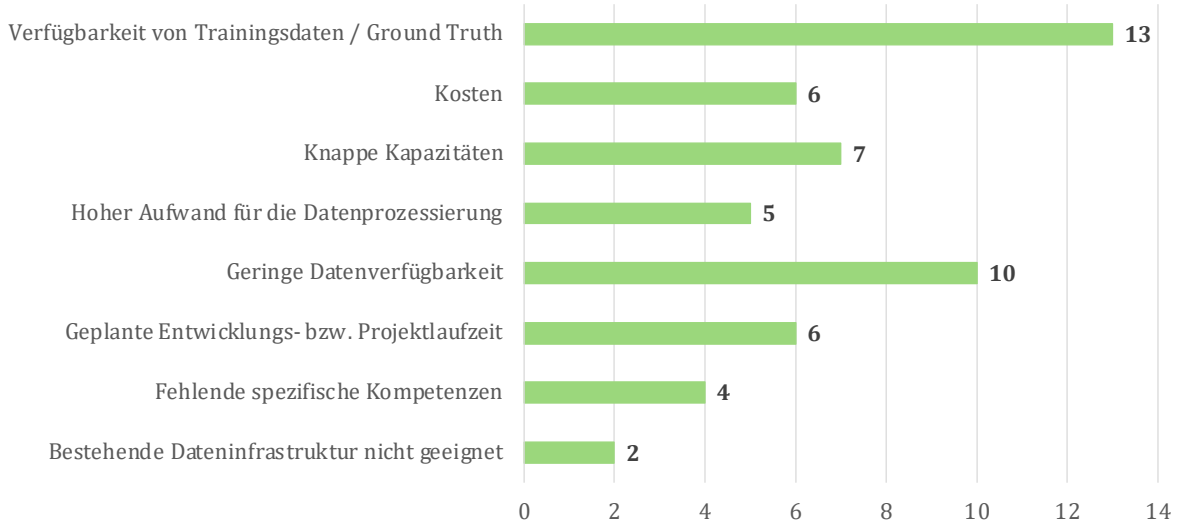
Wie in Kapitel 4.3.1.2 beschrieben, ist das Vorhandensein qualitativ hochwertiger Metadaten eine essenzielle Anforderung an Umweltdaten bzw. umweltrelevante Daten. Als Limitierung wird daher das Fehlen standardisierter Metadatenkataloge genannt, was die Auffindbarkeit und Evaluierbarkeit von Datensammlungen bzw. Einzeldatensätzen erschwert (DLR, 2023; UFZ, 2023; Hochschule Bochum, 2023). In diesem Zusammenhang führen insbesondere fehlende maschinell ansteuerbare Metadatenkataloge (z.B. STAC oder DCAT) zu unzureichenden Datennutzungsmöglichkeiten (UFZ, 2023).

4.4.4.4 Geringe Verfügbarkeit von Datenannotationen

Im Zuge der Onlineumfrage wurde die Verfügbarkeit von Trainingsdaten sowie die Datenverfügbarkeit als wichtigste begrenzende Faktoren für die KI-Modellentwicklung angegeben (Abbildung 11). Zudem spielen die Aspekte Kosten, Kapazitäten und Projektlaufzeit eine entscheidende Rolle.

Abbildung 11: Begrenzende Faktoren der KI-Modellentwicklung

Antworten zu der Frage „Was sind ihre begrenzenden Faktoren bei der KI-Modellentwicklung?" Doppelnennungen waren erlaubt. Insgesamt gab es 18 Rückmeldungen.



Quelle: eigene Darstellung, wetransform GmbH

Unzureichende Datenannotation wurde auch von verschiedensten Organisationen im Rahmen der Stellungnahmen zum FDG genannt (FZI, 2023; Hochschule Bochum, 2023). Diese Herausforderung wird zudem durch die Rückmeldungen, Erfahrungsberichte und Diskussionen der Teilnehmenden in den Stakeholder-Workshops bekräftigt.

Wie oben erwähnt steht der forstwirtschaftliche Sektor vor großen Herausforderungen in Bezug auf das Standardisieren und das Datenteilen. Problematisch ist dabei insbesondere die geringe Bereitschaft zum Teilen von Forstinventurdaten aus wirtschaftlichen und persönlichen Gründen. Vor allem die „Angst vor dem überprüft werden" (Stichwort: Gläserner Wald) stellt ein großes Hindernis dar. Ähnlich wurde die Situation auch im landwirtschaftlichen Sektor beschrieben, da auch hier mittels Inventurdaten Rückschlüsse auf den wirtschaftlichen Wert des Betriebes gezogen werden können.

4.4.4.5 Lücken in der Provenienz[22]

Neben der Datenqualität spielt die Datenherkunft (Provenienz) eine wesentliche Rolle für die Nachnutzung von Daten (DLR, 2023; GFZ & FID GEO, 2023). Dies betrifft im Speziellen die Bereitstellung der Software, die für die Prozessierung der Daten genutzt wurde (DLR, 2023). Das DLR weist in seiner Stellungnahme explizit darauf hin, dass *„diese Lücke in der Provenienz […] den wissenschaftlichen Wert der Daten"* gefährdet (DLR, 2023, S. 4).

4.4.4.6 Heterogenität der bestehenden Datenbanken

Die Auffindbarkeit von Daten wird insbesondere durch die Heterogenität der bestehenden Datenbanken erschwert. Daten sind teilweise auf verschiedene und zum Teil nicht online zugreifbare Repositorien verteilt, wodurch das tatsächliche Datenangebot sehr unübersichtlich ist (siehe auch Kapitel 4.3.1.1).

Die verschiedenen Repositorien basieren auf unterschiedlichen technischen Lösungen, so dass eine effiziente Nutzung erst nach einer gewissen Einarbeitungszeit möglich ist (Thünen Institut, 2023). Zudem stellen die unterschiedlichen Datenformate eine große Herausforderung für die datengetriebene Forschung dar (GFZ & FID GEO, 2023) und verursachen, zusammen mit den verschiedenen Nutzungsbedingungen, zusätzliche Aufwände.

[22] Der Begriff Provenienz wird im Datenmanagement als Dokumentation darüber, woher Daten stammen und mit welchen Prozessen und Methoden (inkl. Software) diese produziert worden, verstanden.

5 Potential annotierter Daten

Qualitätsgeprüfte annotierte Datensätze sind die Grundlage für eine gute KI-Modellbildung und stellen eine treibende Kraft für die Weiterentwicklung KI-gesteuerter Umweltforschung dar. Das folgende Kapitel soll einen Überblick über das politische, wirtschaftliche, wissenschaftliche und technische sowie soziale und ökologische Potential von KI bzw. ML-Algorithmen im Umweltressort geben und sektorübergreifende Potentiale aufzeigen.

Die Potentiale werden vorrangig hinsichtlich des generellen Einsatzes von KI im Umwelt- und Naturschutz analysiert. Wo möglich, wurde jedoch auch explizit zu annotierten Daten, als wichtige Eingangsgröße zur Realisierung des Potentials von KI-Modellen, Stellung genommen.

5.1 Politisches Potential

5.1.1 Förderung von Umwelt- und Naturschutzmaßnahmen und Gestaltung von Umweltgesetzen und -richtlinien

Technologische Fortschritte ermöglichen eine Beschleunigung der Politikgestaltung und tragen dazu bei, dass politische Entscheidungen auf einer fundierteren, datenbasierten Grundlage getroffen werden können (Höchtl, Parycek, & Schöllhammer, 2016). Mithilfe von KI können große Datenmengen ausgewertet und basierend darauf datenbasierte Handlungsempfehlungen erarbeitet werden, welche zur Förderung von Umwelt- und Naturschutzmaßnahmen sowie in weiterer Folge zur Gestaltung von Umweltgesetzen und -richtlinien beisteuern können. Grundlage für eine solide Modellentwicklung und das Ableiten begründeter Handlungsempfehlungen ist das Vorhandensein von hochqualitativen Trainingsdaten. Durch diese können KI- bzw. ML-Methoden wesentlich dazu beitragen, umweltpolitischen Entscheidungsträgerinnen*Entscheidungsträgern ein besseres Verständnis der technischen, ökonomischen oder sozialen Zusammenhänge im jeweiligen Sektor zu geben, um differenzierte und gezielte Maßnahmen zu konzipieren und politische Handlungsansätze besser daran auszurichten. In den folgenden Kapiteln werden Beispiele aus der Landwirtschaft (Kapitel 5.1.1.1) und Forstwirtschaft (Kapitel 5.1.1.2) angeführt, welche dies verdeutlichen.

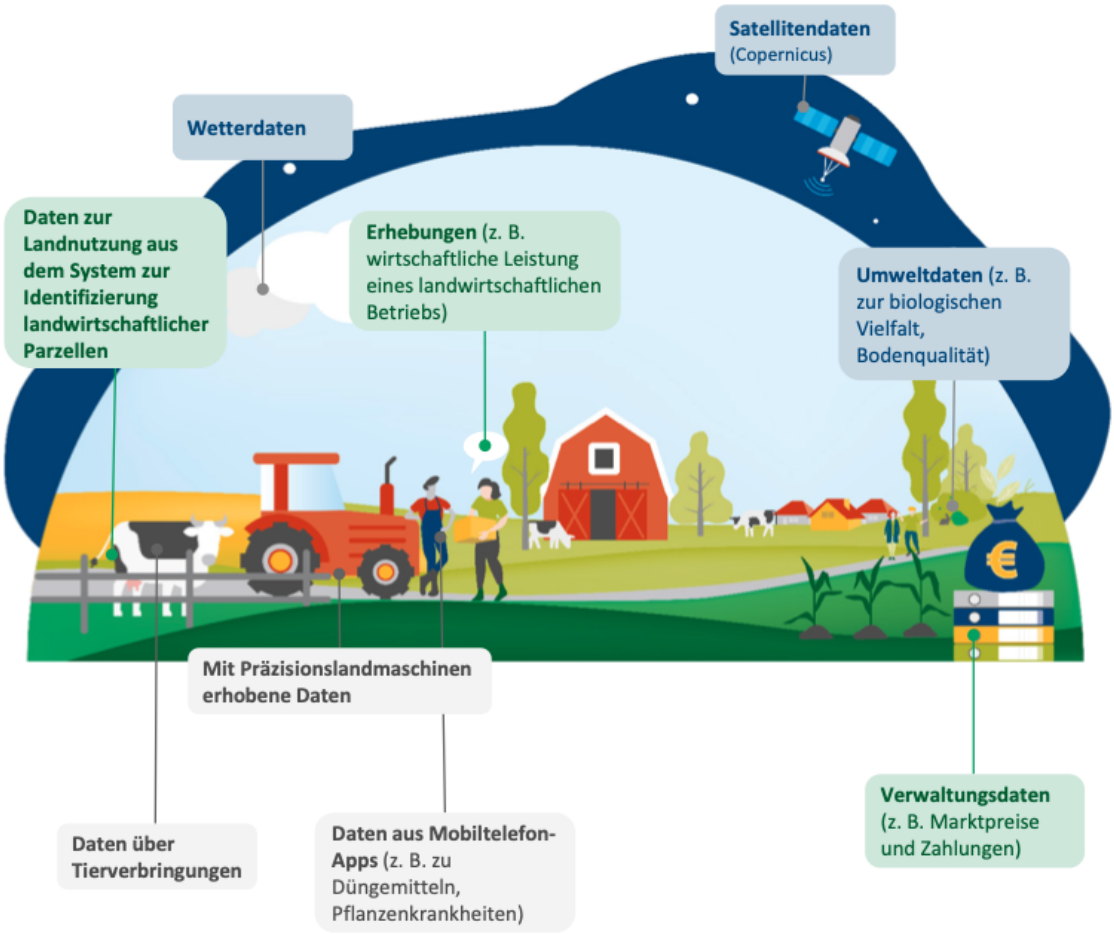
5.1.1.1 Gemeinsame Agrarpolitik der EU

Die Gemeinsame Agrarpolitik (GAP) der EU wurde bereits im Jahre 1962 eingeführt und soll sich nicht nur auf die landwirtschaftliche Erzeugung und die Landwirte, sondern auch auf umwelt- und klimabezogene sowie soziale Aspekte auswirken. In den GAP-Verordnungen sind die allgemeinen Ziele der Verträge[23] weiter spezifiziert.

Im Kontext der GAP erfolgt der Großteil der finanziellen Förderung für die Begünstigten auf Basis der bewirtschafteten landwirtschaftlichen Flächen. Darüber hinaus können Fördermittel als Kostenerstattung für die Umsetzung spezifischer Maßnahmen sowie zur Finanzierung von Investitionen bereitgestellt werden. Die rechtliche Grundlage für die meisten Zahlungen bildet das EU-Recht. Daten über landwirtschaftliche Betriebe werden dabei mittels unterschiedlicher Methoden erhoben und erstellt (Abbildung 12).

[23] https://eur-lex.europa.eu/legal-content/DE/TXT/?uri=celex:12012E/TXT

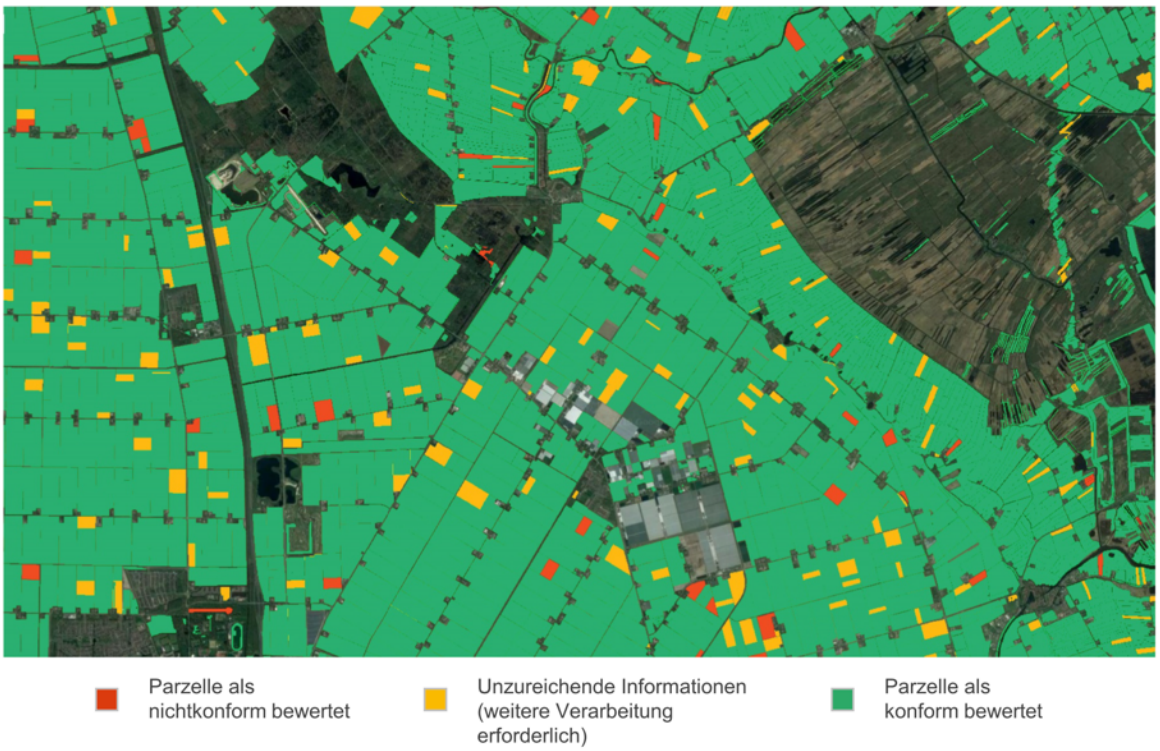
Abbildung 12: GAP-relevante Daten



Quelle: (Europäischer Rechnungshof, Daten in der Gemeinsamen Agrarpolitik: Potenzial von Big Data wird für die Zwecke der Politikbewertung nicht voll ausgeschöpft, 2022)

Eine zentrale Rolle spielen in diesem Kontext Sentinel-Satelliten der Europäischen Raumfahrtbehörde (ESA), welche im Rahmen des Copernicus Programms betrieben werden. Seit 2018 werden Copernicus-Daten eingesetzt, um zu überprüfen, ob bestimmte Landparzellen die Förderkriterien erfüllen, wodurch Feldbesichtigungen (Vor-Ort Kontrollen) ersetzt werden konnten (Europäischer Rechnungshof, 2020). Der Einsatz von KI-Methoden ist dabei essentiell, um abgeleitete Produkte aus den Copernicus-Rohdaten zu erzeugen. Bei Kontrollen durch Monitoring werden Satellitendaten mit den Angaben der Landwirte in ihren Anträgen zusammengeführt. Basierend auf ML-Verfahren erhalten die Zahlstellen für jede Beihilferegelung Informationen über die Art der Kulturen und die landwirtschaftlichen Tätigkeiten auf/in den einzelnen angemeldeten Parzellen/Betrieben. Diese Informationen sind die Basis für die Bewertung (Abbildung 13).

Abbildung 13: Simulation möglicher Ergebnisse bei einer Bewertung von Parzellen (Europäischer Rechnungshof, 2020)



Quelle: Simulation der JRC

Im Jahr 2021 wurden insgesamt 13,1% der Flächen, die Direktzahlungen erhielten, mittels Satellitendaten überprüft. Seit 2023 wird das Flächenmonitoringsystem mittels Copernicus-Satellitendaten standardmäßig in der EU eingesetzt.

5.1.1.2 Forest Stewardship Council® (FSC®)

Der FSC wurde 1993 gegründet und ist eine internationale Non-Profit-Organisation mit Sitz in Bonn. Er schuf das erste weltweite System zur Zertifizierung von nachhaltiger Forstwirtschaft. Stand Juli 2024 sind ca. 1,26 Mio. Hektar (11%) des deutschen Waldes nach FSC-Kriterien zertifiziert.

Im Rahmen seiner Akkreditierung sammelt der FSC eine erhebliche Menge an Daten, die sich auf Nachhaltigkeit, ethische Ressourcennutzung sowie Zertifizierungs- und Compliance-Vorgaben der Kunden beziehen. Darüber hinaus wird ein komplexer, mehrstufiger Bewertungsprozess durchlaufen, um systematisch zu prüfen, inwieweit die Kunden die festgelegten Kriterien für eine FSC-Zertifizierung erfüllen. Dieser strukturierte Ansatz gewährleistet eine umfassende Analyse und Bewertung der Einhaltung von Nachhaltigkeitsstandards.

Innerhalb des FSC ist das Systemintegritätsteam mit der Waldzertifizierung beauftragt. Um die Skalierbarkeit zu gewährleisten, wurden ML-Algorithmen eingesetzt, welche dabei helfen, den Zertifizierungsprozess zu beschleunigen (slalom, 2024). Dadurch konnten die Effizienz, Transparenz und Überwachungsmöglichkeiten des FSC wesentlich verbessert werden.

5.1.2 Erreichung der Nachhaltigkeitsziele

Daten sind die Grundlage, um den Fortschritt der SDGs zu messen und faktenbasierte strategische Entscheidungen zu treffen. Obwohl durch die Arbeit der Inter-Agency Expert Group on SDGs (IAEG-SDGs) die Methoden und Datenverfügbarkeit zur Messung der SDGs verbessert

werden konnten, übersteigen verschiedene Indikatoren nach wie vor die finanzielle und technische Kapazität vieler Länder (Nilashi, Keng Boon, Tan, Lin, & Abumalloh, 2023).

KI und die bessere Verfügbarkeit von annotierten Umweltdaten bzw. umweltrelevanten Daten haben daher ein wesentliches Potential, das Monitoring der SDGs zu unterstützen und letztendlich auch zur Erreichung der Nachhaltigkeitsziele beizutragen. KI bzw. ML-basierte Analysen können traditionelle Methoden ergänzen und unterstützen, um politischen Entscheidungsträgern*Entscheidungsträgerinnen dabei zu helfen, Bereiche zu identifizieren, die sich besonders für die Verfolgung des Fortschritts bei der Erreichung der SDGs eignen, und evidenzbasierte strategische Entscheidungen zu treffen, welche Maßnahmen Vorrang haben sollten (Nilashi, Keng Boon, Tan, Lin, & Abumalloh, 2023). Auf diese Weise kann garantiert werden, dass genügend Ressourcen und Kapazitäten zur Verfügung gestellt werden, um die SDGs effektiv und effizient umzusetzen.

5.2 Wissenschaftliches/technologisches Potential

5.2.1 Besseres Verständnis komplexer Systeme

ML hat sich in den Umweltwissenschaften und -techniken als leistungsfähiges Werkzeug erwiesen, ein besseres Verständnis für die komplexen Vorgänge und vielfältigen Wechselwirkungen der biochemischen Kreisläufe zu entwickeln (Huntingford, et al., 2019). Zudem bietet ML verbesserte Möglichkeiten der Datenanalyse, Vorhersage und Entscheidungsfindung (Binetti et al., 2024). Insbesondere sind ML-Algorithmen in der Lage, die Messwerte verschiedenster Sensoren (z.B. Drohnen- und Satellitenaufnahmen) gemeinsam auszuwerten. Die Integration von ML mit Big-Data-Analytik hat eine genauere und effizientere Verarbeitung komplexer Umweltdatensätze ermöglicht und zu Fortschritten in Bereichen wie Wettervorhersage, Katastrophenmanagement und Fernerkundung geführt (Maganathan, Senthilkumar, & Balakrishnan, 2020; Zhong, et al., 2021). Die Herausforderungen liegen jedoch weiterhin in der Datenqualität, der Interpretierbarkeit der Modelle und der Notwendigkeit von Datenannotationen auf Expertenbasis (Russo, et al., 2021; Liu, Lu, Zhang, Liu, & Jiang, 2022). Trotz dieser Hindernisse hat ML weiterhin ein großes Potenzial, die Umweltüberwachung und das Umweltmanagement zu revolutionieren, indem es neue Möglichkeiten für datengestützte Analysen und Entscheidungen bietet (Malakar, 2022). So diskutieren die Autoren den Einsatz von ML-Techniken, um aus der großen Menge an Umweltdaten, die von verschiedenen Sensoren und Fernerkundungsplattformen gesammelt werden, Erkenntnisse zu gewinnen und Anwendungen zu entwickeln. Dabei können ML-Techniken nützlich sein, um die Beziehungen zwischen Variablen in komplexen Umweltsystemen zu untersuchen, in denen die Beziehungen nicht linear sind und zudem einige Variablen noch unbeachtet sein können. Die Kombination von ML-Techniken und Umweltdaten aus der Fernerkundung und anderen Quellen bietet die Möglichkeit, sogenannte „Umweltintelligenz“-Ansätze zu entwickeln. Basierend darauf können Informationen für die Entscheidungsfindung und Politikberatung und -entwicklung bereitgestellt werden. Es besteht jedoch weiterer Forschungsbedarf, um die Anwendungen der Umweltintelligenz vollumfänglich zu entwickeln und insbesondere die relevanten Variablen für die Modellierung von Umweltsystemen zu identifizieren.

5.2.2 Optimierung von ML-Algorithmen für Umweltdatenanalyse und -überwachung

Bei der Optimierung von ML-Algorithmen kann der Schwerpunkt auf der Verbesserung des Algorithmus, der Verbesserung der Datenqualität oder auf beiden Aspekten liegen. Als komplementärer Ansatz zur modellzentrierten KI entwickelt sich die datenzentrierte KI. Diese konzentriert sich auf die Verbesserung der Datenqualität und nicht der Modellarchitektur, um

die Leistung von KI-Systemen zu steigern (Kumat, Datta, Singh, Singh, & Sharma, 2024; Hamid, 2022). Dieser Paradigmenwechsel erkennt die entscheidende Rolle qualitativ hochwertiger Daten bei der Entwicklung von KI an (Singh, 2023). Studien haben gezeigt, dass datenzentrierte Ansätze im Vergleich zu modellzentrierten Methoden eine höhere Genauigkeit erzielen können (La & La, 2023). Zu den wichtigsten Aspekten der datenzentrierten KI gehören die Bewertung der Datenqualität, die Vorverarbeitung, das Transferlernen und Machine Learning Operations (MLOps) (Singh, 2023). Der datenzentrierte Ansatz ist besonders empfehlenswert, wenn die Potentiale zur Erhöhung der Modellgenauigkeit weitestgehend ausgeschöpft und/oder die damit verbundenen Aufwände unverhältnismäßig sind, da eine Verbesserung der Datenqualität im Vergleich zur weiteren Modelloptimierung zu erheblichen Leistungssteigerungen führen kann (Bhowmik & Arabinda, 2021). Dieses Paradigma betont die systematische Verbesserung von Daten in großem Maßstab und ergänzt die modellzentrierte KI beim Aufbau effizienter KI-Systeme (Jakubik, Vossing, Kuhl, Walk, & Satzger, 2022).

Hochqualitative annotierte Datensätze ermöglichen durch eine verbesserte Modellgenauigkeit präzisere Analysen und können dadurch genauere Vorhersagen für verschiedenste Umweltereignisse erlauben. Dabei spielen die Qualität und Quantität der annotierten Daten eine entscheidende Rolle (Alhazmi, Walaa, Indrek, Lara, & Martin), wobei die Vorteile eines größeren Datensatzes oft die Auswirkungen kleinerer Annotationsfehler und Unausgewogenheit der Annotation überwiegen (Alhazmi, Walaa, Indrek, Lara, & Martin; Schmitt, Ahmadi, & Hansch, 2021). Größere Stichproben neigen dazu, zufällige Verzerrungen zu minimieren und repräsentativere Ergebnisse zu liefern. In solchen Fällen sind kleinere Fehler und Ungenauigkeiten möglicherweise vernachlässigbar oder akzeptabel.

5.3 Wirtschaftliches Potential

5.3.1 Kostenreduktion und Prozessoptimierung

ML bietet zahlreiche Methoden zur Kostenreduktion und Prozessoptimierung in verschiedenen umweltrelevanten Bereichen. Durch die Analyse großer Datenmengen, Mustererkennung und Vorhersagemodelle können ML-Algorithmen dazu genutzt werden, effiziente Lösungen zu entwickeln. Beispielsweise können in landwirtschaftlichen Bereichen Deep Learning (DL)-Algorithmen nicht nur zur genetischen Verbesserung von Nutzpflanzen beitragen (Wang, Cimen, Singh, & Buckler, 2020), sondern sind auch in der Lage, die enormen Datenmengen verschiedenster Sensoren (Fernerkundungsdaten, Sensormessungen) auszuwerten und auf diese Weise präzise Entscheidungen in Bezug auf Bewässerung, Nährstoffe, Krankheiten, Schädlinge und Erträge wesentlich zu unterstützen (Xu, et al., 2021). Dadurch können u.a. der Einsatz von Düngemitteln und Pestiziden optimiert (Ennaji, Vergütz, & El Allali, 2023), der optimale Erntezeitraum vorhergesagt (Sharma, Jain, Gupta, & Chowdary, 2020) oder Bewässerungsstrategien angepasst werden (Umutoni & Samadi, 2024)[24].

Im forstwirtschaftlichen Sektor stellen ML-Methoden ein leistungsfähiges Instrument dar, welches es ermöglicht, riesige Datensätze zu analysieren und aufschlussreiche Vorhersagen zu treffen, wodurch die Waldbewirtschaftung signifikant verbessert werden kann. Beispielsweise können mittels DL-Verfahren Informationen zur aktuellen Baumartenverteilung und zum Baumzustand auf Basis von hochaufgelösten Fernerkundungsdaten wie Luftbildern und Satellitenbildern abgeleitet werden und die Grundlage für Handlungsempfehlungen zu klimaangepassten, nachhaltigen Waldumbaustrategien liefern.

[24] Siehe auch Kapitel 4.1.2.1 für eine Zusammenfassung der landwirtschaftlichen Anwendungsbereiche.

Auch im Sektor Biodiversität haben ML-Algorithmen vielfältige Anwendungsmöglichkeiten und können dabei helfen, die enormen Datenmengen, welche durch Sensoren, Kamerafallen und Satelliten gesammelt und erhoben werden, auszuwerten (siehe auch Kapitel 4.1.1). In diesem Zusammenhang ist auf Thompson (2023, S. 232) zu verweisen, welcher Carl Chalmers zitiert und die Bedeutung von KI für den Naturschutz hervorhebt: *"Without AI, we're never going to achieve the UN's targets for protecting endangered species, (…)".*

5.3.2 Reduzierung von Personalnot und Fachkräftemangel

Der Fachkräftemangel ist eine komplexe Herausforderung und betrifft nahezu alle Sektoren mit teilweise globalen Auswirkungen. Verursacht u.a. durch demografischen Wandel, technologischem Fortschritt, herrschende Arbeitsbedingungen, Abwanderung von Arbeitskräften in andere Branchen und zunehmender Urbanisierung, führt der Fachkräftemangel beispielsweise in der Landwirtschaft zu Produktivitätseinbußen und limitierter Wettbewerbsfähigkeit.

KI kann eine Chance sein, um den Mangel an qualifizierten Fachkräften zu kompensieren und sowohl die Effizienz als auch die Produktivität zu steigern. Tabelle 18 gibt einen Überblick über mögliche Bereiche, in denen KI in den jeweiligen Sektoren dem Fachkräftemangel entgegenwirken kann.

Tabelle 18: Bereiche, in denen KI in den jeweiligen Sektoren unterstützen kann

Sektor	Bereich
Landwirtschaft	• Effizientere Bearbeitung von Feldern durch mit KI ausgestattete Landmaschinen • Optimierung des Einsatzes von Ressourcen (z.B. Personal, Düngemittel, Pestizide, Wasser) durch Präzisionslandwirtschaft • Entscheidungsunterstützungsinstrument für das effiziente Management von Feldern • Effiziente Planung von Maschinen und Personal • Optimierung der Lieferketten • Überwachung der Tiergesundheit, um frühzeitig Krankheiten zu erkennen, wodurch der Bedarf an spezialisierten Arbeitskräften für die tägliche Überwachung der Herden reduziert wird • KI-gesteuerte Fütterungssysteme können die Futtermenge und Zusammensetzung für jedes Tier individuell anpassen, was die Effizienz in der Tierhaltung erhöht und weniger Arbeitskräfte für die Fütterung erfordert
Forstwirtschaft	• Automatisierte Datenerfassung und Analyse mittels Satellitenbildern und Drohnen • Effiziente Planung von Maschinen und Personal • Vorhersagemodelle für Waldwachstum und Schädlingsbekämpfung als Entscheidungsunterstützungsinstrument für die langfristige strategische Planung zur klimaangepassten nachhaltigen Waldbewirtschaftung (vgl. FutureForest) • Unterstützung des Nachhaltigkeitsmanagement
Biodiversität	• KI-basiertes Artenmonitoring anhand von Fotos, Videoaufnahmen oder Audiodaten, um den Bedarf an menschlichen Expertinnen*Experten für die manuelle Datenerfassung zu reduzieren • Überwachung von großen Gebieten mittels Drohnen oder/und Satellitendaten • Effiziente Planung von Schutzmaßnahmen basierend auf Ergebnissen von KI-Vorhersagemodellen • Vorhersagen und Monitoring von Bedrohungen für die Biodiversität (z.B. Wilderer), um präventive Maßnahmen zu entwickeln oder aktiv einzugreifen

5.4 Soziales und ökologisches Potential

5.4.1 Beitrag zu den Nachhaltigkeitszielen

KI und ML-Methoden haben nicht nur das Potential bestehende Datenlücken zu füllen, sondern können auch einen direkten Beitrag zur Erreichung der Nachhaltigkeitsziele leisten (vgl. auch Kapitel 6.1.3). Durch eine bessere Verfügbarkeit qualitativ hochwertiger annotierter Daten und der damit verbundenen effizienteren Analyse großer Datenmengen (siehe Kapitel 5.3.1) kann das Monitoring der SDGs verbessert und unterstützt werden.

In ihrer Studie zeigten Cowls, Tsamados, Taddeo, & Floridi (2021), dass jedes SDG bereits innerhalb von mindestens einem KI-basierten Projekt behandelt wird, jedoch die Verteilung der Projekte nicht gleichmäßig auf die SDGs erfolgt (Abbildung 14).

Abbildung 14: KI-Projekte, die sich mit den SDGs befassen

Eine Stichprobe von 108 KI-Projekten, bei denen festgestellt wurde, dass sie sich weltweit mit den SDGs befassen.

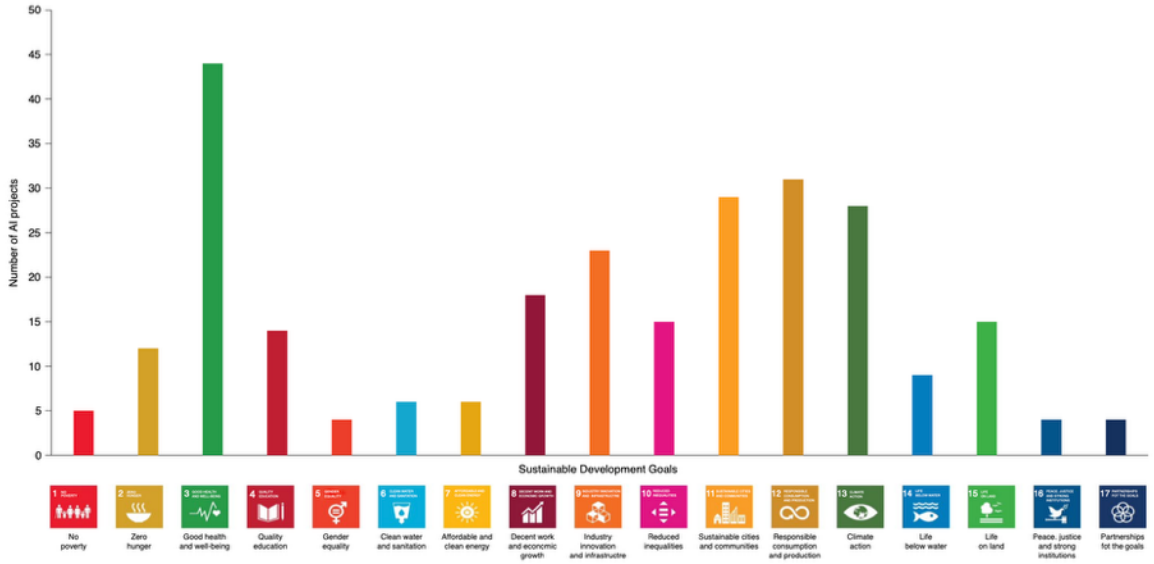


Fig. 1 | Projects addressing the SDGs. A survey sample of 108 AI projects found to be addressing the SDGs globally.

Quelle: (Cowls, Tsamados, Taddeo, & Floridi, 2021)

Verdeutlicht wird dies durch die umfangreiche Datenbank des AI4SDG Think Thank Observatory[25], welches aktuell 13.592 Anwendungsfälle listet. Ausgewählte Beispiele in Bezug auf ML-Methoden sind in Tabelle 19 dargestellt.

Tabelle 19: Beispiele für die Anwendung von ML-Methoden und Umweltdaten zur Unterstützung der Nachhaltigkeitsziele

SDG	Beispielstudie	Quelle
SDG1: Keine Armut	In ihrer Studie nutzten Jean et al. (2016) Convolutional Neural Networks zur Schätzung von Konsumausgaben und Vermögen anhand von hochauflösenden Satellitendaten in 5 afrikanischen Ländern. Die Methode erlaubte es 75 % der Variation der wirtschaftlichen Ergebnisse auf lokaler Ebene zu erklären.	(Jean, et al., 2016)

[25] https://www.ai-for-sdgs.academy/case/134

SDG	Beispielstudie	Quelle
SDG2: Kein Hunger	ML-Methoden werden insbesondere für die Identifizierung, die Auswahl und den Schutz von Pflanzen, die Überwachung des Pflanzenwachstums, die Maximierung und Optimierung von Erträgen, die Empfehlung von Düngemitteln und die Verbesserung der Bodenqualität eingesetzt und tragen damit positiv zu SDG2 bei. Zudem wird ML für die nachhaltige Tierhaltung eingesetzt (z. B. Überwachung der Gesundheit und des Wohlbefindens der Tiere, Vorhersage von Krankheiten und Optimierung von Nährstoffzufuhr, -aufnahme und -verbrauch)	(Sharma, Jain, Gupta, & Chowdary, 2020) (Ennaji, Vergütz, & El Allali, 2023)
SDG6: Sauberes Wasser und Sanitärversorgung	KI und ML-Ansätze werden zur Erkennung von Wasserverschmutzung und für intelligente Wassermanagement-Systeme (SWM) verwendet. ML wird u. a. für die Überwachung und Vorhersage der Wasserqualität eingesetzt.	(Taweesan, Koottatep, & Polprasert, 2024)
SDG7: Bezahlbare und saubere Energie	Neben dem Ausbau von erneuerbaren Energien, können KI und ML dazu beitragen, den Energieverbrauch durch eine Steigerung der Energieeffizienz deutlich zu senken. Dies hat positive Implikationen auf das Klima und die Umwelt.	(Yao, et al., 2023)
SDG13: Maßnahmen zum Klimaschutz	Xu et al. (2024) zeigten, dass Düngemittelmanagement und Bodenbearbeitungspraktiken, die auf lokaler Ebene mithilfe von maschinellem Lernen und Big Data optimiert wurden, die Emissionen aus Düngemitteln um bis zu 38 % reduzieren könnten. Daher könnte der Einsatz von KI bei erfolgreicher Anwendung einen positiven Beitrag zu SDG2 (Ziel 2.4) und SDG13 (Ziel 13.2) leisten.	(Xu, et al., 2024)

5.4.2 Nachhaltige Konsumentscheidungen

In der Digitalagenda des BMUV wird bereits das Zielbild von KI-Anwendungen für den nachhaltigen Konsum definiert und deren Potential beschrieben:

„Nachhaltiger Konsum in der digitalen Welt geht mit gestärkter Selbstbestimmung und Kompetenz der Verbraucherinnen und Verbraucher einher. Vertrauenswürdige Informationen zu den Produkten, Dienstleistungen und ihren Umweltwirkungen sind leicht verfügbar und intelligente Assistenzsysteme erleichtern nachhaltige Konsumentscheidungen im Alltag. Für Plattformen und Onlineangebote gelten klare Regeln. Mehr Transparenz schafft neue Anreize zur Ausweitung des Angebots an nachhaltigen Produkten und Dienstleistungen.“ (BMUV, 2020, S. 19)

KI- und ML-Anwendungen können dazu beitragen, umweltpolitische Prozesse anzustoßen, um die Märkte nachhaltiger und umweltfreundlicher zu gestalten. Zudem können diese Initiativen fördern, dass auch der Handel umweltgerechte Kommunikations-, Informations- sowie Vermarktungsstrategien entwickelt und umsetzt. Wichtig dabei sind stets die Transparenz und verlässliche Informationen zum Herstellungsprozess und zum Ursprung der Produkte. Regulatorische Maßnahmen sind dort notwendig, wo Verbraucherinnen und Verbrauchern diese Informationen im Alltag oft nur unzureichend oder mit erheblichem Aufwand zur Verfügung stehen. Laut BMUV sind Transparenz sowie die Einführung eines digitalen Produktpasses zentrale Elemente für informierte Konsumentscheidungen. Ziel wäre es, den nachhaltigen Konsum für die breite Bevölkerung ohne Hürden zugänglich zu machen. (BMUV, 2020)

6 Wirkungsanalyse

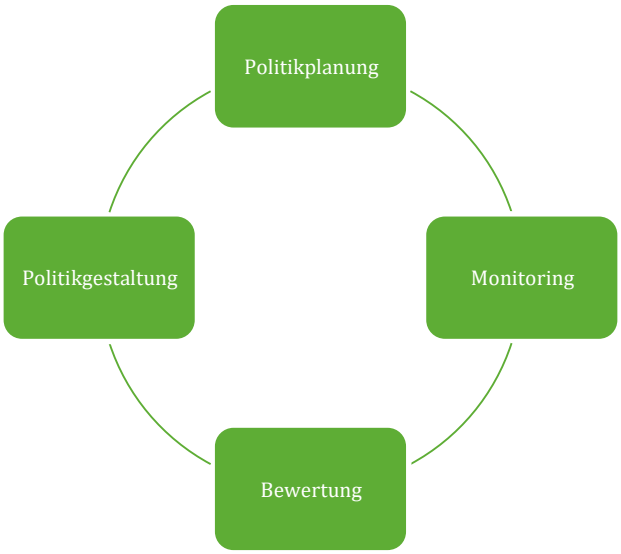
Trotz der vielversprechenden Potentiale haben KI- bzw. ML-Anwendungen unterschiedliche Wirkungen, welche betrachtet werden sollten, bevor diese eingesetzt werden. Positive und negative Wirkungen sind einander abzuwägen und im individuellen Fall zu entscheiden, ob KI die wirksamste Lösung für ein bestimmtes Problem darstellt.

6.1 Positive Wirkungen

6.1.1 Unterstützung von datenbasierten Entscheidungen und Politikumsetzung

In den Leitlinien für eine bessere Rechtsetzung[26] spricht sich die Europäische Kommission für einen evidenzbasierten Ansatz aus. Dadurch sollen politische Entscheidungen auf einer soliden Datengrundlage, insbesondere quantitativer und qualitativer Daten, basieren. Qualitätsgeprüfte annotierte Daten können daher einen wesentlichen Beitrag für faktenbasierte politische Entscheidungen bilden und sind in allen Ebenen eines Politikzyklus wertvoll (siehe Abbildung 15).

Abbildung 15: Etappen eines Politikzyklus



Quelle: (Europäischer Rechnungshof, 2022)

- **Politikplanung:** Die Regelungen für das Monitoring sowie die Bewertung und der Bedarf an Daten werden im Zuge der Politikplanung definiert.
- **Monitoring:** Die Behörden sammeln hauptsächlich quantitative Daten für Indikatoren, die während der Politikentwicklungsphase festgelegt wurden.
- **Bewertung** basieren häufig auf empirischen Analysen der erhobenen Daten, die mithilfe quantitativer Methoden und Modellierung durchgeführt werden. Diese Analysen werden durch qualitative Informationen ergänzt, die beispielsweise durch Befragungen und Umfragen gewonnen werden.

[26] https://commission.europa.eu/document/d0bbd77f-bee5-4ee5-b5c4-6110c7605476_en

- **Politikgestaltung:** Alle relevanten Daten werden gesammelt und im Zuge einer Folgeabschätzung analysiert. Die Ergebnisse werden für die Politikgestaltung verwendet.

6.1.2 Risikominimierung bzw. -vermeidung

Annotierte Daten spielen eine zentrale Rolle bei der Risikovermeidung und Risikominimierung im Umweltsektor, insbesondere im Zusammenspiel mit ML-Verfahren.

Im forstwirtschaftlichen Sektor ermöglicht die gezielte Erhebung und Annotation von Daten über lokale Klima- und Bodenbedingungen, Baumartenverteilung und Wachstumsraten die Entwicklung von klimaangepasstem Waldbau. Auf dieser Grundlage können ML-Modelle Empfehlungen zur Auswahl geeigneter Baumarten geben, die widerstandsfähiger gegenüber klimatischen Veränderungen sind, wodurch das Risiko von Waldschäden, wie z.B. durch Dürren oder Schädlingsbefall, erheblich verringert wird[27].

In der Landwirtschaft unterstützt die Annotation von Daten ebenfalls das Risikomanagement, etwa im Bereich des Tierwohls. Moderne ML-Anwendungen nutzen tierbezogene Daten, um Anzeichen von Stress oder untypischem Verhalten frühzeitig zu erkennen, was die Früherkennung von Krankheiten erleichtert. So können Landwirtinnen*Landwirte präventive Maßnahmen ergreifen und die Gesundheit der Tiere langfristig sichern, was nicht nur das Risiko von Produktionsausfällen mindert, sondern auch ethische und wirtschaftliche Vorteile bietet.

Im Bereich der Biodiversität sind annotierte Daten entscheidend für das Management ökologischer Risiken (Branco, Correia, & Cardoso, 2023). So helfen ML-gestützte Systeme, Wilderer zu identifizieren und damit den Verlust gefährdeter Arten zu verhindern (Thompson, 2023). Zudem ermöglichen sie die Überwachung der Migration von Arten in Echtzeit, was besonders in Zeiten des Klimawandels wichtig ist, da Tiere und Pflanzen ihre Lebensräume verlagern. Auch das Management invasiver Arten wird durch annotierte Daten verbessert: Frühwarnsysteme können Gebiete identifizieren, in denen invasive Arten ökologischen oder wirtschaftlichen Schaden anrichten könnten, sodass entsprechende Gegenmaßnahmen rechtzeitig ergriffen werden können (GPAI, 2022).

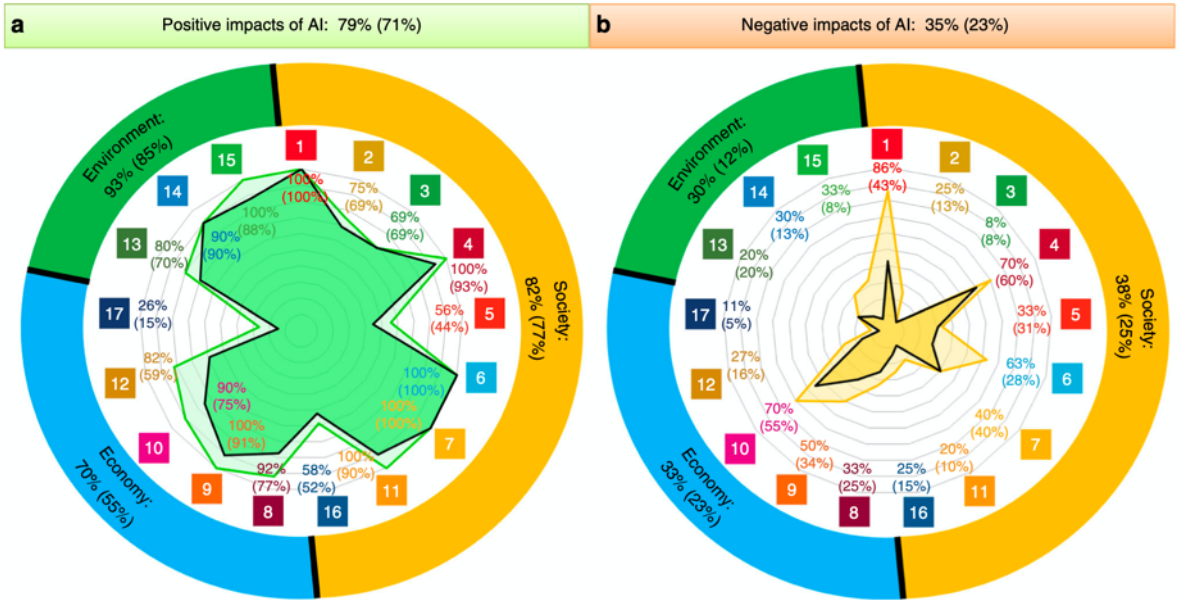
In all diesen Bereichen ermöglichen annotierte Daten die präzise Analyse und Vorhersage komplexer Umweltprozesse, was entscheidend ist, um potenzielle Risiken frühzeitig zu erkennen und wirksame Maßnahmen zur Minimierung von Umweltschäden zu ergreifen.

6.1.3 Förderung der Nachhaltigkeitsziele

In ihrer Studie stellen Vinuesa et al. (2020) fest, dass KI-Methoden insgesamt 134 (79%) der Unterziele aller Nachhaltigkeitsziele fördern können. Die Autoren räumen jedoch ein, dass KI 59 Unterziele negativ beeinflussen kann (siehe Kapitel 6.2.1). Die Zusammenfassung der positiven und negativen Auswirkungen ist in Abbildung 16 visualisiert.

[27] Siehe Projekt FutureForest -KI für den Wald https://future-forest.eu

Abbildung 16: Zusammenfassung der positiven und negativen Auswirkungen von KI auf die verschiedenen SDGs



Quelle: Vinuesa et al. (2020).

In Bezug auf die positiven Wirkungen werden durch KI insgesamt 67 Unterziele (82%) der Dimension Soziales gefördert. Die Autoren postulieren, dass der Einsatz von KI als Wegbereiter für alle Ziele dieser Dimension dienen kann, indem dieser die Versorgung der Bevölkerung mit Nahrung, Gesundheit, Wasser und Energie unterstützt. Für die wirtschaftliche Dimension identifizieren Vinuesa et al. (2020) insgesamt 42 Unterziele (70%), die positiv durch KI beeinflusst werden. Die positive Wirkung ist insbesondere auf die Steigerung der Produktivität durch die Anwendung von KI in verschiedensten Sektoren zurückzuführen. Für 25 Unterziele (93%) der Umweltdomäne wirkt KI fördernd, indem u.a. komplexe und heterogene Daten analysiert und dadurch evidenzbasierte Entscheidungen zum Schutz der Umwelt getroffen werden können. Beispielsweise können KI-basierte Modelle dazu beitragen, den Klimawandel und seine Auswirkungen besser zu verstehen (Callaghan, et al., 2021). Mittels neuronaler Netzwerke und DL-Algorithmen können riesige Datenmengen aus der Fernerkundung klassifiziert werden, um z.B. Landbedeckungsänderungen zu analysieren (Schiller, Költzow, Schwarz, Schiefer, & Fassnacht, 2024). Auch der Schutz und das Monitoring der Biodiversität wird nachweislich durch KI bzw. ML-Anwendungen positiv beeinflusst (Thompson, 2023).

6.2 Mögliche negative Wirkungen bzw. Risiken

Neben den vielversprechenden Potentialen, die KI- bzw. ML-Methoden heben können, und den positiven Wirkungen, sind auch die negativen (direkten und indirekten) Wirkungen sowie die Risiken, welche der Einsatz von KI mit sich bringt, zu berücksichtigen.

6.2.1 Negativer Einfluss auf Nachhaltigkeitsziele

Neben den vielen positiven Aspekten verweisen Vinuesa et al. (2020) auch auf die negative Wirkung von KI auf die SDGs. Die Autoren stellen fest, dass insgesamt 59 Unterziele (35%) negativ durch KI beeinflusst werden. Den größten Anteil nimmt dabei die Dimension Soziales mit 31 Unterzielen (38%) ein. Dies betrifft insbesondere kulturelle Werte und den Wohlstand. Problematisch erscheint insbesondere der hohe Energiebedarf von KI (siehe Kapitel 6.2.2),

welcher sich vor allem auf SDG 7 (Bezahlbare und saubere Energie) und SDG 13 (Maßnahmen zum Klimaschutz) negativ auswirkt.

Für die wirtschaftliche Dimension wurden 20 Unterziele (33%) ermittelt, die durch KI negativ beeinflusst werden. Dies ist insbesondere auf die daraus resultierende Ungleichheit zurückzuführen, welche Implikationen auf SDG 8 (Menschenwürdige Arbeit und Wirtschaftswachstum), SDG 9 (Industrie, Innovation und Infrastruktur) und insbesondere SDG 10 (Weniger Ungleichheiten) hat.

In Bezug auf die Dimension der Umwelt identifizieren Vinuesa et al. (2020) insgesamt 8 Unterziele, für welche KI limitierend wirkt. Hauptursache dafür sind einerseits der hohe Energieverbrauch (siehe Kapitel 6.2.2.1) und andererseits die Annahme, dass KI-basierte Informationen über Ökosysteme die Ausbeutung dieser fördern kann (siehe Kapitel 6.2.3).

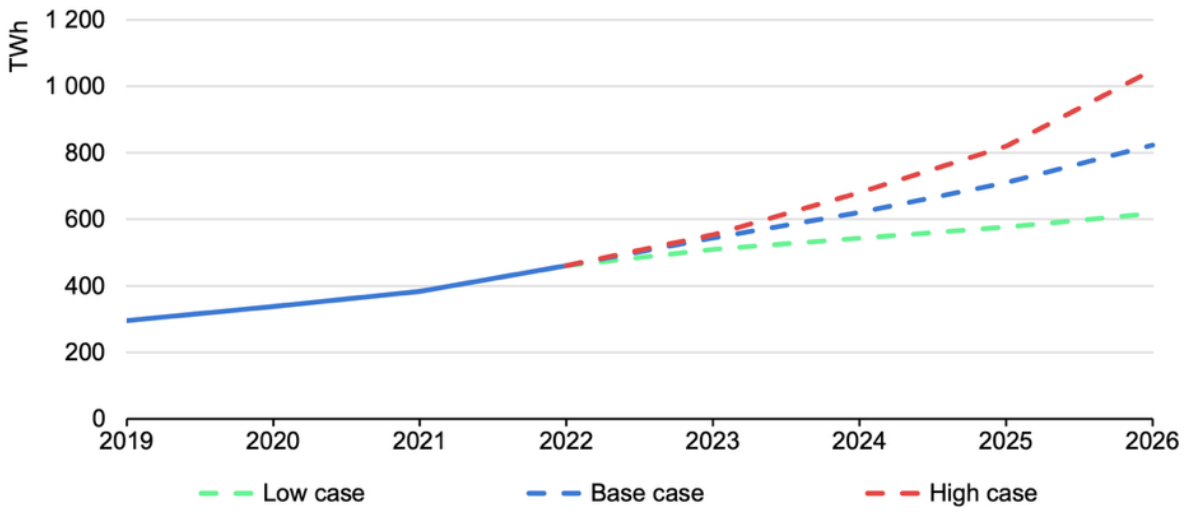
6.2.2 Ressourcenverbrauch und Emissionen

6.2.2.1 Energieverbrauch von Rechenzentren

KI-Technologien erfordern in den meisten Fällen sehr hohe Rechenressourcen, die meist nur in großen Rechenzentren zur Verfügung stehen. Diese Einrichtungen haben einen sehr hohen Energiebedarf und ökologischen Fußabdruck in Bezug auf ihren Kohlendioxidausstoß.

Laut einer Weltbank-Studie zu den Emissionen und dem Energiefußabdruck des IKT-Sektors ist dieser für insgesamt 1,7 Prozent der globalen Emissionen verantwortlich (The World Bank and ITU, 2024). Dabei sind Rechenzentren im Kontext digitaler Transformationsprozesse wesentliche Treiber für den zunehmenden Strombedarf in vielen Regionen. Der europäische Markt für Rechenzentren wuchs im 1. Quartal 2024 im Vergleich zum Vorjahr um fast 20 %. In allen vier großen FLAP-Märkten (Frankfurt, London, Amsterdam und Paris) gab es eine signifikante Entwicklung, wobei Paris mit einem Wachstum von über 40 % im Jahresvergleich führend war (CBRE, 2024). Diese Entwicklung ist nicht zuletzt auf KI zurückzuführen. Daten der Internationalen Energieagentur (IEA) zeigen, dass 2022 Rechenzentren weltweit nahezu 460 TWh an Strom bzw. 2% des globalen Strombedarfs verbrauchten (IEA, 2024). Die IEA schätzt, dass sich der Strombedarf bis 2026 um 160 TWh (low case) bzw. 460 TWh (high case) erhöhen wird (siehe Abbildung 17).

Abbildung 17: Weltweiter Energiebedarf von Rechenzentren, KI und Kryptowährung (2019-2026)



Quelle: IAE (2024)

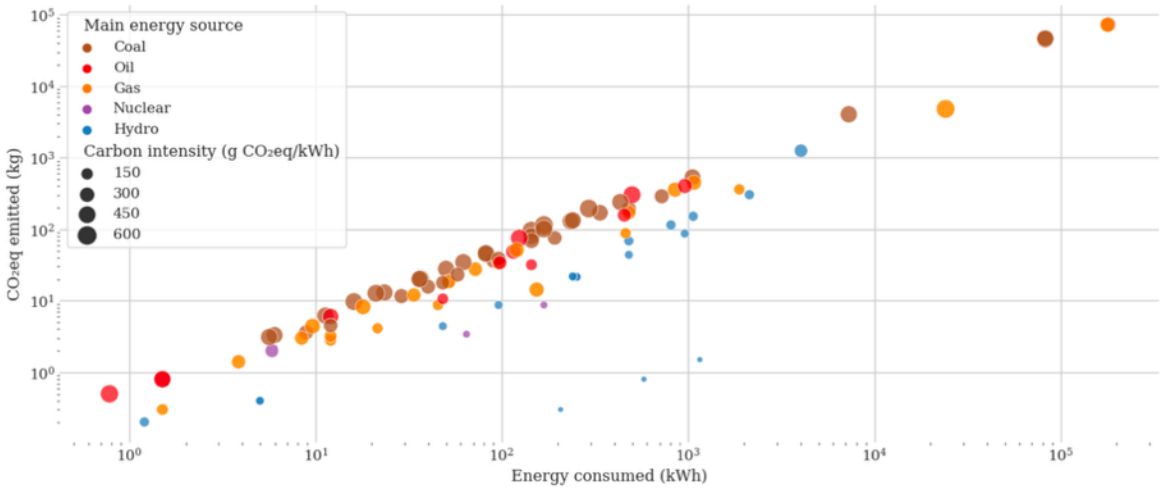
Karnama et al. (2019) setzen daher auf sogenannte „Organic Data Centers“ (OGC), welche nicht nur nachhaltig sind, sondern nachweislich eine positive Wirkung auf die Nachhaltigkeitsziele haben (Karnama, Haghighi, & Vinuesa, 2019).

6.2.2.2 Emissionen von ML-Modellen

Nicht nur der hohe Energieverbrauch, sondern auch die damit einhergehenden hohen Emissionen der ML-Modelle sind als negative Wirkung zu sehen. Beispielsweise kann das Trainieren eines großen Sprachmodells bis zu 315 Tonnen CO_2 verursachen (Ganesh, McCallum, & Strubell, 2019). Des Weiteren verglichen Luccioni & Hernandez-Garcia (2023) die CO_2-Emissionen von 95 ML-Modellen in Bezug auf die verwendete Energiequelle, die Höhe der CO_2-Emissionen, deren zeitliche Entwicklung und ihre Modellleistung (Luccioni & Hernandez-Garcia, 2023). Energiebedarf des Modells und dessen CO_2-Emissionen stehen in einer nahezu linearen Beziehung zueinander. Der Energiebedarf der Modelle reicht dabei von 10 kWh bis über 10.000 kWh. Zudem zeigen die Autoren, dass Modelle, die auf nachhaltigere Energiequellen zurückgreifen, weniger CO_2 emittieren als beispielsweise Modelle, die aus Kohle erzeugten Strom als Energiequelle nutzen (Abbildung 18).

Abbildung 18: Geschätzter Energieverbrauch (kWh) und CO2 (kg) jedes Modells im Datensatz

Darstellung auf einer logarithmischen Skala. Die Farben geben die Hauptenergiequelle und die Größe des Punktes die Kohlenstoffintensität an.



Quelle: (Luccioni & Hernandez-Garcia, 2023)

Außerdem untersuchten Luccioni & Hernandez-Garcia (2023), wie die CO_2-Emissionen der verschiedenen Modelle sich über die Zeit verändern. Die Autoren kommen zu dem Schluss, dass es eine hohe Variabilität gibt und kein Trend erkennbar ist. Jedoch war ein Anstieg der Emissionen von durchschnittlich 487 Tonnen CO_2eq für Modelle der Jahre 2015-2016 auf durchschnittlich 2020 Tonnen für Modelle der Jahre 2020-2022 zu verzeichnen. Es daher davon auszugehen, dass durch immer größere KI-Anwendungen zukünftig mit weiter steigenden Energie- und Ressourcenverbräuchen zu rechnen ist.

6.2.2.3 CO_2 Emissionen der Hardware

Auch der zunehmende Einsatz von KI bzw. ML in nahezu allen Bereichen führt dazu, dass mehr Hardware eingesetzt werden muss und neue Rechenzentren gebaut werden (Savills, 2024). Tabelle 20 gibt einen Überblick des CO_2-Fußabdruckes verschiedener Endgeräte.

Tabelle 20: Treibhausgasemissionen bei der Herstellung, bezogen auf 1 Jahr Nutzungsdauer[28]

Endgerät	Treibhausgasemission
Fernseher	248 kg CO_2e pro Jahr
Laptop	81 kg CO_2e pro Jahr
Desktop Computer	110 kg CO_2e pro Jahr
Smartphone	46 kg CO_2e pro Jahr
Router	68 kg CO_2e pro Jahr

Dabei ist zu beachten, dass dies auch soziale Implikationen hat. Der Großteil der Geräte wird in Ländern mit schlechten Umwelt- und Sozialstandards und unter teils menschenunwürdigen Bedingungen hergestellt. Dadurch kommt es zu einer Verknüpfung von sozialen und ökologischen Aspekten (Jungblut, 2024).

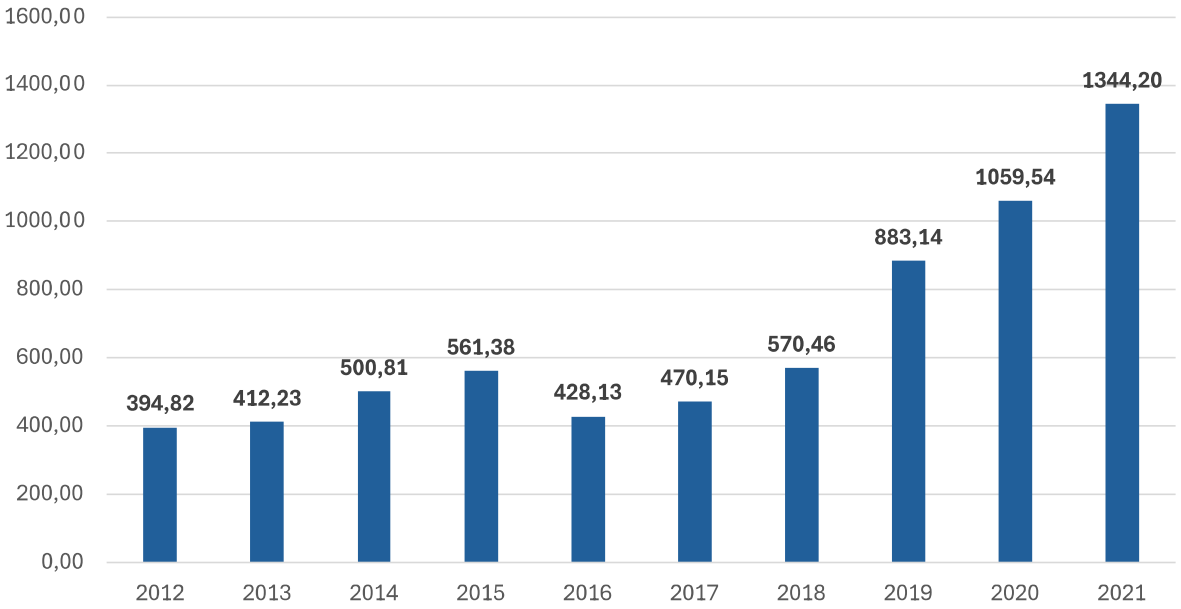
6.2.2.4 Wasserfußabdruck von KI-Systemen

Während Energieverbrauch und CO_2-Emissionen von Rechenzentren bereits seit einiger Zeit im Fokus stehen, wurde dem Wasserverbrauch nur wenig Beachtung geschenkt. Für die Kühlung der Rechenzentren wird meist Trinkwasser bzw. alternativ auch Grauwasser bzw. Passivkühlung benutzt. Insbesondere durch den Verbrauch riesiger Mengen an Trinkwasser kann es dabei zu Nutzungskonflikten kommen (Lizarraga & Solon , 2023; Arellano, Cifuentes, & Rios, 2020)

Abbildung 19 zeigt den jährlichen Wasserverbrauch von Google in The Dalles, welcher nahezu 25% des gesamten Wasserverbrauchs der Stadt ausmacht. Auffallend dabei ist der stetige Anstieg im Verbrauch, welcher nicht zuletzt auf die zunehmende Popularität von KI-Anwendungen zurückzuführen ist (Muhammad, 2022). Insbesondere „Generative AI“ verursachte einen erneuten Anstieg des Verbrauches um 22% bei Google und 34% bei Microsoft im Jahre 2022 (Marx, 2024). Als Beispiel kann das Training von GPT-3 genannt werden, welches durchschnittlich 700.000 Liter sauberes Süßwasser direkt verdampfen lässt (Li, Yang, Islam, & Ren, 2023).

[28] Daten stammen vom *CO2-Fußabdruck-Rechner: https://www.digitalcarbonfootprint.eu/*

Abbildung 19: Jährlicher Wasserverbrauch von Google in The Dalles (2012 – 2021). Angabe in Millionen Liter

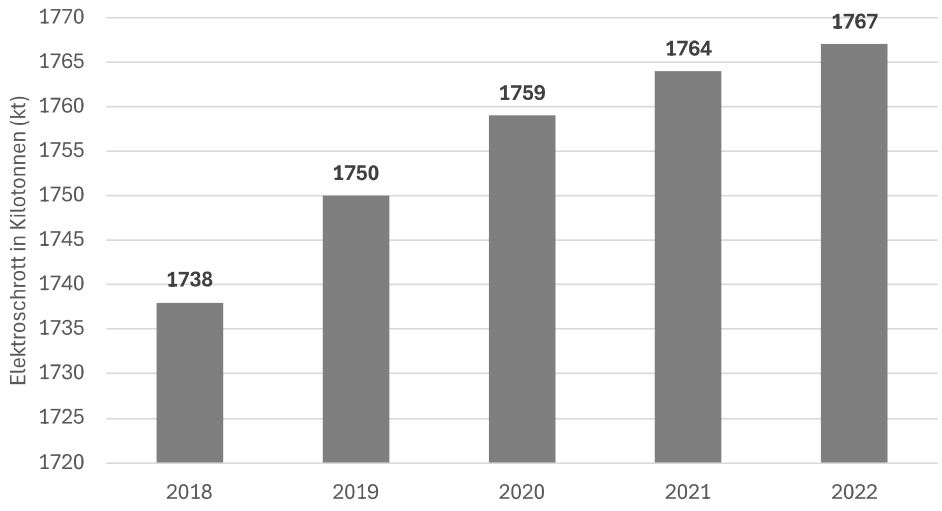


Quelle: (Muhammad, 2022)

6.2.2.5 E-Waste

Der im März 2024 erschiene E-Waste Monitor der Vereinten Nationen dokumentiert, dass die weltweite Erzeugung von Elektroschrott fünfmal mehr steigt als das dokumentierte Recycling. Im Jahre 2022 wurden weltweit 62 Millionen Tonnen an Elektroschrott erzeugt und es wird erwartet, dass dieser Wert auf bis zu 82 Millionen Tonnen im Jahr 2030 steigt, was einen Anstieg von 33% entspricht. Der globale Elektroschrott des IT-Sektors (z.B. Laptops, GPS-Geräte und Router) wurde dabei auf ca. 4,6 Millionen Tonnen geschätzt, von welchen nur 22% recycelt wurden (Baldé, et al., 2024). Aktuell trägt Deutschland dazu jedes Jahr etwa 1,7 Millionen Tonnen bei. Wie Abbildung 20 zeigt, ist die Tendenz dabei steigend (Baldé, et al., 2024).

Abbildung 20: Produzierter Elektroschrott von Deutschland für die Jahre 2018 bis 2022. Angabe in Kilotonnen (kt)

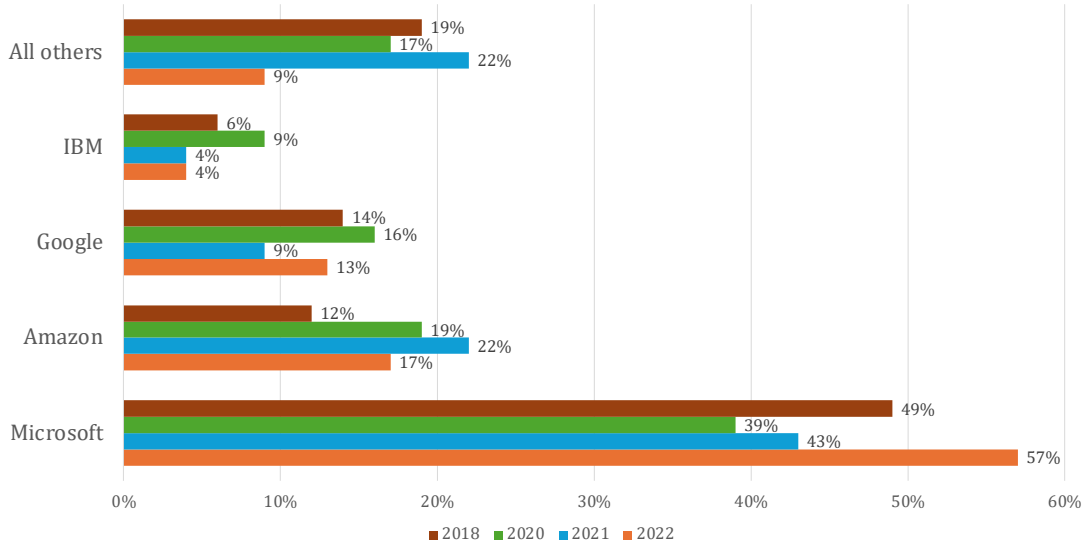


Quelle: (Baldé, et al., 2024)

6.2.3 Ausbeutung der Ökosysteme

In ihrer Studie postulieren Vinuesa et al. (2020), dass KI-bezogene Informationen über Ökosysteme zu einer übermäßigen Ausbeutung von Ressourcen führen können (Vinuesa, et al., 2020). KI-Methoden, welche die Kosten von erneuerbaren Energien senken, können auch die Förderung von Rohöl profitabler machen und die Erschließung neuer Ölfelder wesentlich beschleunigen (Gotsch, et al., 2022). Beispielweise bietet Amazon für Kunden aus der Öl- und Gasindustrie einen Service, um basierend auf ML das nächste Ölfeld vorherzusagen (Peranandam , 2018). Auch Microsoft bietet entsprechende Dienste an, um Vertretungen aus der Öl- und Gasindustrie das Potential von KI und ML-Methoden zu demonstrieren (Microsoft, 2018). Eine Studie von Kimberlite Oilfield Research zeigt, dass Microsoft der führende Cloud-Anbieter für die Öl- und Gasindustrie ist (siehe Abbildung 21) (Kimberlite, 2024).

Abbildung 21: Führende Cloudanbieter der Öl- und Gasindustrie



Quelle: Kimberlite (2024)

Aus diesen Beispielen wird klar ersichtlich, dass wirtschaftliche Interessen dominieren und dadurch auch Aktivitäten gefördert werden, die weder gemeinwohlorientiert noch nachhaltig sind.

6.2.4 Ethische Aspekte

6.2.4.1 Förderung der Ungleichheit

Vinuesa et al. (2020) argumentieren, dass der Einsatz von KI insbesondere die soziale Ungleichheit sowohl weltweit als auch innerhalb eines Landes weiter fördern kann. Dies betrifft beispielsweise den Datenzugang und die Infrastruktur, die Konzentration von Ressourcen sowie den Zugang zu KI-Technologien selbst. In ihrer Studie beziehen sich Vinuesa et al. (2020) auf die starke Abhängigkeit zukünftiger Märkte von Datenanalysen, welche in Ländern mit niedrigen und mittleren Einkommen nicht gleichermaßen verfügbar sind. Die Autoren stellen weiterhin fest, dass die wachsende wirtschaftliche Bedeutung von KI aufgrund der ungleichmäßig verteilten Bildungs- und Computerressourcen weltweit zu größeren Ungleichheiten führen kann.

6.2.4.2 Ausbeutung der „Click-Worker“

Click-Worker, insbesondere in Niedriglohnländern, spielen eine zentrale Rolle beim Annotieren von Daten und der Inhaltsmoderation für KI-Systeme[29]. Trotz der essenziellen Bedeutung ihrer Arbeit für die Entwicklung und Verbesserung von KI-Modellen werden sie häufig unter prekären Bedingungen beschäftigt und nur geringfügig entlohnt. Verschiedene Berichte verdeutlichen die Herausforderungen und Missstände, mit denen Click-Worker konfrontiert sind (Perrigo, 2023); (Perrigo, 2022). Beispielweise zeigt der Bericht von AlgorithmWatch (Algorithmwatch, 2023), dass Click-Worker oft als „unsichtbare Arbeitskräfte“ agieren, indem sie Inhalte annotieren oder moderieren und somit die Grundlage für das Training von KI-Algorithmen schaffen. Ihre Arbeit ist jedoch häufig befristet, schlecht bezahlt und führt aufgrund der repetitiven und psychisch belastenden Natur der Aufgaben oft zu Burnout und psychischen Gesundheitsproblemen. Die Arbeitsbedingungen, die soziale Absicherung und der Zugang zu psychologischer Unterstützung sind in vielen Fällen unzureichend, obwohl diese Arbeit für die KI-Industrie von wesentlicher Bedeutung ist.

[29] Inhaltsmoderation für KI-Systeme bezeichnet den Prozess, durch den Inhalte, die von oder für KI-Systeme bereitgestellt, analysiert oder generiert werden, überwacht, gefiltert und reguliert werden. Ziel der Inhaltsmoderation ist es, sicherzustellen, dass KI-Systeme ethisch, sicher und gesetzeskonform arbeiten.

7 Politische Handlungsempfehlungen

7.1 Förderung der Datenkompetenz und Etablierung von Datenmanagement-Richtlinien

Eine gezielte Förderung der Datenkompetenz im Umweltbereich ist ein wichtiger Baustein für den nachhaltigen Umgang mit Umweltressourcen. Auch heute sind in der Industrie, bei den Verwaltungen und in der Forschung immer noch zahlreiche Datensätze ausschließlich lokal auf Arbeitsplatzrechnern bzw. innerhalb der Organisation verfügbar. Dies reduziert nicht nur deren Zugänglichkeit (vgl. Kapitel 4.3.1.3), es erhöht aufgrund mangelnder Kontrollstrukturen auch die Risiken für unabsichtliche oder unrechtmäßige Weitergabe und -verarbeitung und damit der Verletzung der Rechte Dritter.

Zur Datenkompetenz gehört u.a. der verantwortungsvolle Umgang mit Daten, welcher sicherstellt, dass der Rechtsrahmen eingehalten wird, aber auch, dass die Daten beispielsweise für eine spätere Nachvollziehbarkeit wissenschaftlicher Arbeiten zugänglich und überprüfbar sind. Daten stellen oft ein teuer erworbenes oder hergestelltes Gut dar und sind somit gegen Datenverlust, Verringerung der Datenintegrität sowie gegen andere Gefahren schützenswert.

Entsprechende Richtlinien und Empfehlungen zum Datenmanagement existieren seit vielen Jahren (z.B. (DFG, 2019)), werden jedoch oft inkonsistent angewendet. Hier ist ein starker Ausbau der Datenmanagementkompetenzen zu empfehlen. Zudem ist es wichtig, Personen in den Richtlinien zu schulen und dann diesen Richtlinien einen wesentlich höheren Grad an Verbindlichkeit einzuräumen.

7.2 Etablierung von Datenräumen und Datentreuhändern als digitale Ökosysteme

Es gibt immer noch zahlreiche Silos, abgeschlossene Plattformen sowie projektzentrische Repositories, deren langfristige Perspektive unklar ist. Viele der aktuellen Probleme im Hinblick auf dauerhafte Dateninfrastrukturen würden sich durch den Aufbau offener Datenökosysteme, wie beispielsweise Datenräumen nach International Data Spaces Association (IDSA)- und Gaia-X-Standards, lösen lassen (vgl. Kapitel 4.4.3.2). Solche Ökosysteme sind für datenhaltende und datennutzende Organisationen diskriminierungsfrei und in ihrer Architektur und Governance so angelegt, dass keine Monopole entstehen. Diese Datenräume würden, im Gegensatz zu Open-Data-Portalen, auch die Nachnutzung sensibler Daten ermöglichen.

Weiterhin können Datentreuhänder*innen (vgl. Kapitel 4.4.3.2) als Teil der Datenräume den mitunter schwierigen Betrieb des Datenzugangs übernehmen und die Nutzbarkeit von Datenökosystemen für Forschende und andere Gruppen stark erhöhen.

7.3 Standardisierung als Basis für das Datenteilen

Grundsätzliche Standards, die in jedem Fall einzuhalten sind, um eine Nachnutzung zu ermöglichen, umfassen die Nutzung von offenen Formaten (wie GML, GeoJSON und GeoPackage) und offenen Schnittstellen („APIs“, z.B. OGC API - Features). Insbesondere bei annotierten Daten ist auch die Semantik der Labels sehr wichtig (vgl. Kapitel 4.3.5 und 4.4.4.2). Soweit möglich, sollte hier auf standardisierte Vokabulare, wie beispielsweise den OGC-Standard „TrainingDML-AI” (Yue & Shangguan, 2023), welcher Anforderungen an die Labels und den Labeling-Prozess definiert, zugegriffen werden.

Für die Erreichung von Interoperabilität wird empfohlen, ein Stufenmodell einzufordern und zu etablieren. Solche Stufenmodelle können mit einfach umzusetzenden Mindestanforderungen („ein Stern“) beginnen und mit vollständiger End-to-End-Interoperabilität („fünf Sterne“) enden und wurden beispielweise für Open Data definiert.

7.4 Etablierung eines rechtlichen Rahmens für die objektive Risikobewertung des Datenteilens

Zurzeit wird oft aufgrund diffuser möglicher Gefahren entschieden, Datensätze nicht für die Nachnutzung bereitzustellen (vgl. Kapitel 4.4.1.1). Selbst wenn eine Schutzbedarfsfeststellung durchgeführt wird, werden oft keine evidenzbasierten Gefahrenszenarien herangezogen, sondern es werden Entscheidungen individuell und aus voreingenommener Perspektive bewertet (siehe Diskussion zu Waldinventurdaten in Kapitel 4.2.2.2). Es ist daher dringend zu empfehlen, einen evidenzbasierten Rahmen für die objektive Risikobewertung des Teilens von Daten zu etablieren und dessen Umsetzung, ähnlich wie heute BSI Grundschutz/ISO 27001 und ähnliche, generell zu fordern.

7.5 Verpflichtendes Teilen der Daten von Förderprojekten nach klar definierten Standards

Forschungsprojekte, die mit öffentlichen Mitteln gefördert werden, sollten dazu verpflichtet werden, die im Rahmen des Forschungsvorhabens erhobenen Daten inklusive Metadaten und Annotationen bereitzustellen. Wichtig ist dabei die Aufarbeitung nach den FAIR-Prinzipien, um eine optimale Nachnutzung zu gewährleisten. Dabei müssen die Interessen der Forschenden berücksichtigt werden, welche beispielweise einen exklusiven, aber zeitlich begrenzten, Zugang zur Auswertung der Daten erhalten können. Es gilt hierbei zu überlegen, welche Frist für die Bereitstellung (z.B. 6 Monate nach Projektende) definiert werden sollte.

Den Forschenden bzw. Förderprojekten sollte weiterhin ein Budget für diese Aufgabe sowie Werkzeuge, die dies erleichtern, zur Verfügung gestellt werden (siehe auch Diskussion zu wirtschaftlichen Limitierungen und Hindernissen, Kapitel 4.4.2). Förderprojekte müssten dann direkt im Förderantrag detaillieren, welche Daten sie wie erheben, verarbeiten und managen wollen, und hätten die Auswahl aus einem klaren Satz an Standards, welcher fortwährend an die technische Entwicklung angepasst wird.

Zudem sollten Konsequenzen bei der Nichteinhaltung der Verpflichtung vorgesehen werden. Denkbar wäre beispielsweise das Zurückhalten von Projektmitteln, bis die Forschungsdaten bereitgestellt wurden.

7.6 Etablierung von Anreizstrukturen

Neben den Verpflichtungen sollten auch Anreizstrukturen geschaffen werden (vgl. Kapitel 4.4.2.2). So gibt es bereits den Begriff der Datenspende; würde diese als Spende im steuerlichen Sinne anerkannt werden, könnte dies vor allem für wirtschaftlich orientierte Akteure einen Anreiz darstellen. Ein weiterer denkbarer Ansatz wäre eine bilanzielle Würdigung von Daten; so könnten Daten, die „aktiviert“ werden, indem sie in Infrastrukturen zur Nachnutzung eingebracht werden, als immaterielles Wirtschaftsgut betrachtet werden.

In vielen Datenökosystemen gibt es weitere Anreize, Daten bereitzustellen. So kann z.B. der Zugang zu bestimmten Funktionen oder Datenpools auf die Akteure begrenzt werden, die selbst ihre eigenen Daten eingebracht haben.

Auch weichere Anreize können sich sehr positiv auswirken, beispielweise durch Sichtbarkeit der Organisationen, die Daten teilen, im Vergleich zu denen, die dies nicht tun. So könnten Datensätze, die höheren Anforderungen genügen oder unter permissiveren Lizenzen stehen, besonders hervorgehoben und beispielsweise in Suchen und Portalen zuerst genannt werden.

8 Vorgehensmodell zum Umgang mit wenigen annotierten Daten für die Entwicklung von ML-Anwendungen

Kapitel 8 konzentriert sich auf eine der größten aktuellen Herausforderungen für die Entwicklung von ML-Modellen im Umweltsektor: die Verfügbarkeit von annotierten Daten und welche Methoden Entwickler*innen einsetzen können, falls keine oder nur wenige Daten für ein Training vorhanden sind. Ziel ist die Ableitung von Handlungsempfehlungen für verschiedene Szenarien, die sich in Bezug auf die Verfügbarkeit der gelabelten Daten unterscheiden, die Entwicklung eines Entscheidungsbaums für relevante Annotationsansätze sowie die Validierung des erarbeiteten Vorgehensmodells anhand von zwei konkreten Anwendungsfällen: die Erkennung von Freiflächen-Photovoltaik (PV)-Anlagen und die Baumartenklassifikation in Wäldern. Das entwickelte Vorgehensmodell dient somit als eine praktische Orientierungshilfe für die komplexen Aufgaben guter Modellbildung.

Bereits zu Beginn des Projekts **LabelledGreenData4All** wurde der Schwerpunkt auf Fernerkundungsdaten (Luftbilder, Satellitendaten, Punktwolkendaten) festgelegt, da sie eine wichtige Rolle im Bereich Umwelt- und Nachhaltigkeitsanwendungen spielen. Texte und tabellarische Daten wurden aus Gründen einer notwendigen Fokussierung nicht betrachtet. Wirtschaftliche Aspekte und ökonomische Entscheidungsmodelle wurden in der Literaturrecherche (Kapitel 3) und auch bei der Potential- und Wirkungsanalyse (Kapitel 5 und 6) im Projekt behandelt, wurden aber aus Kapazitätsgründen bei der Entwicklung des Vorgehensmodells nicht berücksichtigt. Entscheidende Kriterien sind zudem Automatisierbarkeit und Skalierbarkeit der untersuchten Annotationsmethoden.

Die folgenden Aspekte wurden bei der Entwicklung und Evaluierung des Vorgehensmodells bearbeitet:

- Eine Plattform zur automatisierten Modellierung und Ausführung von Datenverarbeitungsprozessen für das effiziente Training von KI-Modellen wurde eingerichtet (Kapitel 8.1).
- Verschiedene Annotationsmethoden wurden evaluiert, darunter daten- und modellzentrierte Ansätze wie Pseudo-Labeling und Transfer Learning (Kapitel 8.2).
- Ein Entscheidungsbaum wurde entwickelt, um ML-Methoden basierend auf der Verfügbarkeit annotierter Daten auszuwählen (Kapitel 8.2).
- Eine prototypische Umsetzung und Evaluierung erfolgte durch die Anwendung des entwickelten Vorgehensmodells auf zwei ausgewählte Anwendungsfälle (Kapitel 8.3).

8.1 Setup der technischen Experimentierplattform

Für die Evaluierung der auf Grundlage der Rechercheergebnisse in Kapitel 3 ausgewählten Annotationsmethoden wurde zunächst eine technische Experimentierplattform aufgesetzt. Diese Plattform sollte es ermöglichen, Datenverarbeitungsprozesse (im Folgenden „Workflows“ genannt) zum Training von KI-Netzen und zur Vorverarbeitung von Daten zu modellieren und automatisiert auszuführen. Außerdem sollte sie vorhandene Rechenressourcen bestmöglich ausnutzen, um die Ausführungszeit der Workflows zu verringern.

Im Folgenden werden zunächst die Anforderungen an die Plattform und danach die verwendeten Technologien und die einzelnen Komponenten beschrieben.

8.1.1 Anforderungen

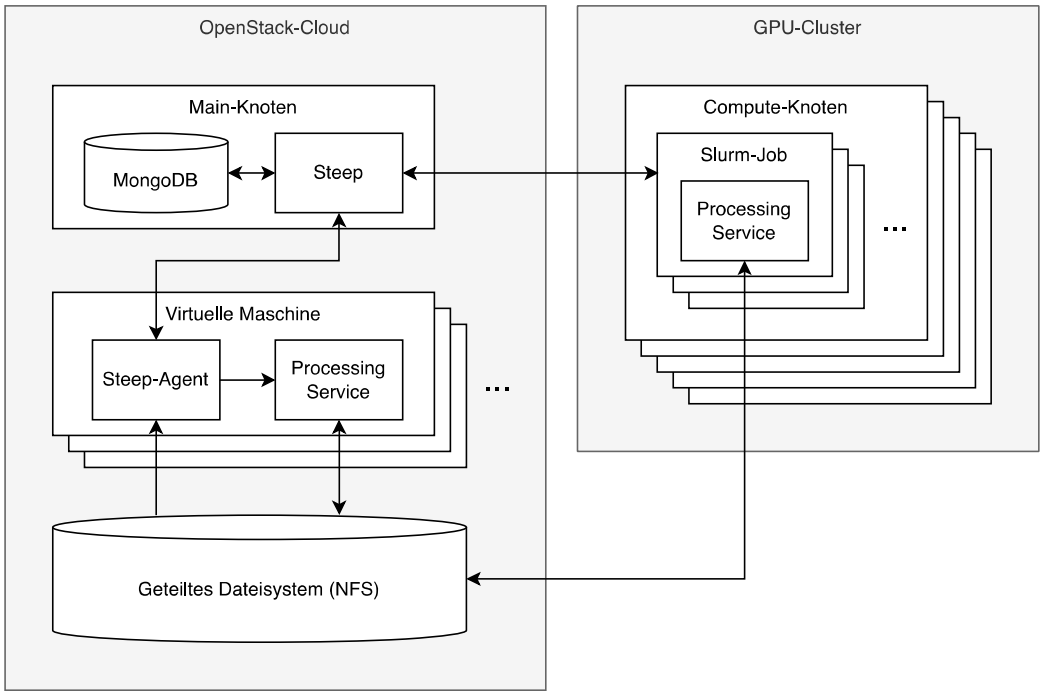
Die folgenden Anforderungen an die Experimentierplattform wurden zu Beginn der Studie identifiziert:

- KI-Trainings dauern oftmals mehrere Tage oder sogar Wochen. Die Plattform sollte eine Möglichkeit bieten, automatisierte Workflows auszuführen, die möglichst autonom laufen, sodass während der Verarbeitung keine menschliche Interaktion mit dem System stattfinden muss.
- Um sicherzustellen, dass langlaufende Workflows nicht durch kurzzeitige Ausfälle (z.B. Netzwerk ist kurz nicht verfügbar oder eine Softwarekomponente stürzt ab) unterbrochen werden und danach vollständig wiederholt werden müssen, sollte die Plattform Mechanismen zur Fehlertoleranz anbieten.
- Die Workflows sollten möglichst schnell abgearbeitet werden z.B. durch Parallelisierung und unter Verwendung von GPUs. Da dies eine begehrte und kostenintensive Ressource ist, sollten die GPUs nur bei Bedarf angefordert und nach Beendigung eines Workflows wieder freigegeben werden.
- Die Plattform sollte eine Möglichkeit bieten, Trainingsdaten hochzuladen und diese in den Workflows zu nutzen. Die Trainingsergebnisse sollten gespeichert und zum Download angeboten werden.

8.1.2 Technologien und Komponenten

Abbildung 22 gibt einen Überblick über die Softwarearchitektur der Experimentierplattform, den darin verwendeten Technologien und Komponenten und wie diese miteinander kommunizieren. Im Folgenden werden die Technologien und Komponenten im Detail beschrieben.

Abbildung 22: Softwarearchitektur der Experimentierplattform



Quelle: eigene Darstellung, Fraunhofer IGD

8.1.2.1 OpenStack-Cloud

Um die oben beschriebenen Anforderungen zu erfüllen, bot sich als Basis für die Plattform eine verteilte Infrastruktur in Form einer Cloud an. Diese bietet folgende Vorteile:

- Durch die Möglichkeit Berechnungen auf mehrere virtuelle Maschinen (VMs) zu verteilen, können Workflows parallelisiert und somit schneller abgearbeitet werden.
- In der Cloud können Rechenressourcen (wie z.B. GPUs) dynamisch und bei Bedarf angefordert und auch wieder freigegeben werden.
- Die Cloud bietet dynamisch reservierbaren Speicher, der für die Trainingsdaten und -ergebnisse genutzt und während des Trainings den verteilten Maschinen zur Verfügung gestellt werden kann.

Für die Umsetzung stand im Fraunhofer IGD bereits eine Cloud zur Verfügung, die auf dem Open-Source-Produkt OpenStack[30] basiert. Damit konnten dynamisch virtuelle Maschinen erzeugt werden, die für das KI-Training genutzt werden konnten. Außerdem bietet OpenStack unter anderem die Möglichkeit, Speicher zu reservieren und das Netzwerk zu konfigurieren, um die virtuellen Maschinen miteinander zu verbinden.

Die Cloud im Fraunhofer IGD nutzt folgende Hardware:

- Compute-/Controller-Knoten:
  - 2x CPU INTEL Xeon E5-2660v4 2,0GHz 14 Core LGA 2011
  - 8x DIMM DDR4-2400 32GB ECC reg
  - 2x Samsung SSD SM863 480GB
  - 4x 10GB Ethernet
- Storage-Knoten:
  - 2x CPU INTEL Xeon E5-2620v4 2, 1 GHz 8 Core LGA 2011-3
  - 8x DIMM DDR4-2400 32GB ECC eg DR
  - 4x Samsung SSD SM863 960GB
  - 14x HDD Hitachi 3TB SATA 64MB 7200 3,5"
  - 2x 1GB Ethernet

Die Compute-Knoten sind darüber hinaus mit einer begrenzten Anzahl an NVIDIA Tesla P40 GPUs mit je 24 GB RAM ausgestattet.

Für die Ausführung der Experimente in dieser Studie wurde die Auswahl an erzeugbaren virtuellen Maschinen auf zwei Typen (sogenannte „Flavors“) eingeschränkt:

- CPU-VMs: 4 virtuelle CPUs, 16 GB RAM
- GPU-VMs: 8 virtuelle CPUs, 64 GB RAM, 1 GPU

[30] https://www.openstack.org/

8.1.2.2 GPU-Cluster

Die OpenStack-Cloud im Fraunhofer IGD besitzt zwar eigene GPU-Ressourcen, dennoch sind diese in der Anzahl der zur Verfügung stehenden physischen Grafikkarten begrenzt. Aus diesem Grund wurde die Cloud um ein dezidiertes GPU-Cluster erweitert, das ebenfalls im Fraunhofer IGD zur Verfügung steht und dessen Fokus auf der Beschleunigung rechenintensiver Prozesse liegt. Das GPU-Cluster setzt sich aus fünf Rechenknoten zusammen, die folgendermaßen ausgestattet sind:

- HPE ProLiant XL675d Gen10 Plus
- 8x Nvidia A100 SXM4 40 GB
- 2x AMD EPYC 7742 64-Core Processor
- 1 TB RAM
- 32 TB SSD storage

Die Nutzung des Clusters geschieht jobbasiert. Einzelne in sich geschlossene Arbeitsschritte können über den Open-Source-Workload-Manager Slurm[31] ausgeführt werden. Komplexe Workflows werden nicht unterstützt.

8.1.2.3 Geteilter Speicher

Zur Speicherung der Trainingsdaten boten sich unterschiedliche Varianten an. In einer Cloud werden große und langlebige Daten häufig in einem so genannten Object Store gespeichert (z.B. basierend auf der S3-Schnittstelle). In diesem Fall liegen die Daten zwar an einer zentralen Stelle, auf die von allen virtuellen Maschinen zugegriffen werden kann, für das KI-Training müssen die einzelnen Datenpakete aber dennoch zunächst auf die lokalen Festplatten der VMs heruntergeladen werden. Erfahrungen aus vergangenen Projekten haben gezeigt, dass dies ein durchaus zeitaufwändiger Prozess sein kann, der die Ausführung des Trainings unnötig verlangsamt.

Als Alternative wurde deshalb ein geteiltes Dateisystem in Form eines NFS-Shares eingerichtet. Dieses bietet eine höhere Performance als der Object Store und kann über übliche Betriebssystemmechanismen in die virtuellen Maschinen sowie in die Knoten des GPU-Clusters eingehängt werden, sodass alle beteiligten Komponenten auf den gleichen Datenbestand zugreifen können.

8.1.2.4 Workflowmanagement: Steep und Processing Services

Für die Automatisierung des Trainings wurden konzeptionell so genannte „Scientific Workflows“ genutzt. Im Gegensatz zu ereignisbasierten Workflows, die häufig für die Automatisierung von Geschäftslogik eingesetzt werden, sind Scientific Workflows datengetrieben und eignen sich dazu, eine vorgegebene Menge an Daten batchbasiert in mehreren Arbeitsschritten (auch Tasks oder Aktionen/Actions genannt), die in einer vordefinierten Abfolge ausgeführt werden, zu verarbeiten, um ein gewünschtes Berechnungsergebnis zu erhalten. Scientific Workflows werden in der Regel deklarativ modelliert und beschreiben die Abhängigkeiten zwischen den Aktionen, wodurch sich der Datenfluss implizit ergibt. Ein Workflowmanagementsystem versucht auf Basis dieser Beschreibung eine sinnvolle Ausführungsstrategie zu finden und nutzt dabei ggf. Parallelisierung, wenn dies möglich ist. So können z.B. Aktionen gleichzeitig zueinander auf

[31] https://slurm.schedmd.com/

unterschiedlichen Prozessoren ausgeführt werden, wenn keine logischen Abhängigkeiten zwischen ihnen bestehen.

Ein weiterer Vorteil von Scientific Workflows besteht darin, dass die einzelnen Aktionen in der Regel durch isolierte, in sich gekapselte Softwarekomponenten implementiert werden, die Daten austauschen, indem sie Dateien lesen und schreiben. Dies entspricht der Microservice-Architektur in der verteilten Softwareentwicklung, bei der lose gekoppelte Komponenten durch Nachrichten miteinander kommunizieren. Sollte eine der Komponente ausfallen, stürzt im Gegensatz zu einem monolithischen System nicht gleich die ganze Anwendung ab. Übertragen auf Workflows bedeutet dies, dass einzelne Arbeitsschritte fehlschlagen können, ohne dass der gesamte Workflow abgebrochen werden muss. Die fehlgeschlagenen Aktionen können zu einem späteren Zeitpunkt noch einmal wiederholt und der Workflow fortgesetzt werden.

Darüber hinaus bietet der serviceorientierte Ansatz die Möglichkeit, die einzelnen Aktionen in unterschiedlichen Technologien (mit verschiedenen Programmiersprachen oder Frameworks) zu implementieren, was nicht möglich wäre, wenn der gesamte Workflow in einer einzelnen monolithischen Anwendung modelliert werden müsste. Dies ermöglicht es z.B. das Training mit KI-Technologien wie Tensorflow[32] und Pytorch[33] über die Programmiersprache Python umzusetzen, während CPU-lastigere Aktionen, wie das Vorverarbeiten der Trainingsdaten, in einer performanteren nativen Programmiersprache (z.B. Rust oder C++) geschrieben werden können. Darüber hinaus können die einzelnen Services von unterschiedlichen Personen unabhängig voneinander implementiert werden, was die Entwicklungszeit des Workflows verkürzt. Einmal entwickelte Services können auch in mehreren Workflows wiederverwendet werden.

Für die Experimentierplattform kam das vom Fraunhofer IGD entwickelte Open-Source-Workflowmanagementsystem Steep[34] zum Einsatz. Steep wurde für den Einsatz in der Cloud entwickelt und bietet neben der reinen Workflowausführung auch die Möglichkeit dynamisch und bei Bedarf virtuelle Maschinen zu erstellen, dort einzelne Workflowteile auszuführen und die Maschinen anschließend wieder freizugeben.

Darüber hinaus ist Steep in der Lage, automatisiert fehlgeschlagene Services zu wiederholen, um so eine hohe Fehlertoleranz zu ermöglichen.

Für die Ausführung interpretiert Steep die deklarative Workflowbeschreibung und erstellt dynamisch zur Laufzeit einen optimierten Ausführungsplan, der sich aus so genannten Prozessketten zusammensetzt. Eine Prozesskette ist eine Abfolge von Aktionen, die sequenziell auf einer virtuellen Maschine ausgeführt werden. Mehrere Prozessketten können dabei parallel von unterschiedlichen Maschinen verarbeitet werden.

Zu diesem Zweck kann Steep dynamisch virtuelle Maschinen in der Cloud erzeugen, die den Anforderungen der Prozessketten entsprechen: benötigt eine Prozesskette z.B. GPU-Ressourcen, wird eine VM mit einer GPU erzeugt; benötigt sie dies nicht, kann eine kostengünstigere VM ohne GPU akquiriert werden. Sollte die VM nach Beendigung einer Prozesskettenausführung nicht innerhalb einer konfigurierbaren Zeit (standardmäßig fünf Minuten) erneut benötigt werden, wird sie automatisch von Steep wieder freigegeben.

Im Gegensatz zu einigen anderen Workflowmanagementsystemen, die auf gerichteten Graphen basieren und die einen statischen Ausführungsplan zu Beginn der Workflowausführung erstellen, bietet Steep durch seinen dynamischen Scheduler letztlich auch die Möglichkeit

[32] https://tensorflow.org/

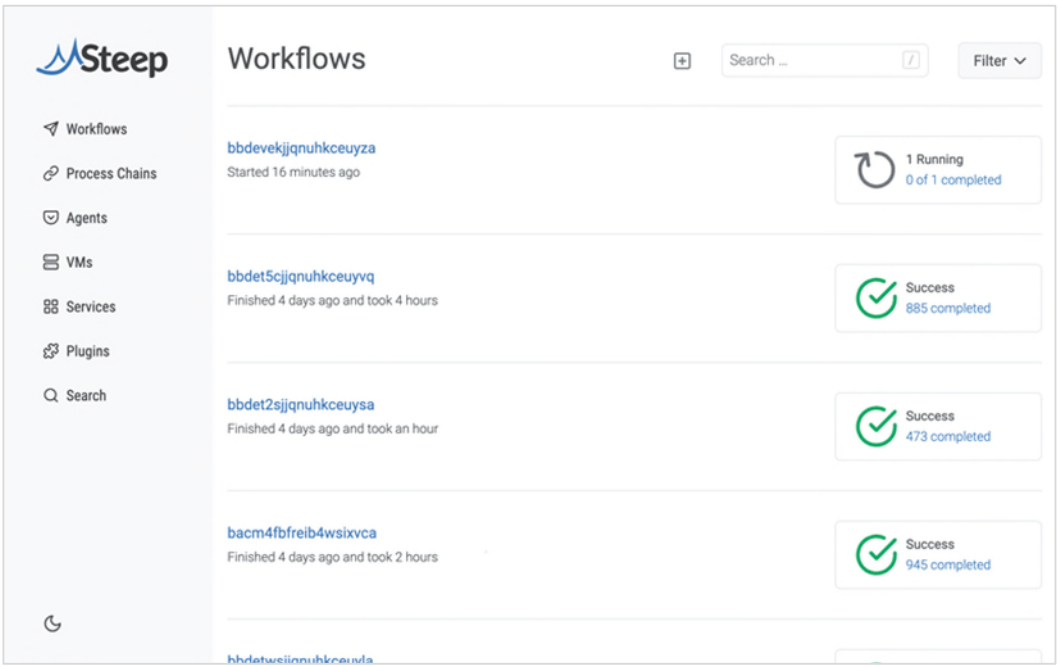
[33] https://pytorch.org/

[34] https://steep-wms.github.io/

ungerichtete Workflowgraphen mit Zyklen zu modellieren. Dies ermöglicht die Ausführung von Schleifen, die eine bestimmte Abfolge von Aktionen wiederholen bis eine definierte Abbruchbedingung eingetreten ist. Für den Fall von KI-Trainings lässt sich dieser Ansatz sehr gut nutzen, um mehrere Epochen durchlaufen zu lassen.

Steep wird bereits seit vielen Jahren produktiv in Forschungsprojekten und in der Industrie eingesetzt, z.B. für die Unterstützung des Glasfaserausbaus in Deutschland[35]. Ein Screenshot der graphischen Benutzeroberfläche von Steep ist in Abbildung 23 zu sehen.

Abbildung 23: Webbasierte graphische Benutzeroberfläche von Steep



Quelle: Fraunhofer IGD

8.1.2.5 Deployment

Für das Deployment des Steep-Main-Knotens und die Konfiguration des Netzwerks in OpenStack wurden die dafür in der Industrie üblichen Open-Source-Tools Terraform[36] und Ansible[37] verwendet. Mit Terraform lässt sich eine Cloud-Konfiguration deklarativ beschreiben und automatisiert anwenden. Ansible ermöglicht es, virtuelle Maschinen zu provisionieren, d.h. zu konfigurieren und Software darauf zu installieren.

Neben Steep wurde auf dem Main-Knoten eine PostgreSQL-Datenbank mit der PostGIS-Erweiterung installiert. Die Konfiguration von Steep[38] wurde entsprechend der Ausführungsumgebung und der Service-Metadaten[39] entsprechend der in Abschnitt 8.3.1 beschriebenen Services angepasst. Für die Kommunikation mit dem GPU-Cluster wurde ein

[35] siehe auch https://steep-wms.github.io/showcase/telekom/

[36] https://www.terraform.io/

[37] https://www.ansible.com/

[38] https://steep-wms.github.io/docs/steepyaml/

[39] https://steep-wms.github.io/docs/servicesservicesyaml/

Steep-Plugin vom Typ „Runtime“ entwickelt[40], das Jobs mittels der Slurm-CLI „sbatch“ startet und die Ausführung überwacht.

Weitere Informationen zur allgemeinen Konfiguration von Steep finden sich in der Dokumentation auf der Website unter https://steep-wms.github.io/docs/.

8.2 Entwicklung eines Vorgehensmodells für Datenannotationen

In Kapitel 3.1 wurden die wichtigsten Methoden zum Umgang mit wenigen annotierten Daten zusammengefasst, wobei sowohl die Stärken als auch die Schwächen der betrachteten Ansätze beschrieben wurden. Diese Bewertung bildet die Grundlage für die Entwicklung eines Entscheidungsbaums, der als zentraler Bestandteil der Studie dient. Der Entscheidungsbaum soll als praktische Orientierungshilfe fungieren und Handlungsempfehlungen für verschiedene Szenarien geben, die sich in Bezug auf die Verfügbarkeit der Daten und die spezifischen Anforderungen für maschinelles Lernen im Bereich Umwelt- und Nachhaltigkeitsanwendungen unterscheiden.

Neben der Verfügbarkeit von gelabelten Daten spielt auch die Qualität der Label eine zentrale Rolle für die Entwicklung robuster ML-Anwendungen. Bisher gibt es jedoch kein standardisiertes Verfahren zur Beschreibung und Beurteilung der Qualität von Datenlabels, das sich explizit auf Annotationen für Umweltdaten und umweltrelevante Daten bezieht (siehe Kapitel 4.3). In die Entwicklung des Vorgehensmodells wurde daher nur die Menge an verfügbaren Labels, nicht jedoch die Qualität als Entscheidungskriterium aufgenommen. Für das hier vorgestellte initiale Vorgehensmodell wird davon ausgegangen, dass die Qualität der verfügbaren Labels gut ist. Sollte sich im Laufe des Trainingsprozesses herausstellen, dass die Ergebnisse aufgrund qualitativ schlechter Labels nicht befriedigend sind, kann im Entscheidungsbaum auf die Methoden für wenige oder nicht vorhandene Label zurückgegriffen werden.

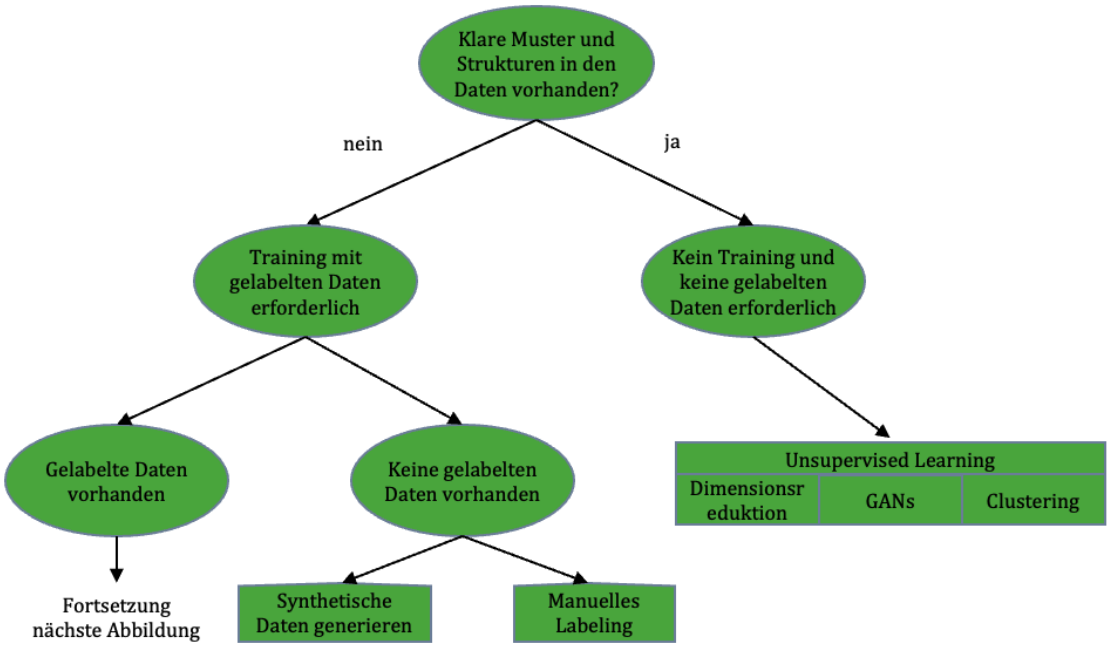
Um einen klaren Überblick zu gewährleisten und das Verständnis für die Entscheidungsfindung zu erleichtern, wurde bewusst darauf verzichtet, den Datentypen explizit zu berücksichtigen. Diese Vorgehensweise wurde gewählt, da die Art der Daten in diesem Zusammenhang als zweitrangig angesehen wird, während der Schwerpunkt auf den zugrundeliegenden Methoden und deren Anwendungsmöglichkeiten liegt. Die Fokussierung auf die Methoden hat den Vorteil, dass die erarbeiteten Empfehlungen breiter anwendbar und auf unterschiedliche Datentypen übertragbar sind.

8.2.1 Keine annotierten Daten vorhanden

In diesem Abschnitt wird das Szenario betrachtet, in dem keine gelabelten Daten für das Training zur Verfügung stehen. Um Missverständnissen vorzubeugen, wird an dieser Stelle darauf hingewiesen, dass es sich hierbei nicht um das generelle Fehlen von Daten handelt, sondern um die für das Training benötigten gelabelten Daten. Die folgende Abbildung 24 zeigt den Teilbaum dieses Szenarios.

[40] Siehe https://steep-wms.github.io/docs/custom-runtimes/

Abbildung 24: Entscheidungsbaum für die Ausgangssituation, dass keine gelabelten Daten erforderlich sind, bzw. erforderlich, aber nicht vorhanden sind



Quelle: eigene Darstellung, Fraunhofer IGD

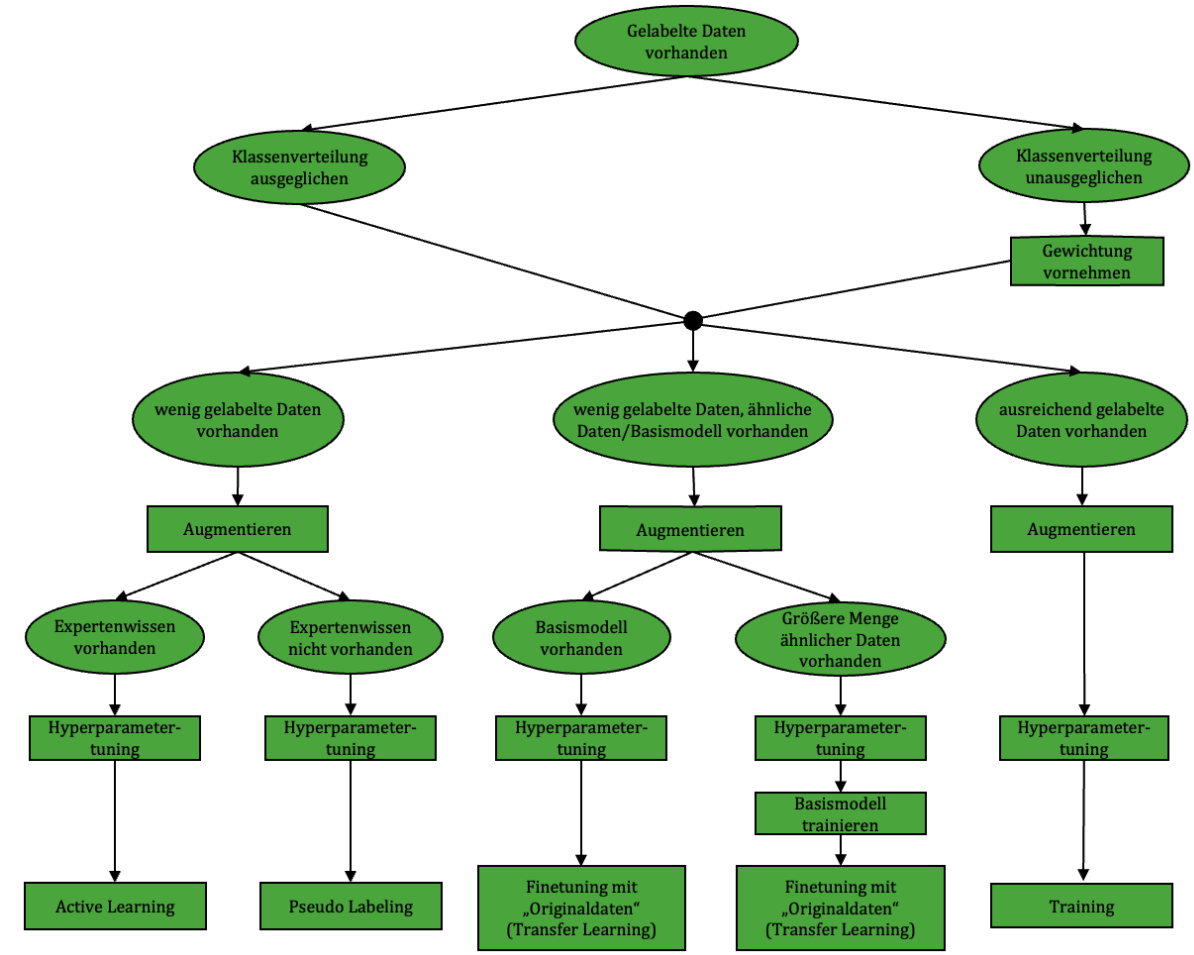
Für jeden Anwendungsfall muss immer zuerst geprüft werden, ob das Trainieren eines ML-Modells notwendig ist oder nicht. Wenn z.B. ausreichend offensichtliche Muster und Strukturen in den Daten vorhanden sind, sind trainierbare Ansätze in der Regel nicht erforderlich. In diesem Fall können verschiedene „unsupervised learning" Verfahren wie Clustering (k-means), Dimensionsreduktion (Principal Component Analysis) oder GANs (DCGAN) eingesetzt werden.

Sollen jedoch hohe Genauigkeiten erreicht werden oder es liegen komplexe Sachverhalte vor, die sich nicht aus den vorhandenen Mustern und Strukturen automatisch ableiten lassen, wird ein Training von ML-Modellen empfohlen. Stehen in diesem Fall keine gelabelten Daten für das Training zur Verfügung, müssen sie zunächst generiert werden. Dies ist entweder durch nachträgliches manuelles Labeling oder durch die Generierung synthetischer Daten möglich.

8.2.2 Annotierte Daten in verschiedenem Umfang sind vorhanden

Dieser Abschnitt behandelt den Fall, dass gelabelte Daten zur Verfügung stehen. Das übergeordnete Entscheidungskriterium für die Auswahl geeigneter ML-Methoden ist die Menge der verfügbaren gelabelten Daten. Dies ist im Teilbaum in Abbildung 25 dargestellt.

Abbildung 25: Entscheidungsbaum für die Ausgangssituation, dass gelabelte Daten in unterschiedlichem Umfang vorhanden sind



Quelle: eigene Darstellung, Fraunhofer IGD

Wenn gelabelte Daten zur Verfügung stehen, müssen diese zunächst analysiert werden. Ein grundlegender Aspekt ist dabei die Klassenverteilung der Labels. Nur eine möglichst homogene Verteilung der Klassen garantiert gute Trainingsergebnisse. Sind die Klassen sehr ungleichmäßig verteilt, müssen die Daten entsprechend gewichtet werden.

Anschließend steht für die Auswahl von geeigneten Methoden die Menge der gelabelten Daten im Vordergrund. Hierbei werden die nachfolgenden drei Fälle unterschieden:

- Wenig annotierte Daten für den gegebenen Anwendungsfall sind vorhanden.
- Wenig annotierte Daten, aber ein Basismodell oder mehr Daten eines ähnlichen Anwendungsfalls sind vorhanden.
- Ausreichend annotierte Daten sind vorhanden.

Da es eine unzureichende Transparenz und Nachvollziehbarkeit des Labeling-Prozesses gibt und die Qualität von Labels nicht automatisch evaluiert werden kann, wurde der Aspekt „Qualität der Labels“ nicht explizit als eigenes Entscheidungskriterium hinzugezogen. Wie bereits erwähnt, besteht die Annahme im Entscheidungsbaum, dass die Qualität der zur Verfügung stehenden gelabelten Daten gut ist. Sollte sich im Laufe des Trainingsprozesses herausstellen,

dass die Labels von minderwertiger Qualität sind, kann im Entscheidungsbaum auf die Methoden für wenige oder nicht vorhandene Labels zurückgegriffen werden.

Unabhängig vom jeweiligen Szenario müssen die verfügbaren Daten für das Training von KI-Anwendungen augmentiert werden. Durch das Augmentieren wird die Menge und Vielfalt der Daten erhöht. Dies ist insbesondere bei einer geringen Datenmenge relevant. Mit Hilfe der augmentierten Daten werden im anschließenden Training robustere und allgemeinere Muster gelernt, was die Genauigkeit erhöht. Ein weiterer Aspekt ist das Hyperparametertuning. Es wird empfohlen, für das Training jeder ML-Anwendung ein Hyperparametertuning durchzuführen, um optimale Trainingsergebnisse zu erzielen. Aus diesem Grund ist das Hyperparametertuning im Entscheidungsbaum ebenfalls enthalten.

8.2.2.1 Wenig annotierte Daten vorhanden

Stehen nur wenige gelabelte Daten zur Verfügung, ist für das weitere Vorgehen entscheidend, ob zusätzliches Fachwissen zur Verfügung steht. Ist dies der Fall, bieten sich Active Learning-Algorithmen an, bei denen der Mensch in das Training eingebunden wird. Dabei wird die KI zunächst mit den zur Verfügung stehenden Daten trainiert. Predicts, bei denen danach hohe Unsicherheiten bestehen, werden vom Menschen „bewertet“ und gegebenenfalls korrigiert. Diese Rückmeldungen werden dann als Labels für die nächsten Trainingsiterationen verwendet. Dieser Ansatz eignet sich auch bei stark inhomogenen Klassenverteilungen der annotierten Daten.

Wenn kein Fachwissen vorhanden ist, wird das Pseudo-Labeling empfohlen. Dieser Ansatz kommt dem Active Learning sehr nahe. Da jedoch kein Fachwissen vorhanden ist, werden die Predicts der KI direkt als Labels in das Training der folgenden Iterationen eingespeist. Beachtet werden sollte allerdings an dieser Stelle die Gefahr von systematischen Fehlinterpretationen. Auf diesen Aspekt wird in Kapitel 8.3.3.1.1 nochmals eingegangen.

8.2.2.2 Wenig annotierte Daten, aber Basismodell oder mehr ähnliche Daten vorhanden

Ist neben der geringen Anzahl an annotierten Daten noch ein Basismodell gegeben oder stehen Trainingsdaten eines ähnlichen Anwendungsfalls zur Verfügung, so bietet sich allgemein das Transfer Learning an. Ist dabei ein trainiertes Basismodell vorhanden, so kann dieses direkt mit den wenigen „Originaldaten“ nachtrainiert werden. Ist kein Basismodell vorhanden, so muss dieses zunächst mit den ähnlichen Daten trainiert werden und kann im nachfolgenden Schritt mit den wenigen Originaldaten nachtrainiert werden.

8.2.2.3 Ausreichend annotierte Daten vorhanden

Steht eine ausreichende Menge an annotierten Daten bereit, so ist das methodische Vorgehen trivial. Nach dem Augmentieren der Daten und dem eingangs erwähnten Hyperparametertuning wird die ML-Anwendung trainiert. Für eine repräsentative Auswertung ist es dabei wichtig, den Datensatz in einen Trainings-, Validierungs-, und Testteil aufzuteilen. Ein mögliches Verhältnis ist dabei 70%-20%-10% (Solawetz, 2020).

8.3 Prototypische Umsetzung und Evaluierung

In der letzten Phase der Studie wurde das in Kapitel 8.2 erarbeitete Vorgehensmodell auf zwei Anwendungsfälle angewendet, die notwendigen Datenverarbeitungsschritte entlang der ausgewählten Pfade auf der bereitgestellten Experimentierplattform durchgeführt und die Ansätze hinsichtlich ihrer Effizienz, Genauigkeit und Skalierbarkeit bewertet. Gemeinsam mit

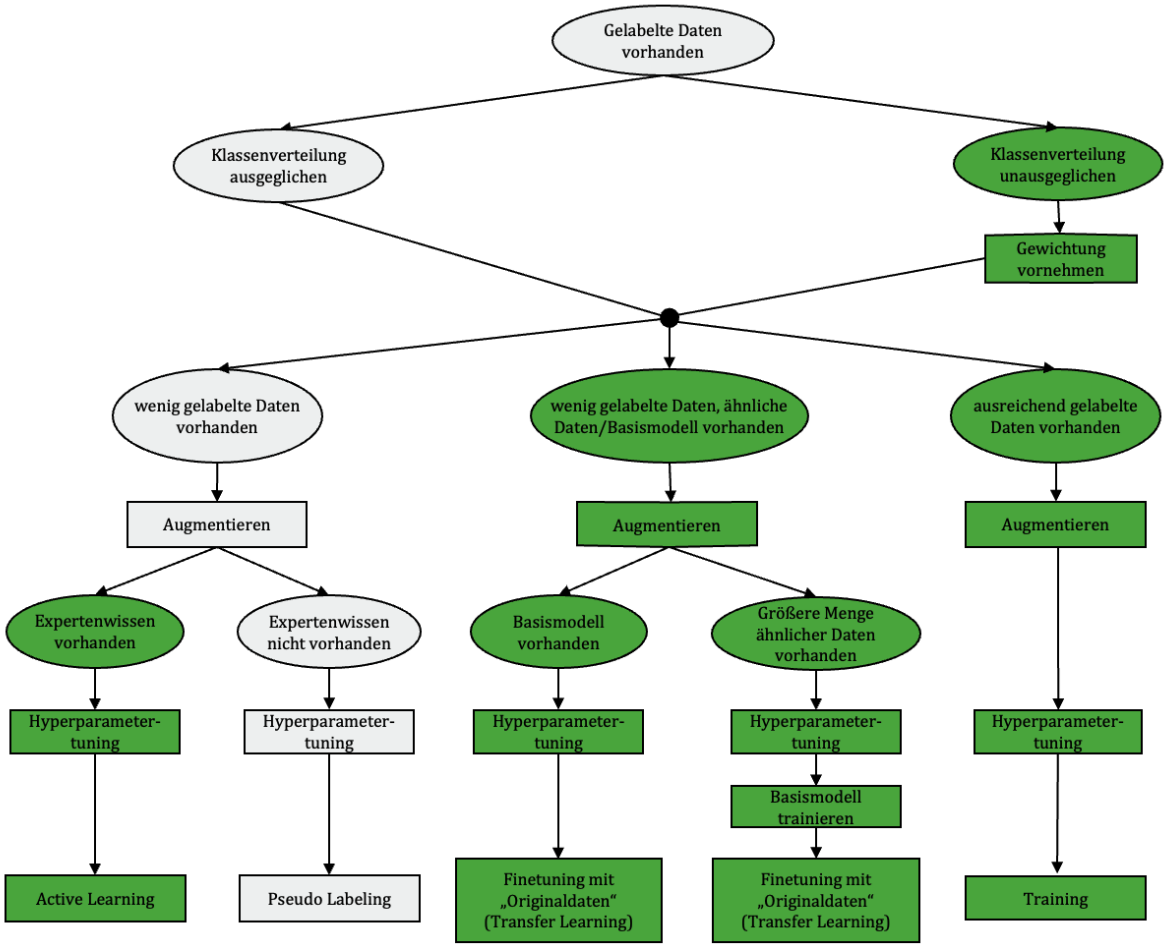
dem UBA wurde entschieden, dass hierfür die Anwendungsfälle „Erkennung von Freiflächen-PV Anlagen“ (***FF-PV***) und „Erkennung von Baumarten in Wäldern“ (***Wald***) dienen sollen.

Die Entwicklung eines Algorithmus zur automatisierten Erkennung und Kartierung bereits vorhandener Freiflächen-Photovoltaikanlagen in Deutschland ist sinnvoll, da sie eine umfassende Datenbasis für die Planung und den Ausbau erneuerbarer Energien schafft. Eine solche Kartierung ermöglicht ein effektives Monitoring des Bestands an Photovoltaikanlagen, was dabei hilft, den Fortschritt der Energiewende zu bewerten, erforderliche Anpassungen vorzunehmen und gezielte Fördermaßnahmen zu entwickeln. Die Automatisierung der Erkennung spart Zeit und Kosten im Vergleich zu manuellen Kartierungen und fördert zudem die Transparenz über den Stand der Photovoltaiknutzung, was politische Entscheidungen unterstützt. Für die Umsetzung des Anwendungsfalls **FF-PV** wurden frei verfügbare gelabelte Daten, die auf der Plattform Zenodo veröffentlicht sind, verwendet (siehe auch Kapitel 8.3.2.1).

Die Entwicklung eines Algorithmus zur deutschlandweiten automatisierten Baumartenklassifizierung ist sinnvoll, weil sie eine präzise und effiziente Erfassung der Biodiversität ermöglicht. Eine solche Klassifizierung unterstützt die Überwachung von Wäldern, und hilft bei der Planung nachhaltiger Forstwirtschaft. Durch die automatisierte Analyse großer Datenmengen können Veränderungen in der Baumartenverteilung schnell erkannt werden, was für die Anpassung an den Klimawandel und die Erhaltung von Ökosystemen entscheidend ist. Ebenso wie bei den Photovoltaikanlagen spart die Automatisierung der Kartierung Zeit und Kosten im Vergleich zu manuellen Kartierungen. Für die Umsetzung des Anwendungsfalls **Wald** wurden die auf dem Portal OpenGeodata.NRW frei verfügbaren Baumartenklassifikationen für NRW als Trainingsdaten verwendet (siehe auch Kapitel 8.3.2.1).

Wirft man einen Blick auf die beiden Anwendungsfälle sowie die dazu zur Verfügung stehenden gelabelten Daten und bezieht den vorgestellten Entscheidungsbaum ein, so liegt mit der Annahme, dass für den jeweiligen Anwendungsfall kein Fachwissen verfügbar ist, für beide Anwendungsfälle eine klare Empfehlung für die Verwendung des *Pseudo-Labelings* vor (siehe Abbildung 26). Aus diesem Grund bildet nachfolgend das Pseudo-Labeling den Hauptteil der prototypischen Umsetzung und wird für beide Anwendungsfälle durchgeführt.

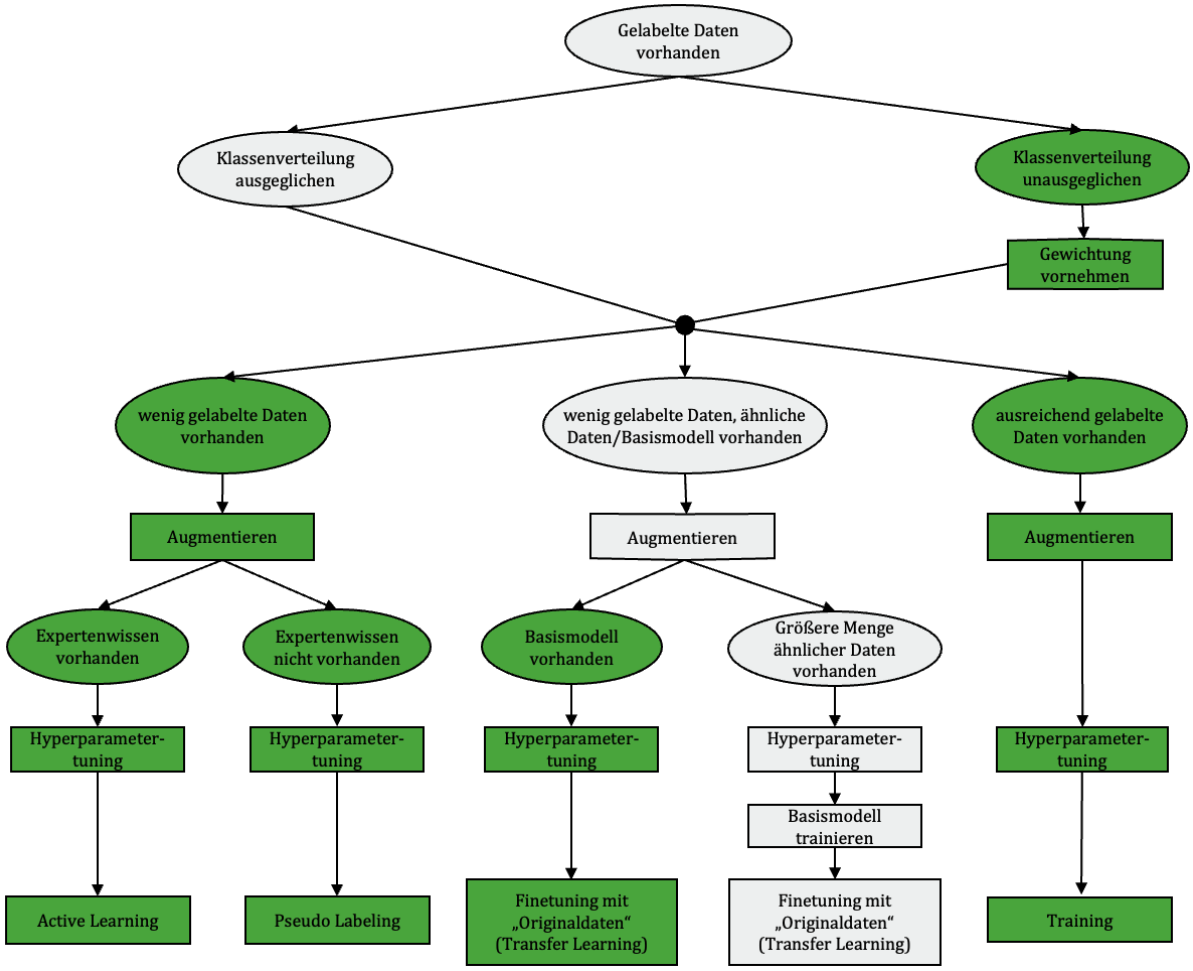
Abbildung 26: Vorgehen (Pseudo-Labeling), das für die Anwendungsfälle Wald und FF-PV umgesetzt wird



Quelle: eigene Darstellung, Fraunhofer IGD

Zum Vergleich wurde im Rahmen der Studie mit dem Transfer Learning aber zusätzlich auch ein modellzentrierter Ansatz durchgeführt und evaluiert. Für die Umsetzung wurden die zur Verfügung stehenden FF-PV-Trainingsdaten angepasst, indem zunächst mit einer größeren Menge ähnlicher Bilddaten (Satellitenbilder anstatt Orthophotos) ein Basismodell trainiert wurde. Anschließend wurde dieses Basismodell mit den „Originaldaten“ nachtrainiert. Dieses Vorgehen entspricht der *Transfer-Learning*-Methodik und ist in Abbildung 27 blau markiert. Genaue Informationen zu den verwendeten Daten sind in Kapitel 8.3.2.1 nachzulesen.

Abbildung 27: Vorgehen (Transfer Learning), das für den FF-PV Anwendungsfall umgesetzt wird



Quelle: eigene Darstellung, Fraunhofer IGD

Somit werden aus methodischer Sicht ein datenzentrierter und ein modellzentrierter Ansatz erprobt und evaluiert. Pseudo-Labeling und Transfer Learning sind von allen vorgestellten Methoden zudem die Ansätze, die das höchste Automatisierungs- und Skalierungspotential aufweisen und für die Umsetzung kein zusätzliches Fachwissen erfordern.

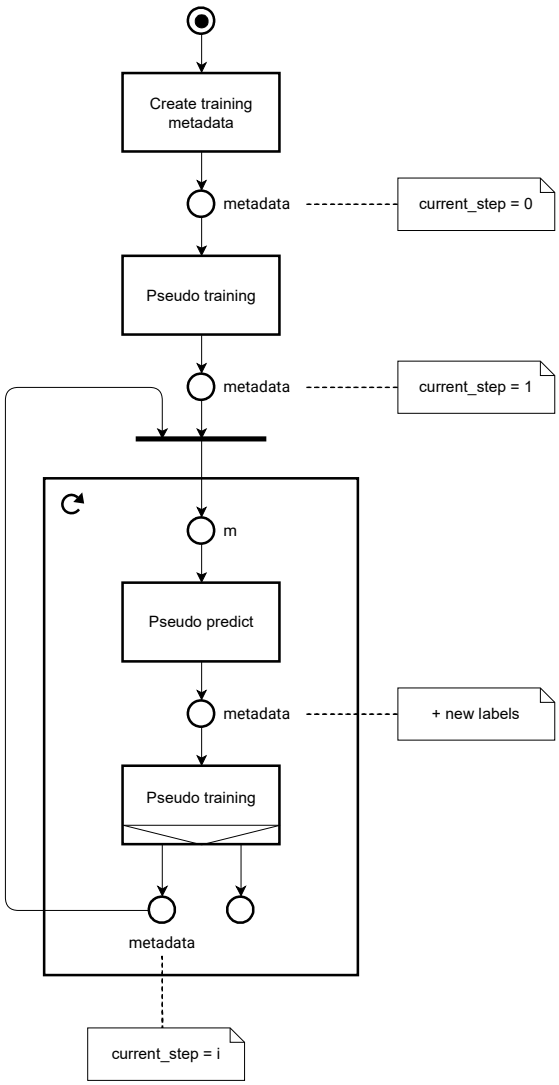
8.3.1 Umsetzung der Scientific Workflows auf der Experimentierplattform

Wie im vorherigen Abschnitt beschrieben wurden das Pseudo-Labeling und Transfer Learning als Methoden für die prototypische Umsetzung ausgewählt. Die technische Umsetzung dieser beiden Methoden auf der Experimentierplattform wird in den nächsten beiden Abschnitten 8.3.1.1 und 8.3.1.2 beschrieben. Sowohl die Erläuterungen in diesem Kapitel als auch die dahinterliegenden Implementierungen wurden dabei absichtlich generisch gehalten, um eine Anwendbarkeit auf andere Anwendungsfälle zu gewährleisten.

8.3.1.1 Pseudo Labeling

Der Pseudo-Labeling-Ansatz wurde, wie in Abschnitt 8.1.2.4 beschrieben, als datengetriebener Scientific Workflow mit dem Workflowmanagement-Tool Steep umgesetzt. Abbildung 28 zeigt schematisch den Ablauf des Workflows mit seinen einzelnen Schritten.

Abbildung 28: Pseudo-Labeling-Workflow in der erweiterten Petri-Net-Notation



Quelle: eigene Darstellung nach Aalst & Hee, 2004 und Krämer et al., 2021, Fraunhofer IGD

Der Workflow besteht aus drei Services: „Create training data“, „Pseudo training” und “Pseudo predict”. Als Parameter können dem Workflow Prozentsätze sowie eine Strategie übergeben werden:

- „evaluate-percent“: Dieser Parameter gibt an, wie viel Prozent des Datensatzes, auf dem das Pseudo-Labeling angewandt wird, für eine spätere Evaluierung des Ansatzes reserviert werden sollen.
- „first-percent“: Mit diesem Parameter wird bestimmt, wie viel Prozent des Datensatzes für das erste Training verwendet werden sollen. Es werden nur Bildpaare aus dem nicht reservierten Teil ausgewählt.
- „increment-percent“: Gibt an, wie viel Prozent des Datensatzes bei jedem Schritt des Pseudo-Labelings zusätzlich verwendet werden sollen. Auch hier werden nur Bildpaare aus dem nicht reservierten Teil ausgewählt.

- „strategy“: Dieser Parameter kann die Werte „sequential“ oder „random“ annehmen und gibt an, ob die Bilder aus dem Datensatz entsprechend den Prozentsätzen in sequenzieller Reihenfolge oder zufällig ausgewählt werden sollen.

Der Workflow startet mit dem Service „Create training data“, der den Datensatz einliest und entsprechend den oben beschriebenen Parametern aufteilt. Die Aufteilung wird in eine JSON-Datei geschrieben, die im Nachfolgenden *Metadaten-Datei* genannt wird. Der Service generiert dabei nicht nur die Aufteilung für den ersten Schritt des Pseudo-Labelings, sondern eine Liste mit allen Schritten. Bei jedem Schritt ergänzt der Service dabei die Liste der Bilder entsprechend dem Parameter „increment-percent“ bis keine Bilder mehr zur Verfügung stehen. Außerdem speichert der Service eine Variable „current_step“ in der Metadaten-Datei und setzt ihren initialen Wert auf 0. Diese Variable wird im weiteren Verlauf des Workflows von den anderen Services genutzt, um auf den aktuellen Eintrag in der Liste der Schritte zuzugreifen.

Als nächstes folgt ein initialer Aufruf des Services „Pseudo-Training“, der mit den im ersten Schritt der Metadaten-Datei referenzierten Bildpaaren ein Training durchführt. Diese Bildpaare entsprechen dem durch den Parameter „first-percent“ vorgegebenen Prozentsatz. Der Service speichert die erzeugten Trainingsgewichte in einer separaten Datei. Abschließend kopiert er die Metadaten-Datei, fügt dabei den Pfad zu den Trainingsgewichten hinzu und erhöht die Variable „current_step“ auf 1. Es ist zu beachten, dass die Trainingsgewichte nicht in Abbildung 28 gezeigt werden, da diese laut Workflow-Definition kein Ausgabe-Parameter des Services sind.

Daraufhin folgt eine Schleife, die die in ihr enthaltenen Services ausführt, bis das Pseudo-Labeling beendet ist. Innerhalb der Schleife wird zunächst der Service „Pseudo Predict“ ausgeführt, der anhand der Gewichte aus dem vorigen Trainingsschritt neue Pseudo-Labels generiert, und zwar genauso viele wie durch den Parameter „increment-percent“ vorgegeben. Der Service schreibt die Pfade zu den neuen Labels für den aktuellen Schritt in eine weitere Kopie der Metadaten-Datei.

Abschließend wird erneut der Service „Pseudo-Training“ ausgeführt, der nun die neuen Labels verwenden kann und ein neues Netz trainiert. Wieder speichert der Service die Gewichte und erstellt eine Kopie der Metadaten-Datei. Insbesondere die Variable „current_step“ wird wieder um 1 erhöht.

Dieser Vorgang wiederholt sich, bis der Service „Pseudo-Training“ erkennt, dass kein weiterer Schritt mehr ausgeführt werden kann. In diesem Fall schreibt er keine neue Metadaten-Datei. Steep erkennt dies und beendet die For-Schleife und somit auch den Pseudo-Labeling-Workflow.

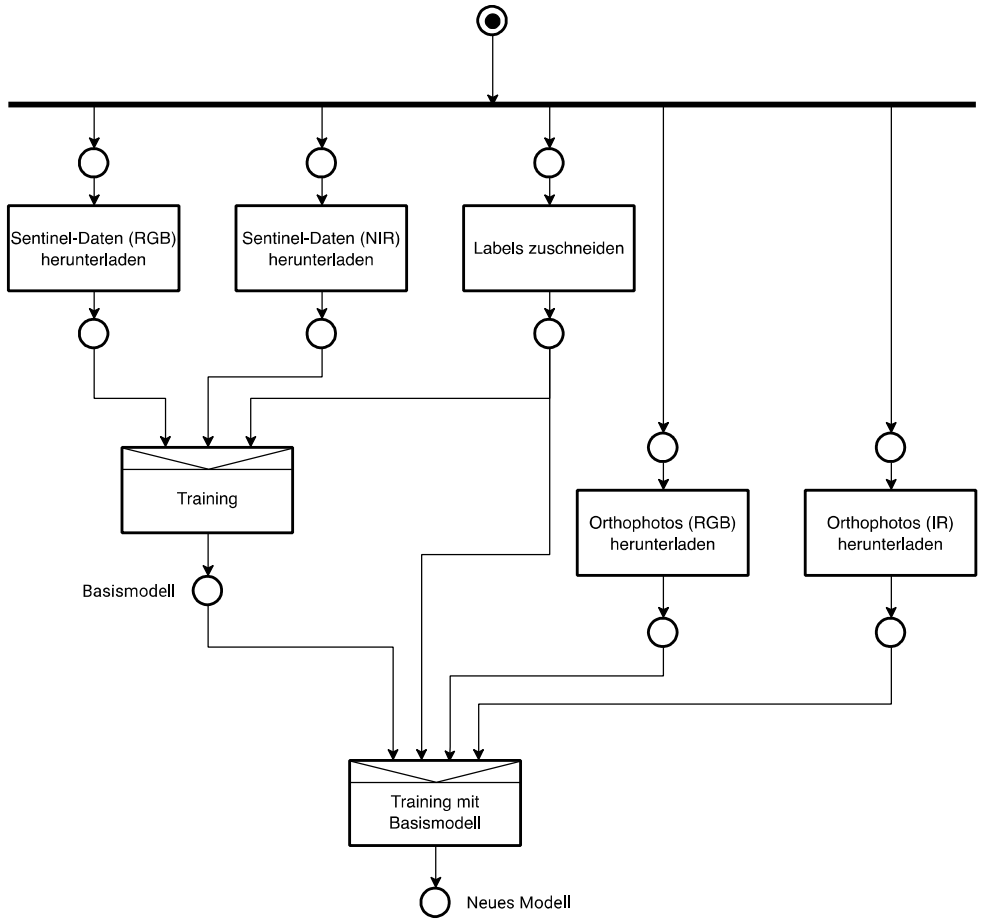
8.3.1.2 Transfer Learning

Transfer Learning ist eine Methode aus dem Bereich des Machine Learnings, bei dem ein zu einem früheren Zeitpunkt trainiertes Modell (das so genannte Basismodell) für eine neue oder artverwandte Aufgabe verfeinert wird. Das Basismodell und das verfeinerte Modell können dabei mit einem ähnlichen Datensatz oder sogar mit unterschiedlichen Datensätzen trainiert werden.

Wie mit Hilfe von Abbildung 27 bereits zu Beginn von Kapitel 8.3 beschrieben, wurde für die Evaluierung des modellzentrierten Ansatzes die Vorgehensweise gewählt, die verfügbaren FF-PV-Trainingsdaten anzupassen und dann zunächst mit einer ähnlichen Menge an Bilddaten (in diesem Fall Satellitenbilder statt Orthophotos) ein Basismodell zu trainieren. Dieses sollte später durch Training mit den Orthophotos verfeinert werden.

Im Folgenden wird dieser Ablauf im Detail beschrieben (siehe Abbildung 29). Zuerst wurden frei verfügbaren Satellitenbilder aus der Copernicus Sentinel-2-Mission als RGB und NIR (Near Infrared) aus dem Jahr 2019 in der Auflösung 10m pro Pixel für einen ausgewählten Ausschnitt heruntergeladen. Dazu wurde der Web Map Service (WMS) des Bundesamts für Kartographie und Geodäsie (BKG) verwendet.[41]

Abbildung 29: Transfer-Learning-Workflow in der erweiterten Petri-Net-Notation



Quelle: eigene Darstellung nach Aalst & Hee, 2004 und Krämer et al., 2021, Fraunhofer IGD

Parallel dazu wurden die Labels für die FF-PV-Anlagen so zugeschnitten, dass zu den Satellitenbildern passende Kacheln entstanden. Mit dieser Datengrundlage wurde ein Training durchgeführt, aus dem das Basismodell entstanden ist.

Zeitgleich wurden passende Orthophotos in RGB und IR (Infrarot) vom DOP-20 WMS des BKG in der Auflösung 20 cm pro Pixel heruntergeladen.[42] Nachdem das Training des Basismodells und

[41] https://gdz.bkg.bund.de/index.php/default/wms-deutschlandmosaik-aus-sentinel-2-daten-wms-sentinel2-de.html
© Europäische Union, enthält Copernicus Sentinel-2 Daten 2019, verarbeitet durch das Bundesamt für Kartographie und Geodäsie (BKG)

[42] https://gdz.bkg.bund.de/index.php/default/wms-digitale-orthophotos-bodenauflosung-20-cm-wms-dop.html
Bereitstellung für Bundesbehörden und Nutzungsberechtigte nach V GeoBund
Geobasisdaten: © GeoBasis-DE / BKG 2024
Nutzungsbedingungen: https://sg.geodatenzentrum.de/web_public/nutzungsbedingungen.pdf

Bereitstellung für Landesbehörden und Nutzungsberechtigte nach V GeoLänder
© GeoBasis-DE / ZSGT 2024

die Downloads abgeschlossen waren, wurde ein erneutes Training mit dem Basismodell und den Orthophotos durchgeführt, aus dem das verfeinerte Modell hervorging.

8.3.2 Evaluierungssetup

Bevor auf die Evaluierungsergebnisse eingegangen wird, wird in diesem Abschnitt ein Überblick über das Evaluierungssetup gegeben. Dabei wird insbesondere auf die Trainingsdaten, die sonstigen Vorbereitungsschritte, sowie die einzelnen Testcases eingegangen.

8.3.2.1 Trainingsdaten

Die folgenden frei verfügbaren gelabelten Daten wurden als Trainingsdaten für die prototypische Umsetzung der beiden Anwendungsfälle „Erkennung von Freiflächen PV-Anlagen“ und „Erkennung von Baumarten im Wald“ sowie dem Referenzfall „Urbane Landnutzungstypen“ verwendet.

► **Deutschlandweite Freiflächen PV-Anlagen:**

Der auf der Plattform Zenodo frei verfügbare Datensatz enthält Geostandorte und Anlagedaten von Erneuerbare-Energie-Anlagen in Deutschland, u.a. auch geoJSON-Dateien mit Polygonen für Freiflächen-Photovoltaik-Anlagen (Manske & Schmiedt, 2023):

- Binärer Datensatz mit zwei Labels (FF-PV-Anlage, keine FF-PV-Anlage)
- Labels bestehen aus exakten Polygonen.

► **Baumarten der Wälder in Nordrhein-Westfalen:**

Das Land Nordrhein-Westfalen stellt über sein Open Geodata Portal einen TIFF-Datensatz mit einer Baumartenklassifikation in NRW frei zur Verfügung (OpenGeodata.NRW, 2024)**:**

- 11 verschiedene Klassen
- Rasterdaten mit räumlicher Auflösung von 10m x 10m

► **Urbane Landnutzungstypen:**

Um die komplexe Methodik des Pseudo-Labelings detaillierter evaluieren zu können, wurde zusätzlich der „ISPRS Potsdam Datensatz“ als Benchmark-Datensatz hinzugezogen. Aus den folgenden Gründen ist eine qualitativ hochwertige Evaluierung des *Pseudo-Labelings* mit den FF-PV- und Walddatensätzen allein nicht möglich:

- Der Datensatz über die FF-PV-Anlagen besitzt zwar eine hohe räumliche Auflösung und ist recht genau. Da es sich aber um einen binären Anwendungsfall handelt (PV-Anlage / keine PV-Anlage) ist ein Rückschluss der Ergebnisse auf komplexere Szenarien nicht möglich.
- Der Baumartendatensatz von NRW ist mit 11 Klassen deutlich komplexer. Aufgrund der geringen Auflösung von 10 m pro Pixel ist er jedoch recht ungenau. Hinzu kommt, dass die Labels nur für NRW vorliegen, so dass für eine Bewertung der Ergebnisse nur wenige Daten zur Verfügung stehen.

Bereitstellung für Dritte durch die ZSGT
© GeoBasis-DE / BKG 2024 (https://www.bkg.bund.de)

Das ISPRS Potsdam Dataset ist ein hochauflösender Referenzdatensatz, der speziell für die Entwicklung und Evaluierung von Algorithmen im Bereich der Bildverarbeitung und des maschinellen Lernens, insbesondere in der Fernerkundung, erstellt wurde (ISPRS, 2018). Es besteht aus Luftbildern, die mit einer hohen räumlichen Auflösung von 5 cm x 5 cm aufgenommen wurden, sowie zugehörigen annotierten Daten, die verschiedene Klassen wie Gebäude, Straßen, Bäume und andere Landnutzungstypen umfassen. Der Datensatz eignet sich gut als Referenzdatensatz für die Entwicklung von KI-Modellen aus mehreren Gründen:

1. **Hohe Auflösung**: Die hochauflösenden Bilder ermöglichen eine präzise Analyse und Klassifikation von Objekten.
2. **Vielfältige Annotationen**: Die umfangreiche und detaillierte Annotation der Bilddaten bietet eine solide Grundlage für das Training von Modellen.
3. **Repräsentativität**: Es spiegelt reale städtische Umgebungen wider und ist somit relevant für zahlreiche Anwendungen in der Stadtplanung, Umweltüberwachung und ähnlichen Bereichen.
4. **Standardisierung**: Als allgemein anerkannter Datensatz wird er häufig in der Forschung verwendet, was den Vergleich von Modellen und Algorithmen erleichtert.

- **Luftbilder und Sentinel-2-Daten des Bundesamts für Kartografie und Geodäsie (BKG):**

Für alle Testcases wurden als zugehörige Bilddaten zu den Labels die Orthophotos in RGB und IR (Infrarot) vom DOP-20 WMS des BKG in der Auflösung 20 cm pro Pixel Bodenauflösung genutzt (siehe Fußnote 42). Beim Transfer Learning wurden für das Training des Basismodells zudem noch die Satellitenbilder aus der Copernicus Sentinel-2-Mission als RGB und NIR (Near Infrared) aus dem Jahr 2019 in der Auflösung 10m pro Pixel verwendet (siehe Fußnote 41).

8.3.2.2 Testcases

Kernziel der Studie war es, Lösungen und Handlungsempfehlungen für den Umgang mit einer geringen Anzahl von gelabelten Daten zu entwickeln, die trotz der wenigen Trainingsdaten die Entwicklung von KI-Anwendungen mit qualitativ guten Ergebnissen ermöglichen. Um mithilfe der prototypischen Umsetzung die Methodiken besser bewerten zu können, wurde eine Vielzahl unterschiedlicher Testszenarien erzeugt. Die einzelnen Testfälle sind in den nachfolgenden Tabellen Tabelle 21, Tabelle 22 und Tabelle 23 aufgeführt und wurden wie folgt thematisch kategorisiert:

- Pseudo-Labeling: Studien durchgeführt mit dem Potsdam-, FF-PV- und Walddatensatz (Tabelle 21)
- Transfer Learning: Studien durchgeführt mit dem FF-PV-Datensatz (Tabelle 22)
- Normales Training: Referenzstudien durchgeführt mit dem Potsdam-, FF-PV- und Walddatensatz (Tabelle 23)

In Tabelle 21 sind dabei alle Versuche aufgeführt, die für die Bewertung des Pseudo-Labeling Ansatzes durchgeführt wurden. Dabei steht in der ersten Spalte die Bezeichnung des Versuchs und in der zweiten Spalte der Name des Anwendungsfalls. Im Rahmen der prototypischen Umsetzung wurde der Pseudo-Labeling Algorithmus in drei Iterationen angewendet. In der ersten Iteration wurde nur mit gelabelten Daten trainiert, in der zweiten Iteration mit den gelabelten Daten und dem ersten Teil der generierten Pseudo-Labels und in der dritten Iteration mit allen gelabelten Daten und allen generierten Pseudo-Labels. In der dritten Spalte der Tabelle steht somit der Anteil verwendeter Labels für die erste Iteration bezogen auf die Menge der Gesamtlabels. In der vierten Spalte steht folglich die Menge verwendeter Pseudo-Labels für die

Iterationen 2 und 3 bezogen auf die Gesamtmenge vorhandener Labels. In der fünften Spalte ist die Anzahl an Epochen angegeben, mit welcher in jeder Iteration trainiert wurde. Und in der letzten Spalte ist angegeben, mit welchem Anteil der Daten eine anschließende Validierung durchgeführt wurde.

Tabelle 22 bezieht sich auf die Versuche des Transfer Learnings. Dafür wurde zunächst ein Basismodell trainiert. Hierzu wurden, anstatt der BKG RGB-IR Orthphotos, Senintel-2 RGB-NIR Aufnahmen zusammen mit den annotierten FF-PV-Labels verwendet. Zwar ist in diesem Fall der Anwendungsfall des Basismodells der gleiche, allerdings besitzen die genannten Sentinel-2 Aufnahmen im Gegensatz zu den BKG Orthophotos eine Auflösung von 10m x 10m. Dies erfordert ein Nachtrainieren und ist somit ein gutes Szenario für die Validierung von Transfer Learning.

Tabelle 23 bezieht sich auf das klassische Training der KI-Anwendungen. Dabei wurde für jeden Datensatz ein „normales“ Training durchgeführt ohne Verwendung zusätzlicher Methoden, wie z.B. Pseudo-Labeling. Das Training klassischer KI-Anwendungen steht zwar nicht im Vordergrund dieses Projektes. Zur Evaluierung der verschiedenen Methodiken werden allerdings Referenzwerte benötigt, die durch das „normale“ Training erzeugt werden.

Tabelle 21: Auflistung aller durchgeführten Versuche zum Pseudo-Labeling

Versuch	Use Case	Anteil Labels	Pseudo-Labels	Epochen	Validierung
Pseudo10_Pot	Potsdam	10 % aller Daten	35% + 35% aller Daten	20+20+20	20 % a. D.
Pseudo20_Pot	Potsdam	20 % aller Daten	30% + 30% aller Daten	20+20+20	20 % a. D.
Pseudo30_Pot	Potsdam	30 % aller Daten	25% + 25% aller Daten	20+20+20	20 % a. D.
Pseudo40_Pot	Potsdam	40 % aller Daten	20% + 20% aller Daten	20+20+20	20 % a. D.
Pseudo50_Pot	Potsdam	50 % aller Daten	15% + 15% aller Daten	20+20+20	20 % a. D.
Pseudo60_Pot	Potsdam	60 % aller Daten	10% + 10% aller Daten	20+20+20	20 % a. D.
Pseudo70_Pot	Potsdam	70 % aller Daten	5% + 5% aller Daten	20+20+20	20 % a. D.
Pseudo10_PV	PV	10 % aller Daten	35% + 35% aller Daten	20+20+20	20 % a. D.
Pseudo20_PV	PV	20 % aller Daten	30% + 30% aller Daten	20+20+20	20 % a. D.
Pseudo30_PV	PV	30 % aller Daten	25% + 25% aller Daten	20+20+20	20 % a. D.
Pseudo40_PV	PV	40 % aller Daten	20% + 20% aller Daten	20+20+20	20 % a. D.
Pseudo50_PV	PV	50 % aller Daten	15% + 15% aller Daten	20+20+20	20 % a. D.

Versuch	Use Case	Anteil Labels	Pseudo-Labels	Epochen	Validierung
Pseudo60_PV	PV	60 % aller Daten	10% + 10% aller Daten	20+20+20	20 % a. D.
Pseudo70_PV	PV	70 % aller Daten	5% + 5% aller Daten	20+20+20	20 % a. D.
Pseudo1000_W	Wald	3000 Labels	1000+1000 Pseudo Labels	20+20+20	-
Pseudo3000_W	Wald	3000 Labels	3000+3000 Pseudo Labels	20+20+20	-

Tabelle 22: Auflistung aller durchgeführten Versuche zum Transfer Learning

Versuch	Use-Case	Basismodell	Labels für das Finetuning	Epochen	Validierung
Transfer5_PV	PV	Sentinel PV	5% aller Daten	30	20% aller Daten
Transfer10_PV	PV	Sentinel PV	10% aller Daten	30	20% aller Daten
Transfer20_PV	PV	Sentinel PV	20% aller Daten	30	20% aller Daten
Transfer80_PV	PV	Sentinel PV	80% aller Daten	30	20% aller Daten

Tabelle 23: Auflistung aller durchgeführten „normalen“ Trainings als Referenzwerte

Versuch	Use Case	Labels	Epochen	Validierung
Train10_Pot	Potsdam	10% aller Daten	60	20% aller Daten
Train20_Pot	Potsdam	20% aller Daten	60	20% aller Daten
Train30_Pot	Potsdam	30% aller Daten	60	-
Train40_Pot	Potsdam	40% aller Daten	60	-
Train50_Pot	Potsdam	50% aller Daten	60	-
Train60_Pot	Potsdam	60% aller Daten	60	-
Train70_Pot	Potsdam	70% aller Daten	60	20% aller Daten
Train5_PV	PV	5% aller Daten	60	20% aller Daten
Train10_PV	PV	10% aller Daten	60	20% aller Daten
Train20_PV	PV	20% aller Daten	60	20% aller Daten
Train30_PV	PV	30% aller Daten	60	-
Train40_PV	PV	40% aller Daten	60	-
Train50_PV	PV	50% aller Daten	60	-
Train60_PV	PV	60% aller Daten	60	-

Versuch	Use Case	Labels	Epochen	Validierung
Train70_PV	PV	70% aller Daten	60	20% aller Daten
Train3000_W	Wald	3000 Labels	60	-
Train10000_W	Wald	10000 Labels	60	-

8.3.2.3 Vorbereitungen für das Training und die Evaluierung

In diesem Abschnitt wird auf weitere Aspekte eingegangen, die bei der prototypischen Umsetzung und deren Evaluierung eine wichtige Rolle spielen. Dabei stehen insbesondere das Augmentieren der Daten und das Hyperparametertuning der Modelle im Vordergrund. Beide Aspekte sind fundamentale Elemente bei der Entwicklung von KI-Anwendungen. Aus diesem Grund wurden für jeden zuvor in den Tabellen aufgeführten Versuch die annotierten Daten augmentiert. Dabei wurden pro Bild 3 augmentierte Bilder erzeugt, da dies bei der hohen Anzahl an durchgeführten Versuchen ein guter Trade-Off zwischen Ergebnisqualität und Rechenzeit ist. Für die Augmentierung wurden 1-3 zufällig ausgewählte geometrische und/oder visuelle Operationen durchgeführt:

- Multiply Brightness
- Gaussian Noise
- Gamma Contrast
- Multiply Saturation
- Gaussian Blur
- Flip Left-Right
- Flip Up-Down
- Rotation

Ebenso wurde für jeden Versuch ein Hyperparametertuning vorgenommen, um die ideale Konfiguration des Netzes sowie der Trainingsparameter zu finden. Nur das garantiert die bestmöglichen Ergebnisse.

Nachfolgend sind alle Parameter aufgeführt, welche konfiguriert wurden:

- Learning Rate [0.00001, 0.0001, 0.001]
- Optimizer [Adam, Nadam, RMSprop]
- Loss Function [CategoricalCrossentropy, SigmoidFocalCrossEntropy, CategoricalFocalCrossentropy]
- Batch Size [4,8,16]
- Tiefe des UNets [4 Encoder Layers, 5 Encoder Layers]
- Dropouts im UNet [nach jedem Layer ein Dropout (true/false)]

Als Grundmodell wurde für alle Versuche ein UNet verwendet. Vorherige Studien haben bereits gezeigt, dass dieses Modell gut geeignet ist, um mit geringen Datenmengen umzugehen (Kocon,

Krämer, & Würz, 2022). Alle KI-basierten Implementierungen wurden mit der Programmiersprache Python und der Keras-Bibliothek durchgeführt.

8.3.3 Evaluierungsergebnisse

Der gewählte Pseudo Labeling-Ansatz wird für die drei Use Cases Potsdam, FF-PV und Wald ausgewertet. Darüber hinaus wird der Transfer Learning-Ansatz für den FF-PV Use Case präsentiert. Die erzielten Ergebnisse werden mit Ergebnissen eines „normalen“ Trainings verglichen, bei dem ausschließlich mit den original gelabelten Daten trainiert wurde.

8.3.3.1 Pseudo-Labeling

8.3.3.1.1 Qualitative Ergebnisse

Zunächst werden die Ergebnisse des Pseudo-Labelings mit den Ergebnissen des normalen Trainings qualitativ verglichen.

Die Tabelle 24 enthält exemplarische Bildausschnitte mit unterschiedlichen Konfigurationen aus dem FF-PV Use-Case.

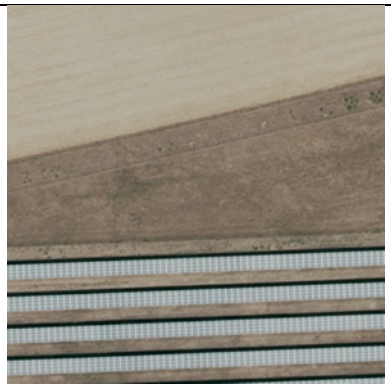
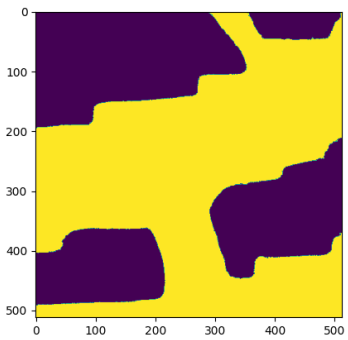
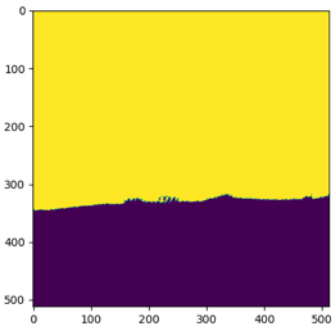
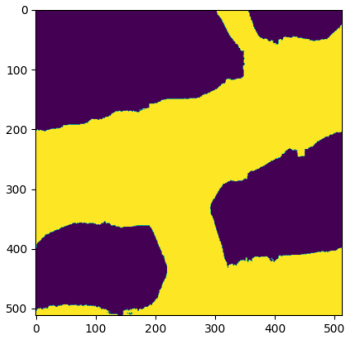
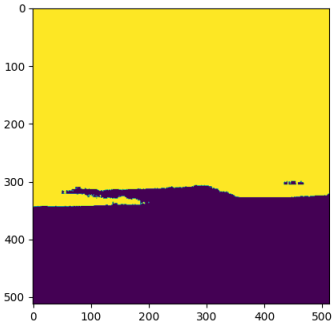
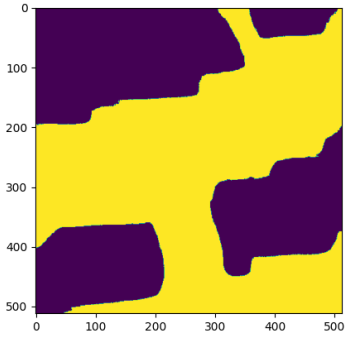
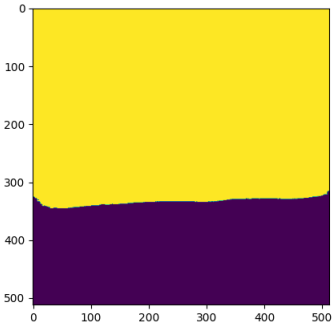
Im linken Beispiel ist bei Train70_PV (normales Training des PV Use-Cases mit 70% der zur Verfügung stehenden Daten) die stufenförmige Anordnung der FF-PV-Anlagen in der Segmentierung sichtbar (Tabelle 24, Zeile 2). Bei Pseudo70_PV (Training mit dem Pseudo-Labeling Ansatz mit 70% der zur Verfügung stehenden Daten) ist in diesem Fall eine Glättung der Objektränder feststellbar (Tabelle 24, Zeile 3). Dass das Pseudo-Labeling, insbesondere bei einer größeren Datenmenge, das Modell auch „in die Irre“ führen kann, ist im rechten Beispiel in der Konfiguration Pseudo70_PV zu erkennen. Während bei Pseudo10_PV die Objektkante noch gut prädiziert wird, weist die Kante bei der Verwendung von 70 % Referenzlabel eine gewisse Unsicherheit auf.

Zusätzlich kann festgestellt werden, dass nicht nur die Anzahl verwendeter Referenzlabel einen Einfluss auf die Ergebnisse hat, sondern auch die verschiedenen Iterationen, bei denen zusätzliche Pseudo-Label generiert werden (Tabelle 25). Ist in Iteration 2 eine glatte Kante sichtbar, ist nach der dritten Iteration wieder eine ausgefranste Kante erkennbar. Dieses Beispiel verdeutlicht zudem die Gefahr möglicher Fehlinterpretationen beim Generieren der Pseudo-Labels. Da die Pseudo-Labels im weiteren Verlauf für das Training verwendet werden, können so Fehler und Unsicherheiten im Modell entstehen.

Insgesamt wird bei diesem Use-Case deutlich, dass insbesondere beim Pseudo-Labeling bereits ein Anteil von 10 % Referenzlabel ausreichend ist, um alle Bereiche mit FF-PV-Anlagen in den hier dargestellten Beispielen zu segmentieren (Zeile 3+4). Die Hinzunahme weiterer Referenzdaten beeinflusst die Ergebnisse nicht weiter positiv. Fairerweise muss aber auch festgehalten werden, dass beim binären FF-PV Use Case auch ein normales Training mit nur 10% der Referenzlabel bereits gute Ergebnisse liefert.

Tabelle 24: Ergebnisse aus FF-PV

Dargestellt sind die Pseudo-Labeling Ergebnisse nach der 3. Iteration für zwei RGB-Beispielbilder (linke und rechte Spalte) mit 70% bzw. 10% Referenzlabel (Pseudo70_PV, Pseudo10_PV) sowie die Ergebnisse eines normalen Trainings mit 70% bzw. 10% Referenzlabel (Train70_PV, Train10_PV). Die Trainingsergebnisse werden jeweils in einem 512 x 512 Pixelbild ausgegeben. Die Einheit an den Bildachsen sind Pixel.

RGB		
Train70_PV		
Pseudo70_PV		
Train10_PV		

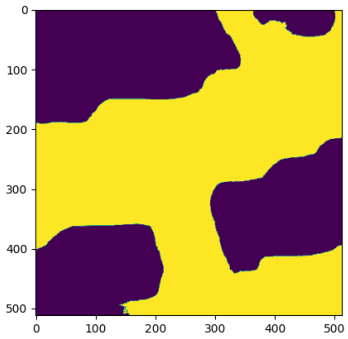
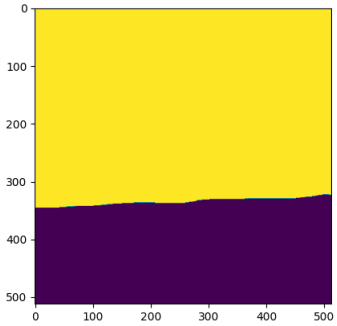
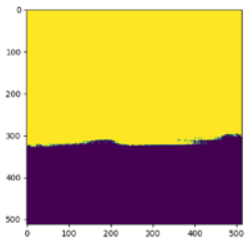
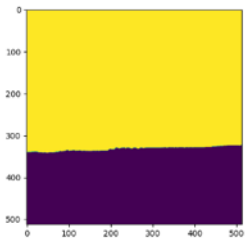
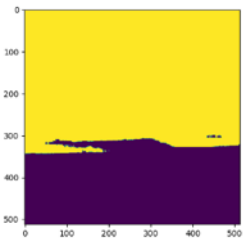
Pseudo10_PV		

Tabelle 25: Verlauf der Prädiktion des Beispiels Pseudo70_PV über die verschiedenen Iterationen hinweg

Die Trainingsergebnisse werden jeweils in einem 512 x 512 Pixelbild ausgegeben. Die Einheit an den Bildachsen sind Pixel.

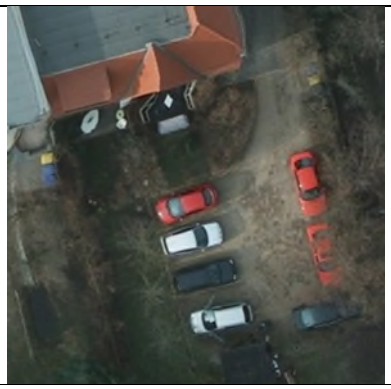
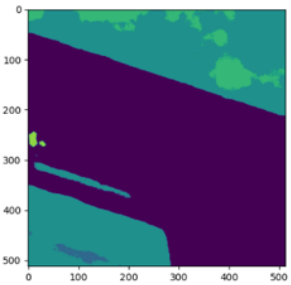
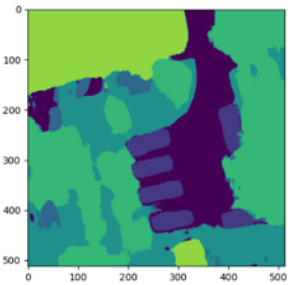
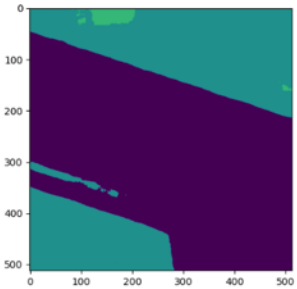
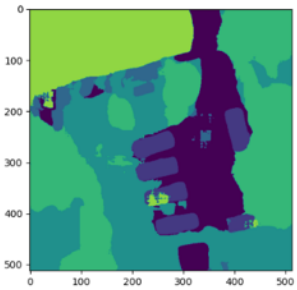
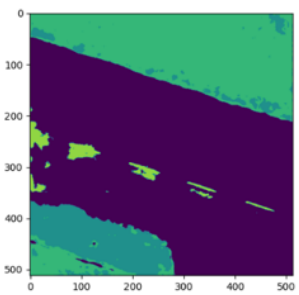
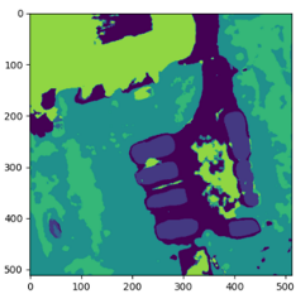
Iteration 1 (noch ohne Pseudo-Label)	Iteration 2	Iteration 3
		

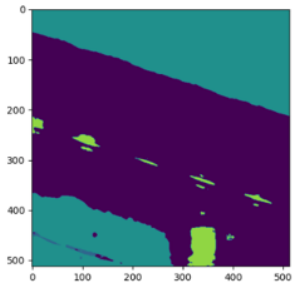
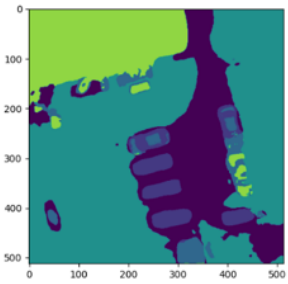
Während FF-PV eine 2-Klassen Klassifikationsaufgabe beschreibt, handelt es sich bei Potsdam und Wald um eine mehrklassige Aufgabe, wodurch die Ergebnisse besser gewertet werden können. Zunächst ist ebenfalls der beschriebene Glättungseffekt durch die Verwendung von Pseudo-Label beim Potsdam Datensatz zu erkennen (Tabelle 26). So werden die detektierten Bäume im linken Beispiel bei Train70_Pot erkannt, durch die Pseudo-Label Pseudo70_Pot jedoch herausgeglättet (Zeile 2 vs. 3). Dies zeigt den Zusammenhang mit Unsicherheiten. Objekträndern sind bei der Segmentierung besonders anfällig für Fehler und weisen eine erhöhte Unsicherheit auf. Aber auch einzelne Objekte, die keine eindeutige Signatur vorweisen können, sind besonders schwierig zu segmentieren und werden häufig falsch erkannt. In dem vorliegenden Beispiel weisen die Bäume in der oberen rechten Bildecke einerseits Mischpixel zwischen den feinen Aststrukturen mit der unterhalb befindlichen Vegetation auf, andererseits besteht zusätzlich eine spektrale und strukturelle Ähnlichkeit mit dem Untergrund. Dies erschwert die korrekte Klassenzuordnung. Durch die Verwendung der Pseudo-Label wird ungeachtet dieser Unsicherheiten die Klasse mit der größten Wahrscheinlichkeit ausgewählt. Sobald das Modell jedoch mittels der Pseudo-Label lernt, mehrdeutige Pixel auf eine bestimmte Weise zu interpretieren, kann eine damit einhergehende Informationslücke nicht geschlossen werden und es muss mit systematischen Fehlinterpretationen gerechnet werden.

Was beim Vergleich von Train10_Pot und Pseudo10_Pot deutlich wird, ist die Tatsache, dass in den Ergebnissen von Pseudo10_Pot weniger Unsicherheiten vorhanden sind als in Pseudo10_Pot. Dieses Ergebnis belegt qualitativ den positiven Einfluss von Pseudo-Labels bei einer geringen Menge annotierter Daten.

Tabelle 26: Ergebnisse des Potsdam Datensatzes

Dargestellt sind die Pseudo-Labeling Ergebnisse nach der 3. Iteration für zwei RGB-Beispielbilder (linke und rechte Spalte) mit 70% bzw. 10% Referenzlabel (Pseudo70_Pot, Pseudo10_Pot) sowie die Ergebnisse eines normalen Trainings mit 70% bzw. 10% Referenzlabel (Train70_Pot, Train10_Pot). Die Trainingsergebnisse werden jeweils in einem 512 x 512 Pixelbild ausgegeben. Die Einheit an den Bildachsen sind Pixel.

RGB		
Train70_Pot		
Pseudo70_Pot		
Train10_Pot		

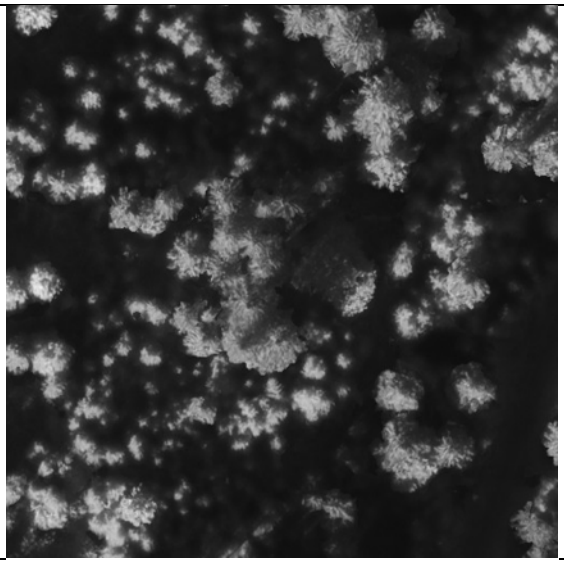
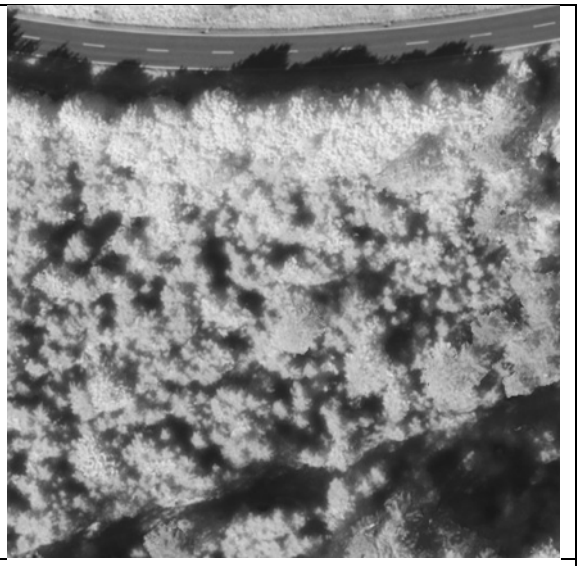
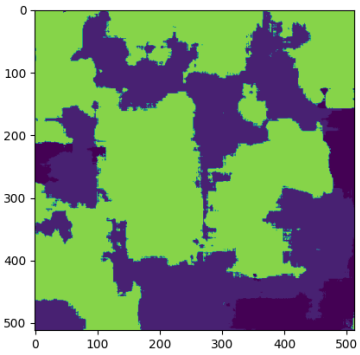
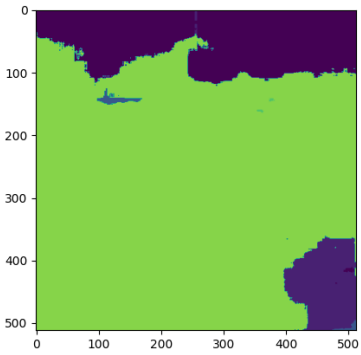
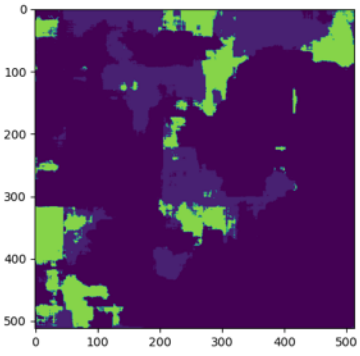
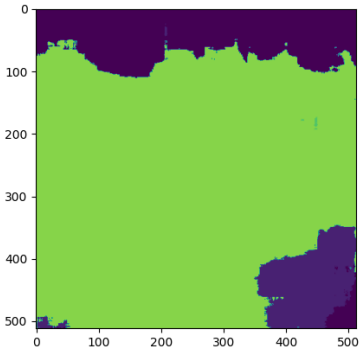
Pseudo10_Pot		

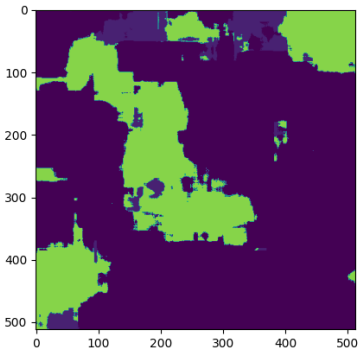
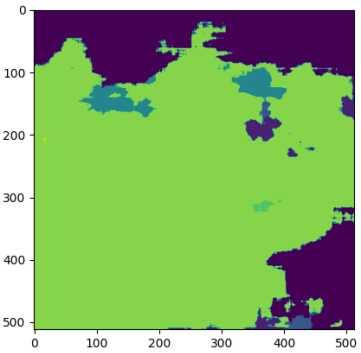
Der Wald-Use Case wurde bereits im Vorfeld als realitätsnaher Anwendungsfall identifiziert. Damit einhergehend kommen Bedingungen zum Tragen wie eine geringe Menge an Referenzlabeln sowie durch die geringere räumliche Auflösung auch Label von geringer Qualität. Daher handelt es sich bei einer Baumartenklassifikation basierend auf RGB-NIR Daten aufgrund des dichten Bewuchses und der hohen spektralen Ähnlichkeiten bereits um eine komplexe und somit schwer zu erlernende Aufgabe.

Tabelle 27 zeigt die Schwierigkeit, zuverlässige Ergebnisse in diesem Fall zu erzeugen. Insbesondere bei Waldflächen mit heterogenem Baumartenbestand ist die Qualität der Prädiktion gering (Beispiel linke Spalte,Tabelle 27), wohingegen Flächen mit homogenerer Verteilung der Baumarten deutlich bessere Ergebnisse liefern (rechte Spalte, Tabelle 27). In diesem Fall kann auch hier die Glättungswirkung der Pseudo-Label beobachtet werden, wenn nicht nur mit 3000 Referenzlabeln, sondern zusätzlich noch mit Pseudo-Labeln trainiert wird (Pseudo3000_W vs. Train3000_W). Dabei bestätigen sich die zuvor gemachten Erkenntnisse. Insbesondere beim Betrachten der Straße im oberen Bereich des rechen Bildes wird sichtbar, dass diese bei Pseudo3000_W deutlich besser vom Wald abgegrenzt wird und mit weniger Unsicherheiten verbunden ist als bei Train3000_W.

Tabelle 27: Ergebnisse des Wald Datensatzes.

Dargestellt sind die Pseudo-Labeling Ergebnisse nach der 3. Iteration für zwei NIR-Beispielbilder (linke und rechte Spalte) mit 1000 Referenzlabel (Pseudo1000_W) und 3000 Referenzlabel (Pseudo3000_W) sowie die Ergebnisse eines normalen Trainings mit 3000 Referenzlabel (Train3000_W). Die Trainingsergebnisse werden jeweils in einem 512 x 512 Pixelbild ausgegeben. Die Einheit an den Bildachsen sind Pixel.

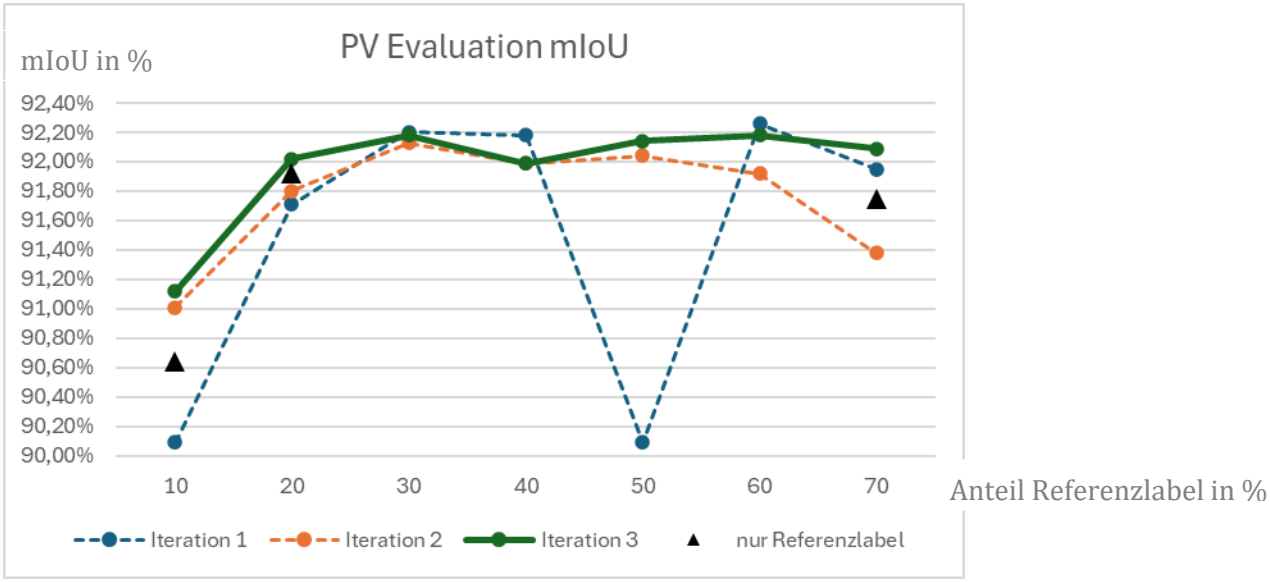
NIR		
Pseudo1000_W		
Pseudo3000_W		

Train30000_W		

8.3.3.1.2 Quantitative Ergebnisse

Die vorgestellten qualitativen Ergebnisse ermöglichen einen ersten Eindruck, jedoch erzielt die quantitative Auswertung mit unabhängigen Validierungsdaten eine umfassendere Analyse. So zeigt Abbildung 30 den Verlauf des meanIoU steigendem Anteil von Referenzlabeln beim Pseudo-Labeling und normalen Training für den FF-PV Anwendungsfall. Die schwarzen Dreiecke markieren dabei die Ergebnisse des normalen Trainings, wobei die drei Kurven die Ergebnisse einer Iteration des Pseudo-Labelings darstellen. Bereits bei einem geringen Anteil von 20 % Referenzlabel kann beim Pseudo-Labeling eine Genauigkeit von 92 % erreicht werden. Im Vergleich zum normalen Training, kann das Training mit dem Pseudo-Label Ansatz im Anwendungsfall FF-PV leicht verbesserte Ergebnisse erzielt.

Abbildung 30: Evaluation Pseudo-Labeling FF-PV-Anwendungsfall

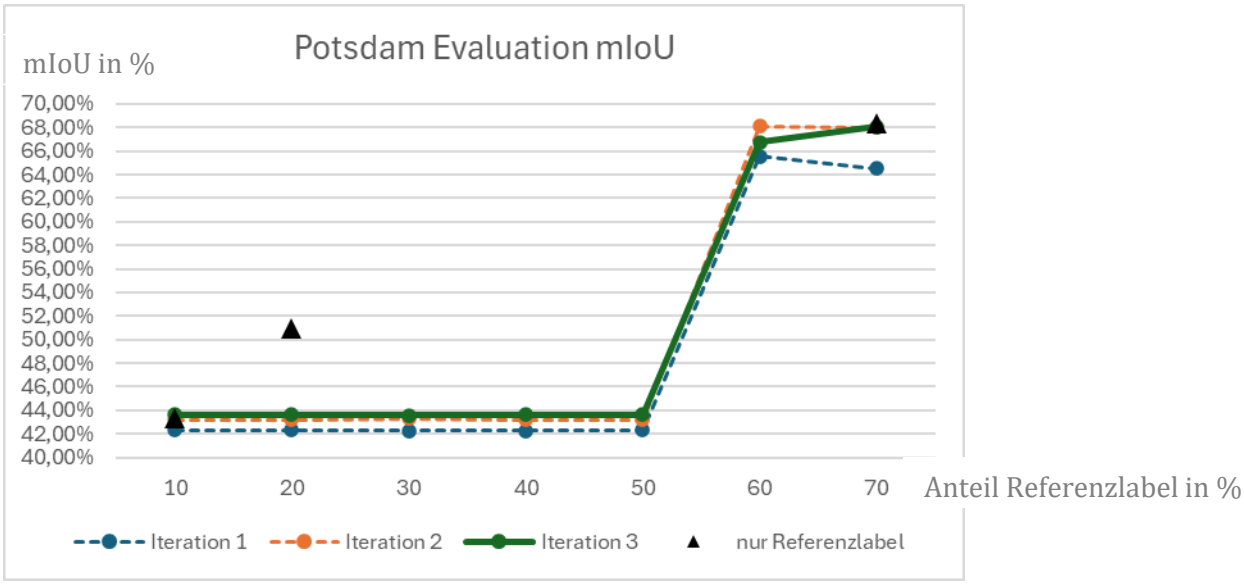


Quelle: eigene Darstellung, Fraunhofer IGD

Im Gegensatz zu den FF-PV-Ergebnissen benötigt der komplexere Potsdam Anwendungsfall mindestens 60 % Referenzlabel, um eine meanIoU von mehr als 64 % zu erreichen. Damit ist der Vorteil der Pseudo-Label deutlich geringer einzuschätzen als bei Anwendungsfall FF-PV. Allerdings ist es möglich, mit dem Pseudo-Label Ansatz sowohl mit einem Referenzanteil von 10 % als auch 70 % die Genauigkeit der Ergebnisse des alleinigen Referenzlabel-Ansatzes zu

erreichen. Somit ist durch die Verwendung der Pseudo-Label kein Nachteil entstanden. Insgesamt bleibt der meanIoU unter einem Wert von 70 %. Zu den möglichen Gründen gehören hier die teilweise mangelhafte Qualität der Referenzlabel sowie eine Abhängigkeit von den zufällig gewählten Referenzdaten. Interessant an dieser Stelle ist, dass im Gegensatz zum FF-PV-Anwendungsfall, aber auch zu den qualitativen Ergebnissen des Potsdam Datensatzes, das normale Training von Potsdam bei einer geringen Datenmenge ohne Pseudo-Labeling besser abschneidet.

Abbildung 31: Evaluation Pseudo-Labeling Potsdam Anwendungsfall



Quelle: eigene Darstellung, Fraunhofer IGD

Da für die Bewertung im Anwendungsfall Wald keine zusätzlichen Referenzlabel zur Verfügung stehen, werden stattdessen die Testergebnisse aus dem Training für die Bewertung herangezogen. Die Ergebnisse in Tabelle 28 zeigen einen großen Unterschied zwischen den beiden Metriken Accuracy und meanIoU. Dabei ist zu beachten, dass bereits beim Training und auch bei der Validierung Referenzlabel zum Einsatz kommen, die ursprünglich eine geringere räumliche Auflösung besaßen. Erschwerend kommt hinzu, dass die zu segmentierenden Klassen eine hohe Ähnlichkeit aufweisen, so dass aufgrund der unzureichenden Auflösung der Referenzlabels diese nur schwer aufgelöst werden können. Eine sinnvolle quantitative Bewertung der Ergebnisse für Wald ist daher aufgrund der geringen Datenmenge, aber auch aufgrund der geringen räumlichen Auflösung nicht möglich. Es wäre interessant die hier durchgeführten Berechnungen mit höher aufgelösten Referenzlabeln (< 1 m) durchzuführen, um zu testen, ob sich dadurch bessere Ergebnisse erzielen lassen.

Tabelle 28: Testgenauigkeiten Wald Anwendungsfall

	Pseudo Labeling		Normales Training	
Labelmenge	1000	3000	3000	10000
mIoU	44,43	35,13	23,04	25,98
Acc	88,55	91,35	68.98	72,86

8.3.3.2 Transfer Learning

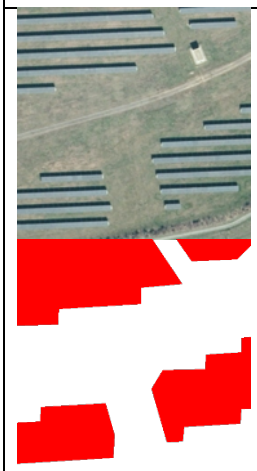
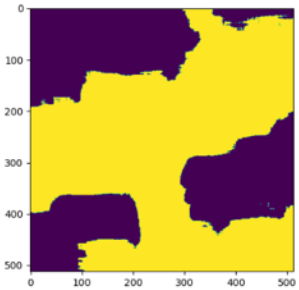
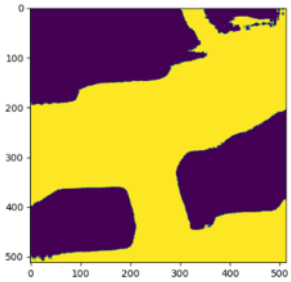
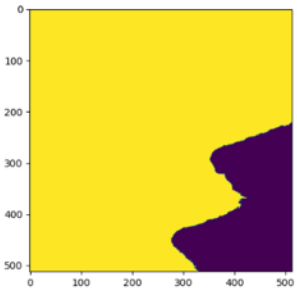
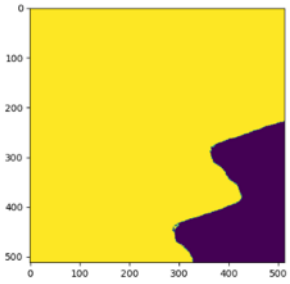
In diesem Abschnitt werden die Ergebnisse des Transfer Learnings qualitativ und quantitativ bewertet. Ziel des Transfer Learnings ist es, vor allem durch die Verwendung und das Nachtrainieren eines bereits bestehenden Basismodells, mit einer geringeren Anzahl an Daten, aber auch geringeren Dauer des Trainings, gute Ergebnisse zu erzielen. Dies ist besonders ressourcenschonend.

8.3.3.2.1 Qualitative Ergebnisse

Im Vergleich zu dem datenzentrierten Ansatz des Pseudo-Labeling enthält Tabelle 29 Beispiele für den modellzentrierten Ansatz des Transfer Learning für den Anwendungsfall FF-PV. Bei nur einem geringen Anteil von Referenzlabeln weist die Segmentierung deutlich „ausgefranste“ Objekträndern auf. Dies kann durch die Hinzunahme von mehr Referenzdaten ausgeglichen werden. Nichtsdestotrotz werden auch bei einer geringen Datenmenge die FF-PV-Anlagen gut erkannt. Besonders interessant ist an dieser Stelle die benötigte Zeit des Trainings, um diese Ergebnisse zu erzielen, was in der folgenden quantitativen Bewertung betrachtet wird.

Tabelle 29: Ergebnisse für Transfer Learning bei FF-PV.

Dargestellt sind die Transfer Learning Ergebnisse anhand der zwei Trainingspaare (RGB + Ground Truth Map) in der linken Spalte mit einem Anteil von jeweils 10 % Referenzlabel (mittlere Spalte) und 70 % Referenzlabel (rechte Spalte). Die Trainingsergebnisse werden jeweils in einem 512 x 512 Pixelbild ausgegeben. Die Einheit an den Bildachsen sind Pixel.

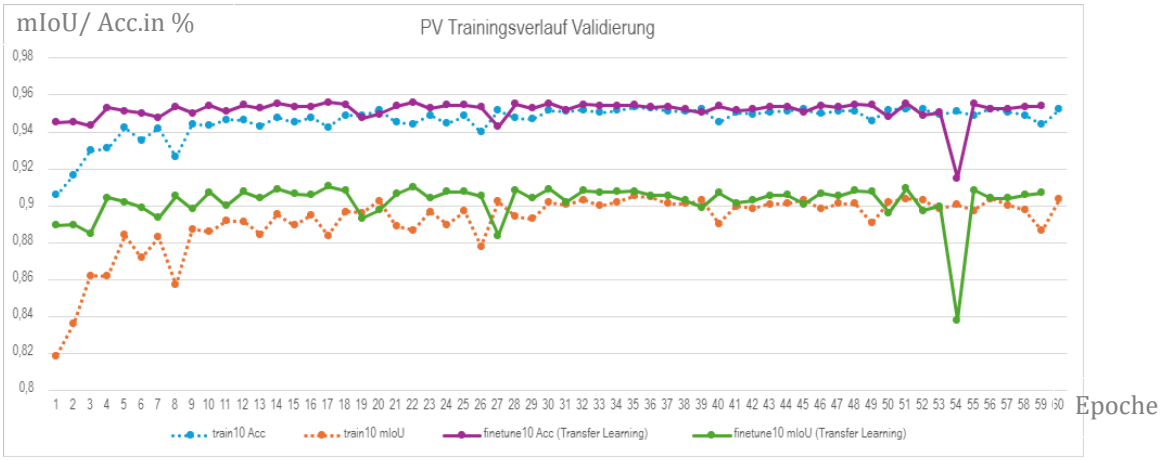
RGB	Transfer10_PV	Transfer70_PV
		
		

8.3.3.2.2 Quantitative Ergebnisse

Bei der Analyse des Einflusses des Transfer Learnings ist anhand der Abbildung 32 zu erkennen, dass bei der Verwendung eines vortrainierten Basismodells deutlich schneller eine Sättigung erzielt wird. Durch die Nutzung von Sentinel-2 Daten ist das Basismodell bereits in der Lage, charakteristische Merkmale aus Luftbildern zu extrahieren. Somit ermöglicht die Verwendung des Transfer Learning eine verkürzte Trainingszeit und ist damit deutlich ressourcenschonender. Bereits nach der jeweils 5. Epoche werden beim Transfer Learning die jeweiligen Maxima erreicht (siehe lila und grüne Kurven in Abbildung 32). Beim normalen Training werden diese Genauigkeiten erst nach 30-40 Epochen erreicht (siehe orange und blaue Kurve in Abbildung 32). Bedenkt man nun, dass beim Training eine Epoche, je nach KI-Modell und Datenmenge, mehrere Stunden bis Tage benötigt, wird die Stärke des Transfer Learnings bezüglich Ressourcenschonung sehr deutlich. Hinzu kommt, dass die Ergebnisse des Transfer Learnings besser sind als die des normalen Trainings.

Die hier vorgestellten quantitativen Ergebnisse dienten zur Verdeutlichung der relevantesten Beobachtungen der umgesetzten Methodiken. Die Ergebnisse aller durchgeführten Versuche sind im Zwischenbericht 3 aufgelistet (Klien, et al., unveröffentlicht).

Abbildung 32: Vergleich der Ergebnisse zwischen normalem Training (trainXXAcc) und Transfer Learning (finetuneXXAcc)



Quelle: eigene Darstellung, Fraunhofer IGD

8.3.3.3 Auswahl des KI-Modells, Hyperparametertuning und Datenaugmentierung

Auch wenn die Auswahl des KI-Modells, das Hyperparametertuning oder das Datenaugmentieren nur kleine Elemente der prototypischen Umsetzung sind, sind diese Aspekte elementar wichtig beim Training von KI-Anwendungen. Bereits in (Kocon, Krämer, & Würz, 2022) wurde die Wichtigkeit der Auswahl eines passenden KI-Modells gezeigt. In dieser Studie zur Waldtypenklassifizierung erzielte das UNet eine 14 % bessere Accuracy als ein FCN-8 Netz. Dies veranschaulicht sehr deutlich, dass Daten allein keine Garantie sind für gute Ergebnisse, wenn das Modell nicht passt.

Ein wichtiger Aspekt dabei ist auch das Hyperparametertuning. Die Parameter beeinflussen Details des Netzes sowie das Training. Wie bereits zuvor beschrieben, wurden beim PV Use-Case MeanIoUs von ca. 92 % erreicht. Wie wichtig für dieses gute Ergebnis das Hyperparametertuning ist, zeigt ein Testdurchlauf mit „falschen“ Parametern, bei dem eine

MeanIoU von nur 16% erzielt wurde. Dabei wurden die exakt gleichen Daten verwendet; lediglich die Hyperparameter waren anders. Da es nicht „die eine“ ideale Hyperparameterkonfiguration gibt, ist es wichtig, die Parameter immer individuell zu optimieren.

Die Wichtigkeit vom Datenaugmentieren wurde ebenfalls in (Kocon, Krämer, & Würz, 2022) dargestellt. In den Studien konnte allein durch das Augmentieren eine Verbesserung der Accuracy um bis zu 16% aufgezeigt werden.

8.4 Fazit und Empfehlungen

Eckpunkte für die Entwicklung des Vorgehensmodells

Ein wesentlicher, hervorzuhebender Aspekt für gute Modellbildung ist die Notwendigkeit, dass überhaupt eine bestimmte Menge an gelabelten Daten zur Verfügung steht. Stehen keine gelabelten Daten zur Verfügung, besteht die Möglichkeit, Daten synthetisch zu generieren (siehe Abbildung 24 in Kapitel 8.2). Dabei werden allerdings nicht die vorhandenen Bilddaten gelabelt, sondern künstlich neue Daten erzeugt. Dieses Vorgehen ist mit einem hohen Aufwand und mit dem Risiko verbunden, dass diese Daten die Realität möglicherweise nicht genau widerspiegeln. Dieser Aspekt wurde im Rahmen des Projekts auch in einem der Community-Workshops mit dem Vortrag „Daten als das neue Gold“ hervorgehoben. Ähnlich wie Gold lassen sich gelabelte Daten nicht einfach generieren. Für ML-Anwendungen bedeutet das im Kern: Ohne qualitativ hochwertige Daten entfaltet KI nicht ihr volles Potenzial und bleibt weitgehend ineffektiv.

Doch wie kann die Qualität annotierter Daten bewertet werden? Die in Kapitel 4.3 dokumentierten Rechercheergebnisse haben gezeigt, dass es zwar einige vielversprechende Ansätze gibt, aber bisher noch kein standardisiertes Verfahren zur Beschreibung der Qualität von Annotationen und Labels den Weg in die breite Anwendung gefunden hat. Damit gibt es auch noch keine Grundlage für eine einheitliche qualitative Bewertung von Annotationen, bevor mit dem Training von ML-Modellen begonnen wird. Meist wird die Qualität der Annotation erst bei der Auswertung und Validierung der Trainingsergebnisse deutlich. Der Faktor Qualität wird daher in dieser ersten Version des Entscheidungsbaums nicht als Entscheidungskriterium berücksichtigt; eine Erweiterung in diesem Punkt wäre aber eine weitere sinnvolle Entscheidungsunterstützung für die Entwickler*innen.

Prototypische Umsetzung

Für die prototypische Umsetzung wurden zwei Anwendungsfälle (*FF-PV* und *Wald*) ausgewählt, für die im begrenzten Umfang gelabelte Daten zur Verfügung standen. Der Fokus bei der Umsetzung lag weiterhin auf dem Ziel, eine hohe Effizienz im Entwicklungsprozess von ML-Anwendungen zu erreichen. Ein wichtiges Kriterium für die Auswahl der Annotationsmethoden war daher deren Automatisierbarkeit und Skalierbarkeit sowie die Möglichkeit, auch ohne Fachwissen gute Trainingsergebnisse zu erzielen. Auf Basis dieser Ausgangssituation wurden entlang des in Kapitel 8.2 vorgestellten Entscheidungsbaums die beiden Methoden Pseudo-Labeling und Transfer Learning für die prototypische Umsetzung ausgewählt. Für beide Methoden wurde ein entsprechender Workflow auf der eigens dafür eingerichteten Experimentierplattform aufgesetzt, um deren Effizienz, Genauigkeit und Skalierbarkeit im Rahmen der exemplarischen Umsetzung der Anwendungsfälle zu evaluieren.

Die Ergebnisse aus beiden Anwendungsfällen *FF-PV* und *Wald* für das Pseudo-Labeling zeigen, dass insbesondere bei geringer Verfügbarkeit von gelabelten Daten bereits mit einem geringen

Anteil von Referenzlabels gute Ergebnisse erzielt werden können, was die Methode effizient macht. Jedoch besteht die Gefahr, dass bei Pseudo-Labels systematische Fehlinterpretationen entstehen, die die Modellgenauigkeit beeinträchtigen, da sie zu Unsicherheiten führen können. In den Anwendungsfällen wurde der positive Einfluss von Pseudo-Labels bei der Segmentierung von Bildern bestätigt, insbesondere bei binären Klassifikationsaufgaben wie der Erkennung von FF-PV-Anlagen.

Transfer Learning zeigte sich als ressourcenschonende Methode, da es durch die Nutzung eines vortrainierten Basismodells die Trainingszeit signifikant verkürzt. Anhand der exemplarischen Umsetzung für den *FF-PV*-Anwendungsfall konnte gezeigt werden, dass die Ergebnisse im Vergleich zu einem normalen Training nicht nur schneller erreichbar waren, sondern auch präziser, was die Eignung des Ansatzes für ähnliche Anwendungsfälle bestätigt.

Ein zentraler Punkt in der Diskussion ist die Bedeutung der Wahl von ML-Modellen, Hyperparametertuning und Datenaugmentierung, die sich in der praktischen Umsetzung als entscheidende Faktoren für die Qualität der Ergebnisse erwiesen haben. Die Ergebnisse aus Kapitel 8 heben daher die Notwendigkeit hervor, diese Aspekte bei jeder Entwicklung von ML-Anwendungen zu berücksichtigen, um die bestmöglichen Resultate zu erzielen.

Empfehlungen

Insgesamt zeigt die prototypische Umsetzung des Entscheidungsbaums, dass eine sinnvolle Entscheidungsunterstützung bei einer entsprechenden Ausgangslage für ein ML-Entwicklungsvorhaben möglich ist und dass daten- und modellzentrierte Ansätze effektiv kombiniert werden können, um die Herausforderungen der Arbeit mit begrenzten Datenmengen im Bereich der Umweltforschung zu adressieren. Allerdings mussten aus Gründen der Fokussierung bei der Entwicklung des Vorgehensmodells einige Aspekte außer Acht gelassen bzw. konnten im Rahmen der prototypischen Umsetzung nicht alle Annahmen überprüft und evaluiert werden. Es bleiben daher noch einige offene Fragen, die in weitergehenden Untersuchungen adressiert werden sollten und die zu einer kontinuierlichen Verbesserung und Erweiterung des Entscheidungsbaums beitragen können.

Der Entscheidungsbaum eröffnet die Möglichkeit, die (im Rahmen dieser Studie) nur für zwei Anwendungsfälle getesteten Methoden auf andere Datensätze und Anwendungsfälle zu übertragen. Hierbei ist es entscheidend, die Übertragbarkeit der Ansätze sorgfältig zu prüfen, um sicherzustellen, dass die Ergebnisse in unterschiedlichen Kontexten valide bleiben. Für einige spezifische Szenarien könnte eine universelle Anwendbarkeit der Methoden nicht gegeben sein, weshalb gezielte Anpassungen notwendig sind. Eine weitergehende Systematisierung und Spezifizierung bestimmter Anwendungsfälle sind daher empfehlenswert, um an dieser Stelle eine differenzierte Betrachtung zu unterstützen.

Der Einsatz von Fachwissen im Prozess der Annotation und Verifikation von Daten und Ergebnissen im maschinellen Lernen ist ein wichtiger Faktor zur Sicherung der Modellqualität, ist aber auch meist nur mit hohem Aufwand und Effizienzverlusten realisierbar. Für die prototypische Umsetzung wurden Methoden ausgewählt, die ohne den Einsatz von Fachwissen verwendet werden können, was potenziell Zeit und Kosten im Gesamtprozess einspart. Ohne die qualitative Überprüfung durch Expertinnen*Experten besteht jedoch ein höheres Risiko von Fehlbewertungen bei der Annotation und damit auch eine mögliche Verstärkung des Bias beim Training des Modells. Bei der Auswahl von Methoden ohne Nutzung von Fachwissen ist es daher

wichtig, anwendungsfallspezifische Abschätzungen der Effekte vorzunehmen, d.h. abzuwägen, in welchen Fällen die Risiken von Fehlbewertungen und Verzerrungen ggf. akzeptiert werden können. Alternativ wird dann ggf. einer der kollaborativen Ansätze zur Datenannotation (wie z.B. Active Learning) empfohlen, um Fehlinterpretationen zu vermeiden bzw. zu reduzieren.

Weiterhin ist es wünschenswert, den in dieser Studie entwickelten ersten Entwurf des Entscheidungsbaums schrittweise weiterzuentwickeln und zu erweitern. Ziel ist es, möglichst viele Entscheidungskriterien zu integrieren, die für eine qualitativ hochwertige und nachhaltige Modellbildung von Bedeutung sind. In diesem Zusammenhang sollen drei wesentliche Punkte hervorgehoben werden:

- **Qualität der Labels / Annotationen**: Qualitativ hochwertige Annotationen sind entscheidend für eine gute Modellbildung. Bisher lässt sich dieser Aspekt jedoch nur schwer formal erfassen. In der Fachcommunity existieren bereits verschiedene Ansätze und Initiativen, die sich dieser Herausforderung widmen. Diese Bemühungen sollten personell und/oder finanziell unterstützt werden, um zeitnah eine formale Grundlage, wie beispielsweise ein einheitliches „Qualitätsprüfsiegel“ oder einen verpflichtenden Metadatenstandard zur Bewertung der Labelqualität zu schaffen. Auf dieser Basis könnte der Entscheidungsbaum weiterentwickelt werden.
- **Berücksichtigung anderer Datentypen**: Zu Beginn des Projektes wurde festgelegt, dass der Fokus auf Methoden zur Auswertung von Fernerkundungsdaten (Luftbilder, Satellitendaten, Punktwolkendaten) liegt. Eine breitere Anwendbarkeit des Vorgehensmodells auf weitere Datentypen erfordert die Berücksichtigung von Ansätzen, die über die in dieser Studie untersuchten, bildbasierten Ansätze hinausgehen. Es wird daher empfohlen, weitere Studien durchzuführen, die sich in ähnlicher Form mit der Identifikation und Bewertung von Methoden und Ansätzen für Anwendungsfälle mit textuellen und tabellarischen Daten beschäftigen, und diese Ergebnisse als Erweiterung in das Vorgehensmodell aufzunehmen.
- **Einbindung von ökonomischen Entscheidungsmodellen**: Wirtschaftliche Aspekte und ökonomische Entscheidungsmodelle sind in der Literaturarbeit (vgl. Kapitel 3) und auch bei der Potential- und Wirkungsanalyse (vgl. Kapitel 5 und 6) im Projekt behandelt worden, wurden aber aus Kapazitätsgründen bei der Entwicklung des Vorgehensmodell nicht berücksichtigt. Hier werden weitergehende Studien zum Thema Rechenzeit und Ressourceneinsatz für die Entwicklung von ML-Anwendungen empfohlen, auch unter dem Aspekt der Ressourcenschonung und Nachhaltigkeit. Zu diesem Thema sind derzeit kaum anwendbare Studien verfügbar und damit auch keine Grundlage, von der ein standardisiertes Vorgehen abgeleitet werden könnte. In jedem Fall ist es wünschenswert und damit auch eine Empfehlung, dass diese Aspekte in kommenden Forschungsvorhaben eine stärkere Beachtung finden sollten und letztendlich als wichtige Entscheidungspunkte bei einer Weiterentwicklung des Vorgehensmodells mit aufgenommen werden.

9 Zusammenfassung

9.1 Qualitätsgesicherte und interoperable Daten als Engpass im Umweltbereich

Die derzeit bestehenden technischen, rechtlichen und infrastrukturellen Barrieren beschränken die sektorübergreifende Nutzung von Umweltdaten und umweltrelevanten Daten (einschließlich Annotationen). Herausfordernd gestalten sich dabei insbesondere Defizite in den Bereichen Datenqualität, Interoperabilität und Datenstandardisierung. Das grundlegende Problem ist derzeit noch immer der erschwerte Zugang zu Umweltdaten und umweltrelevanten Daten. Dabei bilden die FAIR-Prinzipien einen zentralen Aspekt. Ein FAIRer Umgang erleichtert das Auffinden (*Findability*), und führt zu einer Verbesserung des Zugangs (*Accessibility*), der Interoperabilität (*Interoperability*) und der Nachnutzung (*Reuse*). FAIR bedeutet dabei nicht notwendigerweise „Open" – auch FAIRe Daten können immer noch eingeschränkt zugänglich und nutzbar sein.

Die Verbesserung der übergreifenden Datennutzung erfordert eine Kombination aus strukturierter und interoperabler Datenspeicherung mit notwendigen Sicherheits- und Datenschutzmaßnahmen, welche eine Transparenz- und Effizienzsteigerung im Umgang mit Daten und eine Stärkung des Innovationspotentials und des wissenschaftlichen Fortschritts ermöglicht. Um die Nutzbarkeit der Daten zu verbessern, ist es zudem unabdingbar, dass die Bereitschaft zum Datenteilen gesteigert wird. Dies kann gelingen, indem existierende Hürden abgesenkt und die Mehrwerte des Datenteilens klar herausgestellt werden.

Der bessere Zugang zu Umweltdaten und umweltrelevanten Daten führt insbesondere in den Umweltwissenschaften zu einer Steigerung des Innovationspotentials. Annotierte Daten spielen dabei eine besondere Rolle, da qualitativ hochwertige Daten einen direkten Einfluss auf das Training von KI-Modellen, deren Leistungsfähigkeit und Ergebnisqualität haben.

9.2 Potentiale und Synergieeffekte

Die Analyse der Potenziale in den untersuchten Sektoren zeigt ein differenziertes Bild. In der **Landwirtschaft** sind viele digitale Potenziale bereits weitgehend ausgeschöpft. Die Vielfalt und Heterogenität der Anwendungsbereiche in diesem Sektor schränkt die Wiederverwendbarkeit der Daten erheblich ein. Es bestehen aber signifikante Möglichkeiten, insbesondere im Hinblick auf einen effizienteren und ressourcenschonenderen Einsatz von Düngemitteln oder Pestiziden.

Im **forstwirtschaftlichen Sektor** zeigen sich hohe Potenziale, vor allem im Kontext der Digitalisierung und der datenbasierten Entscheidungsunterstützung. Die Voraussetzung hierfür sind jedoch eine sektorweite Standardisierung und Bereitstellung insbesondere von Inventurdaten, sowie die Bereitschaft aller beteiligten Stakeholder zum Datenteilen.

Im Bereich **Biodiversität** hingegen erscheinen die Potenziale vergleichsweise gering, da es kaum Möglichkeiten zur Mehrfachnutzung der Daten gibt und direkte wirtschaftliche Vorteile in der Regel ausbleiben.

Sektorübergreifende Synergieeffekte bestehen insbesondere im Bereich der **Fernerkundungsdaten**. Diese bieten vielversprechende Ansätze für sektorübergreifende Anwendungen und sollten in zukünftigen Projekten gezielt weiterentwickelt werden.

9.3 Anknüpfungspunkte für weitere Forschungsvorhaben

Die Ergebnisse der Studie bieten eine solide Grundlage für **weitere Forschungsvorhaben**. Zentrale offene Fragestellungen beinhalten:

- **Erweiterung der Untersuchung auf andere Datentypen:** Der Schwerpunkt des Vorhabens wurde auf die Auswertung von Fernerkundungsdaten gelegt. Für eine breitere Anwendbarkeit sollten weitere Studien zu Methoden (für textuelle und tabellarische Daten) durchgeführt und in das entwickelte Vorgehensmodell integriert werden.
- **Entwicklung eines einheitlichen Verfahrens zur Beurteilung der Datenqualität für das Training:** Qualitativ hochwertige Annotationen sind entscheidend für die Modellbildung, jedoch schwer formal messbar. Dies betrifft insbesondere die Qualität der Referenzlabel. Derzeit existieren bereits Bestrebungen innerhalb der Fachcommunity zur Entwicklung von Ansätzen, die sich mit der Bewertung der Labelqualität befassen. Diese Initiativen sollten unterstützt werden, um formale Standards wie ein Qualitätsprüfsiegel oder Metadatenstandards zu etablieren und den Entscheidungsbaum weiterzuentwickeln.
- **Integration von Kriterien der Nachhaltigkeit und Wirtschaftlichkeit in ökonomische Entscheidungsmodelle:** Wirtschaftliche Aspekte und ökonomische Entscheidungsmodelle wurden im Projekt zwar analysiert, aber nicht in die Entwicklung des Vorgehensmodells einbezogen. Weitere Studien zum Ressourceneinsatz und Nachhaltigkeit bei ML-Anwendungen sind notwendig, um diese Aspekte künftig als Entscheidungsgrundlage im Vorgehensmodell zu berücksichtigen.

10 Quellenverzeichnis

Aalst, W. van der, & Hee, K. M. van. (2004). *Workflow management: Models, methods and systems* (1. MIT Press paperback ed). MIT Press.

Abid, N. M. (2021): Burnt Forest Estimation from Sentinel-2 Imagery of Australia using Unsupervised Deep Learning. *Digital Image Computing: Techniques and Applications (DICTA)*, 01–08. https://doi.org/10.1109/DICTA52665.2021.9647174

Ahman, Z. (2024): *How Machine Learning is Advancing Precision Agriculture*. https://www.tabsgi.com/how-machine-learning-is-advancing-precision-agriculture/ (Abgerufen am 08.08.2024)

AI for Good. (2023): *Detection and attribution of biodiversity change: a role for AI | AI for Biodiversity*. https://www.youtube.com/watch?v=WAoUxRzzhZE (Abgerufen am 08.08.2024)

Al-Dweik, A., Mayhew, M., Muresan, R., Shirook, A., & Shami, A. (2017): Using Technology to Make Roads Safer: Adaptive Speed Limits for an Intelligent Transportation System. *IEEE Vehicular Technology Magazine*, DOI:10.1109/MVT.2016.2634462

Algorithmwatch. (2023): *Moderatoren: Ausgebeutet, um KI zu trainieren.* https://sustain.algorithmwatch.org/moderatoren/ (Abgerufen am 11.11.2024)

Alhazmi, K., Walaa, A., Indrek, S., Lara, P., & Martin, S. (kein Datum): Effects of annotation quality on model performance. *International Conference on Artificial Intelligence in Information and Communication (ICAIIC)*, 10.1109/icaiic51459.2021.9415271

Arellano, A., Cifuentes, L., & Rios, C. (2020): *Las zonas oscuras de la evaluación ambiental que autorizó "a ciegas" el megaproyecto de Google en Cerrillos*. https://www.ciperchile.cl/2020/05/25/las-zonas-oscuras-de-la-evaluacion-ambiental-que-autorizo-a-ciegas-el-megaproyecto-de-google-en-cerrillos/ (Abgerufen am 11.09.2024)

Bachmann, N., Tripathi, S., Brunner, M., & Jodlbauer, H. (2022): The Contribution of Data-Driven Technologies in Achieving the Sustainable Development Goals. *Sustainability, 14*(5). DOI: https://doi.org/10.3390/su14052497

Baldé, C. P., Kuehr, R., Yamamoto, T., McDonald, R., D'Angelo, E., Althaf, S., Bel, G., Deubzer, O., Fernandez-Cubillo, E., Forti, V., Gray, V., Herat, S., Honda, S., Iattoni, G., Khetriwal, D.S., di Cortemiglia, V.L., Lobuntsova, L., Nnorom, I., Pralat, N., Wagner, M. (2024): *Global E-waste Monitor.* Geneva/Bonn: International Telecommunication Union (ITU) and United Nations Institute for Training and Research (UNITAR). https://ewastemonitor.info/wp-content/uploads/2024/03/GEM_2024_18-03_web_page_per_page_web.pdf

BAM. (2022): *Fire Science: Neues Brandmanagement- system soll Gefahr von Waldbränden deutlich redzuieren.* https://www.bam.de/Content/DE/Pressemitteilungen/2022/Infrastruktur/2022-03-31-waldbraende.html (Abgerufen am 04.08.2024)

Bansal, M. A., Sharma, R., Kathuria, M. (2022): *A systematic review on data scarcity problem in deep learning: solution and applications.* ACM Computing Surveys (CSUR), 54(10s), 1-29. DOI: https://doi.org/10.1145/3502287

Bhowmik, P., & Arabinda, S. (2021): A Data-Centric Approach to Improve Machine Learning Model's Performance in Production. *International Journal of Engineering and Advanced Technology*, https://doi.org/10.35940/ijeat.a3201.1011121

Binetti, M. S., Massarelli, C., & Uricchio, V. F. (2024). Machine Learning in Geosciences: A Review of Complex Environmental Monitoring Applications. *Machine Learning and Knowledge Extraction*, *6*(2), 1263–1280. https://doi.org/10.3390/make6020059

Bitkom. (2023): *Stellungnahme Forschungsdatengesetz: Antwort des Bitkom auf die Konsultation des BMBF zum Forschungsdatengesetz,*. https://www.bitkom.org/sites/main/files/2023-04/2304Bitkom-Stellungnahme-Forschungsdatengesetz.pdf (Abgerufen am 04.08.2024)

BMUV. (2020): *Die Umweltpolitische Digitalagenda: Wie ein Problem zur Lösung wird.* https://www.oekologische-plattform.de/wp-content/uploads/2020/03/digitalagenda_bf.pdf (Abgerufen am 03.06.2025)

BMUV. (2023): *Künstliche Intelligenz für Umwelt und Klima*. Von Künstliche Intelligenz als Chancentreiber: https://www.bmuv.de/WS5881 (Abgerufen am 03.06.2025)

Branco, V. V., Correia, L., & Cardoso, P. (2023): The use of machine learning in species threats and conservation analysis. *Biological Conservation, 283*. https://doi.org/10.1016/j.biocon.2023.110091

Callaghan, M., Schleussner, C.-F., Nath, S., Lejeune, Q., Knutson, T. R., Reichstein, M., Hansen, G., Theokritoff, E., Andrijevic, M., Brecha, R.J., Hegarthy, M., Jones, C., Lee, K., Lucas, A., van Maanen, N., Menke, I., Pfleiderer, P., Yesil, B., Minx, J.C. (2021): Machine-learning-based evidence and attribution mapping of 100,000 climate impact studies. *Nature Climate Change, 11*, 966-972. https://doi.org/10.1038/s41558-021-01168-6

Carrasco, V., Arenas, J., Huijse, P., Espejo, D., Vargas, V., Viveros-Munoz, R., Poblete, V.; Venier, M., Suarez, E. (2023): Application of Deep Learning to Enforce Environmental Noise Regulation in an Urban Setting. *Sustainability, 15*(4). https://doi.org/10.3390/su15043528.

Catalli, F., Hochenegger, F., Reitz, T., Klien, E., & Kocon, K. (unveröffentlicht): *Bedarfsermittlung, Potential- und Wirkungsanalyse annotierter Umweltdaten.* Umweltbundesamt, Dessau-Roßau.

CBRE. (2024): *Global Data Center Trends 2024*. https://www.cbre.com/insights/reports/global-data-center-trends-2024 (Abgerufen am 03.06.2025)

Chadli, K., Botterwreck, G., Saber, T (2024): The Environmental Cost of Engineering Machine Learning-Enabled Systems: A Mapping Study. In Proceedings of the 4th Workshop on Machine Learning and Systems (pp. 200-207). https://doi.org/10.1145/3642970.3655828

Chen, J., Liang, X., Liu, Z., Gong, W., Chen, Y., Hyyppä, J., Kukko, A., Wang, Y. (2024): Tree species recognition from close-range sensing: A review. *Remote Sensing of Environment, 313*. https://doi.org/10.1016/j.rse.2024.114337

Chen, L., Zhang, Y., Lin, Y., Jiang, M., Huang, Y., & Lei, Y. (2021): Consistency-based semi-supervised learning for point cloud classification. 2021 4th International Conference on Pattern Recognition and Artificial Intelligence (PRAI), 440-445. https://doi.org/10.1109/prai53619.2021.9551055

Chen, X., Mersch, B., Nunes, L., Marcuzzi, R., Vizzo, I., Behley, J. (2022): Automatic Labeling to Generate Training Data for Online LiDAR-Based Moving Object Segmentation. IEEE Robotics and Automation Letters, 7(3), 6107–6114. https://doi.org/10.1109/LRA.2022.3166544

Cowls, J., Tsamados, A., Taddeo, M., & Floridi, L. (2021): A definition, benchmark and database of AI for social good initiatives. *nature machine intelligence, 3*, 111-115. https://doi.org/10.1038/s42256-021-00296-0

Cui, Y. (2023): *Weakly Supervised Point Cloud Semantic Segmentation with Graph Convolutional Networks.* https://www.proquest.com/openview/adfc3b6f984fcccddca6ccedfe29039f/1?pq-origsite=gscholar&cbl=18750&diss=y (Abgerufen am 03.06.2025)

Desai, S. &. Ghose, D. (2022): Active Learning for Improved Semi-Supervised Semantic Segmentation in Satellite Images. 2022 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 1485–1495. https://doi.org/10.1109/WACV51458.2022.00155.

de Vries, A. (2023): *The growing energy footprint of artificial intelligence.* Joule, 7(10), 2191-2194. https://doi.org/10.1016/j.joule.2023.09.004

DFG. (2019): *Leitlinien zur Sicherung guter wissenschaftlicher Praxis.* https://zenodo.org/records/14281892 (Abgerufen am 03.06.2025)

DFG. (2023): *Öffentliche Konsultation zum Forschungsdatengesetz*. https://www.bmbf.de/SharedDocs/Downloads/DE/gesetze/forschungsdatengesetz/stellungnahmen/df/DeutscheForschungsgemeinschaftAnlage.pdf?__blob=publicationFile&v=3 (Abgerufen am 03.06.2024)

DLR. (2023): *Öffentliche Konsultation zum Forschungsdatengesetz*. https://www.bmbf.de/SharedDocs/Downloads/DE/gesetze/forschungsdatengesetz/stellungnahmen/DeutschesZentrumfurLuft-RaumfahrteV.pdf?__blob=publicationFile&v=3 (Abgerufen am 03.06.2025)

Domingues, T., Brandão, T., & Ferreira, J. C. (2022): Machine Learning for Detection and Prediction of Crop Diseases and Pests: A Comprehensive Survey. *Agriculture, 12*(9). 1350. https://doi.org/10.3390/agriculture12091350

DZSF. (2023): *Öffentliche Konsultation zum Forschungsdatengesetz*. https://www.bmbf.de/SharedDocs/Downloads/DE/gesetze/forschungsdatengesetz/stellungnahmen/DeutschesZentrumfurSchienenverkehrsforschung.pdf?__blob=publicationFile&v=3 (Abgerufen am 03.06.2025)

Elsevier. (kein Datum): Scopus. *https://www.scopus.com/*. (Abgerufen am 03.06.2025)

Ennaji, O., Vergütz, L., & El Allali, A. (2023): Machine learning in nutrient management: A review. *Artificial Intelligence in Agriculture, 9*, 1-11. https://doi.org/10.1016/j.aiia.2023.06.001

Europäischer Rechnungshof. (2020): *Nutzung neuer Bildgebungstechnologien zur Überwachung der Gemeinsamen Agrarpolitik: Fortschritte insgesamt kontinuierlich, bei der Klima- und Umweltüberwachung jedoch langsamer.* https://www.eca.europa.eu/lists/ecadocuments/sr20_04/sr_new_technologies_in_agri-monitoring_de.pdf (Abgerufen am 03.06.2025)

Europäischer Rechnungshof. (2022): *Daten in der Gemeinsamen Agrarpolitik: Potenzial von Big Data wird für die Zwecke der Politikbewertung nicht voll ausgeschöpft.* https://www.eca.europa.eu/Lists/ECADocuments/SR22_16/SR_Big_Data_in_CAP_DE.pdf (Abgerufen am 03.06.2025)

European Commission (2022): Implementing regulation - C(2022)9562 https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=intcom:C%282022%299562 (Abgerufen am 03.06.2025)

European Commission. (2018): *Cost-Benefit analysis for FAIR research data - Cost of not having FAIR research data.* https://op.europa.eu/en/publication-detail/-/publication/d375368c-1a0a-11e9-8d04-01aa75ed71a1 (Abgerufen am 03.06.2025)

European Union. (2022): *Regulation (EU) 2022/868 of the European Parliament and of the Council of 30 May 2022 on European data governance and amending Regulation (EU) 2018/1724 (Data Governance Act)*. http://data.europa.eu/eli/reg/2022/868/oj (Abgerufen am 03.06.2025)

Ferraglioni, G. (2024): *Interview: Biodiversity space isn't ready for AI due to data gaps, expert says*. https://carbon-pulse.com/293270/ (Abgerufen am 03.06.2025)

Forzieri, G., Girardello, M., Ceccherini, G., Spinoni, J., Feyen, L., Hartmann, H., Beck, P.S.A., Camps-Valls, G., Chirici, G., Mauri, A., Cescatti, A. (2021): Emergent vulnerability to climate-driven disturbances in European forests. *Nat Commu, 12*. https://doi.org/10.1038/s41467-021-21399-7

Fraunhofer Gesellschaft. (2023): *Öffentliche Konsultation zum Forschungsdatengesetz*. https://www.bmbf.de/SharedDocs/Downloads/DE/gesetze/forschungsdatengesetz/stellungnahmen/Fraunhofer-GesellschaftzurFoerderungderangewandtenForschungeV.pdf?__blob=publicationFile&v=3 (Abgerufen am 03.06.2025)

FZI. (2023): *Öffentliche Konsultation zum Forschungsdatengesetz*. https://www.bmbf.de/SharedDocs/Downloads/DE/gesetze/forschungsdatengesetz/stellungnahmen/ForschungszentrumInformatik.pdf?__blob=publicationFile&v=3 (Abgerufen am 03.06.2025)

Ganesh, A., McCallum, A., & Strubell, E. (2019): Energy and policy considerations for deep learning in NLP. *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics.* https://doi.org/10.48550/arXiv.1906.02243

GFZ, & FID GEO. (2023): *Öffentliche Konsultation zum Forschungsdatengesetz*. https://www.bmbf.de/SharedDocs/Downloads/DE/gesetze/forschungsdatengesetz/stellungnahmen/GeoForschungsZentrumGFZ.pdf?__blob=publicationFile&v=3 (Abgerufen am 03.06.2025)

Golyadkin, M. G. (2024): Refining the ONCE Benchmark With Hyperparameter Tuning. *IEEE Access, 12*, 3805–3814. https://doi.org/10.1109/ACCESS.2023.3348750

Gotsch, M., Martin, N., Eberling, E., Shirinzadeh, S., Kuhlmann, D., Petschow, U., & Pentzien, J. (2022): *Der Beitrag von Big Data, KI und digitalen Plattformen auf dem Weg zu einer Green Economy - Einsatzbereiche und Transformationspotenziale.* Dessau-Roßlau: Umweltbundesamt. https://www.umweltbundesamt.de/sites/default/files/medien/479/publikationen/texte_85-2022_digitalisierung_als_transformationsmotor_fuer_eine_green_economy.pdf (Abgerufen am 03.06.2025)

GPAI. (2022): *Biodiversity & Artificial Intelligence, Opportunities and Recommendations.* Global Partnership on AI. https://gpai.ai/projects/responsible-ai/environment/biodiversity-and-AI-opportunities-recommendations-for-action.pdf (Abgerufen am 03.06.2025)

Halevy, A., Norvig, P., & Pereira, F. (2009). The Unreasonable Effectiveness of Data. *IEEE Computer Society*, 8-12. doi: 10.1109/MIS.2009.36

Hamid, O. (2022): From Model-Centric to Data-Centric AI: A Paradigm Shift or Rather a Complementary Approach? *2022 8th International Conference on Information Technology Trends (ITT)*, 196–199. DOI: 10.1109/ITT56123.2022.9863935

Heguerte, L. B., Bugeau, A., & Lannelongue, L. (2023). *How to estimate carbon footprint when training deep learning models? A guide and review*. https://doi.org/10.48550/ARXIV.2306.08323

Hochschule Bochum. (2023): *Öffentliche Konsultation zum Forschungsdatengesetz*. https://www.bmbf.de/SharedDocs/Downloads/DE/gesetze/forschungsdatengesetz/stellungnahmen/HochschuleBochum.pdf?__blob=publicationFile&v=3 (Abgerufen am 03.06.2025)

Höchtl, J., Parycek, P., & Schöllhammer, R. (2016): Big data in the policy cycle: Policy decision making in the digital era. *Journal of Organizational Computing and Electronic Commerce, 26*(1-2), 147-169. https://doi.org/10.1080/10919392.2015.1125187

Hughes, A. C., Orr, M. C., Ma, K., Costello, M. J., Waller, J., Provoost, P., Yang, Q., Zhu, C., Qiao, H. (2021): Sampling biases shape our view of the natural world. *Ecography, 44*, 1259-1269. https://doi.org/10.1111/ecog.05926

Huntingford, C., Jeffers, E. S., Bonsall, M. B., Christensen, H. M., Lees, T., & Yang, H. (2019): Machine learning and artificial intelligence to aid climate change research and preparedness. *Environ. Res. Lett., 14*. https://doi.org/10.1088/1748-9326/ab4e55

IEA. (2024): *Electricity 2024 Analysis and forecast to 2026.* IEA, Paris https://www.iea.org/reports/electricity-2024 (Abgerufen am 03.06.2025)

ISPRS. (2018): 2D Semantic Labeling Contest - Potsdam. *ISPRS Test Project on Urban Classification, 3D Building Reconstruction and Semantic Labeling*. https://www.isprs.org/education/benchmarks/UrbanSemLab/2d-sem-label-potsdam.aspx (Abgerufen am 08.08.2024)

Jacobs, M. (2020): *Using Machine Learning for carbon emission optimization in Transport and Logistics*. https://medium.com/appanion/using-machine-learning-for-carbon-emission-optimization-in-transport-and-logistics-5cbf5ff831a (Abgerufen am 03.06.2025)

Jakubik, J., Vossing, M., Kuhl, N., Walk, J., & Satzger, G. (2024): Data-Centric Artificial Intelligence. Bus Inf Syst Eng 66, 507–515. https://doi.org/10.1007/s12599-024-00857-8

Jean, N., Burke, M., Xie, M., Davis, W. M., Lobell, D. B., & Ermon, S. (2016): Combining satellite imagery and machine learning to predict poverty. *Science, 353*, 790-794. DOI:10.1126/science.aaf7894

Jungblut, S.-I. (2024): *Wie wird der Energiefresser Künstliche Intelligenz nachhaltiger?* https://reset.org/wie-wird-der-energiefresser-ki-nachhaltiger/ (Abgerufen am 03.06.2025)

Karnama, A., Haghighi, E. B., & Vinuesa, R. (2019): Organic data centers: A sustainable solution for computing facilities. *Results in Engineering*. https://doi.org/10.1016/j.rineng.2019.100063

Kimberlite. (2024): *Microsoft is the Leading Cloud Provider for the Oil & Gas Industry*. https://www.kimberliteresearch.com/single-post/microsoft-is-the-leading-cloud-provider-for-the-oil-gas-industry-2 (Abgerufen am 03.06.2025)

Klien, E., Kocon, K., Krämer, M., Catalli, F., Hochenegger, F., & Reitz, T. (unveröffentlicht): *Datenannotation für ML-Anwendungen im Umweltsektor - Stand der Wissenschaft und Technik.* Umweltbundesamt, Dessau-Roßau.

Klien, E., Kocon, K., Krämer, M., Hochenegger, F., Reitz, T., & Catalli, F. (unveröffentlicht): *ML-Anwendungen mit wenig Daten. Entwicklung und prototypische Umsetzung eines Vorgehensmodells zum Umgang mit wenigen annotierten Daten für die Generierung von ML-Anwendungen im Umweltbereich.* Umweltbundesamt, Dessau-Roßlau.

Klompenburg, T., Kassahun, A., & Catal, C. (2020): Crop yield prediction using machine learning: A systematic literature review. *Computer and Electronics in agriculture, 177*. Article 105709. https://doi.org/10.1016/j.compag.2020.105709

Kocon, K. K. (2022): *Comparison of CNN-based segmentation models for forest type classification.* AGILE: GIScience Series, 3, 1–6. https://doi.org/10.5194/agile-giss-3-42-2022

Kölle, M., Walter, V., Schmohl, S., & Soergel, U. (2023). LEARNING ON THE EDGE: BENCHMARKING ACTIVE LEARNING FOR THE SEMANTIC SEGMENTATION OF ALS POINT CLOUDS. *ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, *X-1/W1-2023*, 945–952. https://doi.org/10.5194/isprs-annals-X-1-W1-2023-945-2023

Köller, C. (2022): *Intelligentes Verkehrsmanagement reduziert CO2-Emissionen*. https://www.springerprofessional.de/verkehrsmanagement/emissionen/intelligentes-verkehrsmanagement-reduziert-co2-emissionen/23257802 (Abgerufen am 03.06.2025)

Krämer, M., Würz, H. M., & Altenhofen, C. (2021). Executing cyclic scientific workflows in the cloud. Journal of Cloud Computing, 10(25), 1–26. https://doi.org/10.1186/s13677-021-00229-7

Kumar, S., Tiwari, S., Tewatia, H., Jain, A., & Popli, J. (2023): DynaTrack: Machine Learning Aided Variable Speed Limit System. *Procedia Computer Science, 218*, 83-92. https://doi.org/10.1016/j.procs.2022.12.404

Kumari, S. K. (2021): *An ensemble approach for classification and prediction of diabetes mellitus using soft voting classifier.* International Journal of Cognitive Computing in Engineering, 2, 40-46. https://doi.org/10.1016/j.ijcce.2021.01.001

Kumat, S., Datta, S., Singh, V., Singh, S., & Sharma, R. (2024): Opportunities and Challenges in Data-Centric AI. *IEEE Access*, 33173–33189. doi: 10.1109/ACCESS.2024.3369417

La, H., & La, K. (2023): Comparing model-centric and data-centric approaches to determine the efficiency of data-centric AI. *Journal of Emerging Investigators*. DOI:10.59720/22-112

LeCun Y., C. C. (2024): THE MNIST DATABASE. *https://www.kaggle.com/datasets/hojjatk/mnist-dataset* (Abgerufen am 03.06.2025)

Lee, D.-H. (2013): Pseudo-label: The simple and efficient semi-supervised learning method for deep neural networks. *Workshop on challenges in representation learning, ICML. Vol. 3. No. 2*. https://citeseerx.ist.psu.edu/document?repid=rep1&type=pdf&doi=798d9840d2439a0e5d47bcf5d164aa46d5e7dc26 (Abgerufen am 03.06.2025)

Li, P., Yang, J., Islam, M., & Ren, S. (2023): Making AI Less "Thirsty": Uncovering and Addressing the Secret Water Footprint of AI Models. *arXiv*. https://doi.org/10.48550/arXiv.2304.03271

Li, Z. Z. (2022): *Multi-stage pseudo-label iteration framework for semi-supervised land-cover mapping.* In IGARSS 2022-2022 IEEE International Geoscience and Remote Sensing Symposium (pp. 4607-4610). IEEE. DOI:10.1109/IGARSS46834.2022.9884345

Liu, X., Lu, D., Zhang, A., Liu, Q., & Jiang, G. (2022): Data-Driven Machine Learning in Environmental Pollution: Gains and Problems. *Environmental Science & Technology, 56*(4). https://doi.org/10.1021/acs.est.1c06157

Lizarraga, C., & Solon , O. (2023): *Thirsty Data Centers Are Making Hot Sommers Even Scarier*. https://www.bloomberg.com/news/articles/2023-07-26/extreme-heat-drought-drive-opposition-to-ai-data-centers?embedded-checkout=true (Abgerufen am 03.06.2025)

Luccioni, A. S., & Hernandez-Garcia, A. (2023): Counting Carbon: A Survey of Factors Influencing the Emissions of Machine Learning. *arXiv preprint arXiv:2302.08476*. https://doi.org/10.48550/arXiv.2302.08476

Luers, A. K., Koomey, J., Masanet, E., Gaffney, O., Creutzig, F., Ferres, J. L., Horvitz, E. (2024): *Will AI accelerate or delay the race to net-zero emissions?* Nature, 628(8009), 718-720. doi: https://doi.org/10.1038/d41586-024-01137-x

Maganathan, T., Senthilkumar, S., & Balakrishnan, V. (2020): Machine Learning and Data Analytics for Environmental Science: A Review, Prospects and Challenges. *IOP Conf. Ser.: Mater. Sci. Eng, 955*. https://doi.org/10.1088/1757-899X/955/1/012107

Malakar, N. (2022): Environmental Intelligence using Machine Learning Applications. *Journal of Nepal Physical Society, 8*(2). https://doi.org/10.3126/jnphyssoc.v8i2.50148

Manca, F., Benedetti-Cecchi, L., Bradshaw, C., Cabeza, M., gustafsson, C., Norkko, A. M., Roslin, T. V., Thomas, D. N., White, L., Strona, G.(2024): Projected loss of brown macroalgae and seagrasses with global environmental change. *Nat Commun, 15*(5344). https://doi.org/10.1038/s41467-024-48273-6

Manske, D., & Schmiedt, J. (2023): Geo-locations and System Data of Renewable Energy Installations in Germany. *https://zenodo.org/records/8188601*. (Zenodo, Hrsg.) doi:10.5281/zenodo.6920930 (Abgerufen am 03.06.2025)

Marx, P. (2024): *The Backlash to the AI-Fueled Data Center Boom*. https://www.youtube.com/watch?v=k5xbjhYpklo&list=PLAR_6-tD7IZV6D2ud_q905-riF4QAp3Jy&index=18 (Abgerufen am 03.06.2025)

Microsoft. (2018): *Microsoft demonstrates the power of AI and Cloud to Oil and Gas players, at ADIPEC 2018*. https://news.microsoft.com/en-xm/2018/11/12/microsoft-demonstrates-the-power-of-ai-and-cloud-to-oil-and-gas-players-at-adipec-2018/ (Abgerufen am 03.06.2025)

Mirpulatov, I., Illarionove, S., Shadrin, D., Burnaev, E. (2023): Pseudo-Labeling Approach for Land Cover Classification Through Remote Sensing Observations With Noisy Labels. *IEEE Access,*(11), 82570–82583. https://doi.org/10.1109/ACCESS.2023.3300967

Mizzaro, S., Nazzi, E., Vassena, L. (2009): *Collaborative annotation for context-aware retrieval.* In Proceedings of the WSDM'09 Workshop on Exploiting Semantic Annotations in Information Retrieval (pp. 42-45). https://doi.org/10.1145/1506250.1506260

Mo, L., Zohner, C., Reich, P., Liang, J., de Miguel, S., Nabuurs, G., Renner, S. S., van den Hoogen, J., Araza, A., Herold, M., Mirzagholi, L., Ma, H., Averill, C., Phillips, O.L., Gamarra, J. G. P., Hordijk, I., Routh, D., Abegg, M., Yao, Y. C. A., Alberti, G. … Crowther, T. (2023): Integrated global assessment of the natural forest carbon potential. *Nature*, 92-101. https://doi.org/10.1038/s41586-023-06723-z

Muhammad, Z. (2022): *Google's Data Centers Used 4 Billion Gallons of Water in 2021*. https://www.digitalinformationworld.com/2022/12/googles-data-centers-used-4-billion.html (Abgerufen am 03.06.2025)

MVKU. (2023): *Abschlussbericht: Luftschadstoffprognose in Berlin mit Machine-Learning-Methoden.* Berlin: Senatsverwaltung für Mobilität, Verkehr, Klimaschutz und Umwelt. https://www.google.com/url?sa=t&source=web&rct=j&opi=89978449&url=https://www.berlin.de/weniger-dicke-luft/_assets/projekte-und-massnahmen/luftschadstoffprognose/luftschadstoffprognose-in-berlin.pdf&ved=2ahUKEwjb7N2FkdeNAxXjZ_EDHfLMH9wQFnoECBMQAQ&usg=AOvVaw0ngcItA3n_2S9jhL9kSDFJ (Abgerufen am 03.05.2025)

Natarajan, S., Shanmurthy, P., Arockiam, D., Balusamy, B., Selvarajan, S. (2024):Optimized machine learning model for air quality index prediction in major cities in India. *Sci Rep, 14*. https://doi.org/10.1038/s41598-024-54807-1

NFDI. (2023): *Stellungnahme zur öffentlichen Konsultation zum Forschungsdatengesetz*. https://www.nfdi.de/wp-content/uploads/2023/05/NFDI-Stellungnahme-zum-Forschungsdatengesetz.pdf (Abgerufen am 03.06.2025)

Nieradzik L, Sieburg-Rockel, J., Helmling, S., Keuper, J., Weibel, T., Olbrich, A., Stephanie, H. (2024): Automating Wood Species Detection and Classification in Microscopic Images of Fibrous Materials with Deep Learning. Microscopy and microanalysis: the official journal of Microscopy Society of America, Microbeam Analysis Society, Microscopical Society of Canada, 30(3), 508–520. https://doi.org/10.1093/mam/ozae038

Nilashi, M., Keng Boon, O., Tan, G., Lin, B., & Abumalloh, R. (2023): Critical Data Challenges in Measuring the Performance of Sustainable Development Goals: Solutions and the Role of Big-Data Analytics. *Harvard Data Science Review, 5*(3). https://doi.org/10.1162/99608f92.545db2cf

OpenGeodata.NRW. (2024): Baumartenklassifikation in NRW als TIFF . *Fernerkundung im Wald NRW*. https://www.opengeodata.nrw.de/produkte/umwelt_klima/wald_forst/fernerkundung/ (Abgerufen am 03.06.2025)

Peiyuan, L., Tianfang, X., Shiqi, W., & Zhi-Hua, W. (2022): Multi-objective optimization of urban environmental system design using machine learning. *Computers, Environment and Urban Systems, 94*. DOI: 10.1016/j.compenvurbsys.2022.101796

Peranandam , C. (27. September 2018): *Your Guide to AI and Machine Learning at re:Invent 2018*. https://aws.amazon.com/blogs/machine-learning/your-guide-to-ai-and-machine-learning-at-reinvent-2018/ (Abgerufen am 03.06.2025)

Perera, A. S., Witharanma, C., Manos, E., Liljedahl, A.K. (2024): Hyperparameter Optimization for Large-Scale Remote Sensing Image Analysis Tasks: A Case Study Based on Permafrost Landform Detection Using Deep Learning. *IEEE Access, 12*, 43062–43077. https://doi.org/10.1109/ACCESS.2024.3379142

Perez, L. & Wang, J. (2017): *The effectiveness of data augmentation in image classification using deep learning.* Convolutional Neural Networks Vis. Recognit, 11(2017), 1-8. https://doi.org/10.48550/arXiv.1712.04621

Perrigo, B. (2022): *Inside Facebook's African Sweatshop.* https://time.com/6147458/facebook-africa-content-moderation-employee-treatment/ (Abgerufen am 11.11.2024)

Perrigo, B. (2023): *Exclusive: OpenAI Used Kenyan Workers on Less Than $2 Per Hour to Make ChatGPT Less Toxic.* https://time.com/6247678/openai-chatgpt-kenya-workers/ (Abgerufen am 11.11.2024)

Petrović, B., Bumbálek, R., Zoubek, T., Kuneš, R., Smutný, L., & Bartoš, P. 2024, March 1). Application of precision agriculture technologies in Central Europe-review. Journal of Agriculture and Food Research. Elsevier B.V. https://doi.org/10.1016/j.jafr.2024.101048

Plattform Lernende Systeme. (2020): *KI-Geschäfts- modelle für Reisen und Transport: Mehr Wirtschaftlich- keit und Nachhaltigkeit in der Mobilität der Zukunft (Whitepaper). https://www.acatech.de/publikation/ki-geschaeftsmodelle-fuer-reisen-und-transport-mehr-wirtschaftlichkeit-und-nachhaltigkeit-in-der-mobilitaet-der-zukunft/ (*Abgerufen am *03.06.2025)*

Rajadell, O. & García-Sevilla, P. (2013): Training Selection with Label Propagation for Semi-supervised Land Classification and Segmentation of Satellite Images. In P. Latorre Carmona, J. S. Sánchez, & A. L. N. Fred (Eds.), Pattern Recognition - Applications and Methods (Vol. 204, pp. 181–192). Springer Berlin Heidelberg. https://doi.org/10.1007/978-3-642-36530-0_15

Ramírez-Delgado, J., Di Marco, M., Watson, J., & et al. (2022): Matrix condition mediates the effects of habitat fragmentation on species extinction risk. *Nat Commun*(13), https://doi.org/10.1038/s41467-022-28270-3

Rammer, W., & Seidel, R. (2019): Harnessing deep learning in ecology: An example predicting bark beetle outbreaks. *Frontiers in Plant Science*, doi.org/10.3389/FPLS.2019.01327

Reitz, T., Escriu, J., Minghini, M., & Kotsev, A. (In review): Interoperability - Technical Report. JRC

Russo, S., Besmer, M. D., Blumensaat, F., Bouffard, D., Disch, A., Hammes, F., Hess, A., Lürig, M., Matthews, B., Minaudo, C., Morgenroth, E., Tran-Khac, V., Villez, K. (2021): The value of human data annotation for machine learning based anomaly detection in environmental systems. *Water Research, 206*. https://doi.org/10.1016/j.watres.2021.117695.

Sahana, M., Areendran, G., & Sajjad, H. (2022): Assessment of suitable habitat of mangrove species for prioritizing restoration in coastal ecosystem of Sundarban Biosphere Reserve, India. *Sci Rep* , https://doi.org/10.1038/s41598-022-24953-5

Savills. (2024): *European Data Centres.* https://www.savills.com/research_articles/255800/362604-0 (Abgerufen am 03.06.2025)

Schiller, C., Költzow, J., Schwarz, S., Schiefer, F., & Fassnacht, F. E. (2024): Forest disturbance detection in Central Europe using transformers and Sentinel-2 time series. *Remote Sensing of Environment, 315*. https://doi.org/10.1016/j.rse.2024.114475

Schmitt, M., Ahmadi, S., & Hansch, R. (2021): There is No Data Like More Data - Current Status of Machine Learning Datasets in Remote Sensing. *IEEE International Geoscience and Remote Sensing Symposium IGARSS*, http://dx.doi.org/10.1109/IGARSS47720.2021.9555129

Seidl, R., Klonner, G., Rammer, W., Essl, F., Moreno, A., Neumann, M., Dullinger, S. (2018): Invasive alien pests threaten the carbon stored in Europe's forests. *Nat Commun, 9*(1626). https://doi.org/10.1038/s41467-018-04096-w

Sharma, A., Jain, A., Gupta, P., & Chowdary, V. (2020): Machine Learning Applications for Precision Agriculture: A Comprehensive Review. *IEEE Access*, DOI:10.1109/ACCESS.2020.3048415

Singh, P. (2023): Systematic review of data-centric approaches in artificial intelligence and machine learning. *Data Science and Management, 6*(3), 144–157. https://doi.org/10.1016/j.dsm.2023.06.001

slalom. (2024): *Mit modernen Daten und KI den Waldschutz beschleunigen*. https://www.slalom.com/de/de/customer-stories/forest-stewardship-council?utm_cta=website-pricing-page-resources-pricing-guide (Abgerufen am 03.06.2025)

Solawetz, J. (4. September 2020): Train, Validation, Test Split for Machine Learning. *Roboflow Blog*. https://blog.roboflow.com/train-test-split/ (Abgerufen am 03.06.2025)

Statistika. (2024): *Annual water consumption of Taiwan Semiconductor Manufacturing Company (TSMC) from 2016 to 2023, by user*. https://www.statista.com/statistics/1313040/tsmc-water-consumption-by-user/#:~:text=In%202023%2C%20the%20Taiwanese%20semiconductor,to%20the%20stability%20of%20production. (Abgerufen am 03.06.2025)

Sujatha, M., & Jaidhar, C. D. (2024): Machine learning-based approaches to enhance the soil fertility—A review. *Expert Systems with Applications, 240*. https://doi.org/10.1016/j.eswa.2023.122557

Sylvain, J., Drolet, G., Thiffault, E., & Anctil, F. (2024): High-resolution mapping of tree species and associated uncertainty by combining aerial remote sensing data and convolutional neural networks ensemble. *International Journal of Applied Earth Observation and Geoinformation*, https://doi.org/10.1016/j.jag.2024.103960

Taweesan, A., Koottatep, T., Kanabkaew, T., Polprasert, C. (2024): Application of machine learning in sanitation management prediction: Approaches for achieving sustainable development goals. *Environmental and Sustainability Indicators, 22*. https://doi.org/10.1016/j.indic.2024.100374

The World Bank and ITU. (2024): *Measuring the Emissions & Energy Footprint of the ICT Sector: Implications for Climate Action.* Washington, D.C. and Geneva.: The World Bank and ITU. https://documents.worldbank.org/en/publication/documents-reports/documentdetail/099121223165540890/p17859702a98880540a4b70d57876048abb (Abgerufen am 03.06.2025)

Thompson, T. (2023): How AI can help to save endagered species. *Nature, 623*, 232-233. doi: https://doi.org/10.1038/d41586-023-03328-4

Thünen Institut. (2023): *Öffentliche Konsultation zum Forschungsdatengesetz*. https://www.bmbf.de/SharedDocs/Downloads/DE/gesetze/forschungsdatengesetz/stellungnahmen/thuenen-institut/Thuenen-Institut.pdf?__blob=publicationFile&v=3 (Abgerufen am 08.08.2024)

Tuia, D., Kellenberger, B., Beery, S., Costelloe, B. R., zuffi, S., Risse, B., Mathis, A., Mathis, M. W., van Langevelde, F., Burghardt, T., Kays, R., Klinck, H., Wikelski, M., Couzin, I. D., van Horn, G., Crofoot, M. C., Stewart, C. V., Berger-Wolf, T. (2022): Perspectives in machine learning for wildlife conservation. *Nat Commun, 13*(792). https://doi.org/10.1038/s41467-022-27980-y

UFZ. (2023): *Öffentliche Konsultation zum Forschungsdatengesetz - Fragebogen*. https://www.bmbf.de/SharedDocs/Downloads/DE/gesetze/forschungsdatengesetz/stellungnahmen/HelmholtzZentrumfurUmweltforschungGmbH-UFZ.pdf?__blob=publicationFile&v=3 (Abgerufen am 03.06.2025)

Umutoni, L., & Samadi, V. (2024): Application of machine learning approaches in supporting irrigation decision making: A review. *Agricultural Water Management, 294*. https://doi.org/10.1016/j.agwat.2024.108710

Vinuesa, R., Azizpour, H., Leite, I., Balaam, M., Dignum, V., Domisch, S., Felländer, A., Langhans, S.D., Tegmark, M., Nerini, F. F. (2020): The role of artificial intelligence in achieving the Sustainable Development Goals. *Nature Communications, 11*(233), 1-10. https://doi.org/10.1038/s41467-019-14108-y

Walter, A., Finger, R., Huber, R., & Buchmann, N. (2017): Smart farming is key to developing sustainable agriculture. *Agricultural Sciences, 114*(24), 6148-6150. https://doi.org/10.1073/pnas.1707462114

Wang, H., Cimen, E., Singh, N., & Buckler, E. (2020): Deep learning for plant genomics and crop improvement. *Current Opinion in Plant Biology, 54*, 34-41. https://doi.org/10.1016/j.pbi.2019.12.010

Wurm, D., Zielinski, O., Lübben, N., Ramesohl, S., Jansen, M., Heinz, R., Morinmg, A., Just, V., Spreiter, L., Damm, P., Hiontermann, R., Hinterholzer, S., Grothey, T., Bach, B., Dowie, U., Schuklze, M. (2022): *Wege in eine ökologische Machine Economy: ein systemischer Blick auf die Umweltwirkungen digitaler Technologien.* Institut für Zukunftsstudien und Technologiebewertung. https://nbn-resolving.org/urn:nbn:de:bsz:wup4-opus-85390 (Abgerufen am 03.06.2025)

Xu, P., Li, G., Zheng, Y., Fung, J., Chen, A., Zeng, Z., Shen, H., Hu, M. Mao, J., Zheng, Y., Cui, X., Guo, Z., Chen, Y., Feng, L., He, S., Zhang, X., Lau, A. K. H., Houlton, B. (2024): Fertilizer management for global ammonia emission reduction. *Nature, 626*, 792–798. https://doi.org/10.1038/s41586-024-07020-z

Xu, Y., Liu, X., Cao, X., Huang, C., Liu, E., Qian, S., Liu, X., Wu, Y., Dong, F., Qiu, C., Qui, J., Hua, K., Su, W., Wu, J., Xu, H., Han, Y., Fu, C., Yin, Z., Liu, M., … Zhang, J. (2021): Artificial intelligence: A powerful paradigm for scientific research. *The Innovation, 2*(100179). https://doi.org/10.1016/j.xinn.2021.100179

Yang, S., Xiao, W., Zhang, M., Guo, S., Zhao, J., Shen, F. (2022): *Image Data Augmentation for Deep Learning: A Survey.* arXiv preprint arXiv:2204.08610. https://doi.org/10.48550/arXiv.2204.08610

Yao, Z., Lum, Y., Johnston, A., Mejia-Mendoza, L. M., Zhou, X., Wen, Y., . Aspuru-Guzik, A., Sargent, E. H., She, Z. W. (2023): Machine learning for a sustainable energy future. *Nat Rev Mater, 8*, 202-215. https://doi.org/10.1038/s41578-022-00490-5

Yue, P., & Shangguan, B. (2023): *OGC Training Data Markup Language for Artificial Intelligence (TrainingDML-AI) Part 1: Conceptual Model Standard*. (OGC) https://docs.ogc.org/is/23-008r3/23-008r3.html (Abgerufen am 03.06.2025)

Zha, D., Bhat, Z. P., Lai, K., Yang, F., Jiang, Z., Zhong, S., Hu. X. (2023): *Data-centric artificial intelligence: A survey.* arXiv preprint arXiv:2303.10158. https://doi.org/10.48550/arXiv.2303.10158

Zhong, S., Zhang, K., Bagheri, M., Burken, J. G., Gu, A., Li, B., Ma, X., Marrone, B. L., Ren, Z. J., Schrier, J., Shi, W., Tan, H., Wang, T., Wang, X., Wong, B. M., Xiao, X., Yu, X., Zhu, J. Zhang, H. (2021): Machine Learning: New Ideas and Tools in Environmental Science and Engineering. *Environmental Science and Technology, 395*. https://doi.org/10.1021/acs.est.1c01339