

Wald5Dplus

Sarah Hauser^{*,1,2}, Andreas Schmitt^{1,2}

¹Munich University of Applied Sciences (MUAS), Geoinformatics Department, Karlstraße 6, D-80333 Munich, Germany.

²Institute for Applications of Machine Learning and Intelligent Systems (IAMLIS) at MUAS, Lothstraße 34, D-80335 Munich, Germany. sarah.hauser@hm.edu; andreas.schmitt@hm.edu

***Corresponding and presenting author**

Die Verfügbarkeit hochqualitativer Referenzdaten ist entscheidend für Anwendungen von Künstlicher Intelligenz (KI) und maschinellem Lernen (ML) in der Forstwirtschaft, wo eine detaillierte und genaue Überwachung von strukturellen und ökologischen Dynamiken unabdingbar ist. Während KI und ML vielversprechend bei der Ableitung von Waldattributen aus Satellitendaten sind, hängt ihr Erfolg maßgeblich von der Verfügbarkeit robuster, multi-sensorischer und multi-temporaler Datensätze ab [1]. Die Forstwirtschaft steht somit vor einer doppelten Herausforderung: Einerseits eröffnet die Verfügbarkeit immer größerer Datenmengen neue Möglichkeiten, andererseits fehlen geeignete Trainingsdatensätze, um diese Informationen effizient in KI-gestützte Anwendungen zu überführen. Bestehende Ansätze scheitern häufig daran, heterogene Daten wie Radar- und optische Informationen konsistent zu fusionieren, ohne dabei die Datenqualität zu beeinträchtigen. Ein weiteres Problem ist der eingeschränkte Zugang zu Benchmark-Datensätzen, die für methodische Weiterentwicklungen und Validierungen essenziell sind. Solche Datensätze ermöglichen es, neue Ansätze zu testen, zu vergleichen und bestehende Algorithmen weiterzuentwickeln. Das Fehlen einer solchen offenen Datenressource behindert die Integration von KI-Technologien in die Forstwirtschaft erheblich.

Wald5Dplus begegnet diesen Herausforderungen durch die Bereitstellung eines Open-Access-Benchmark-Datensatz [2], der speziell entwickelt wurde, um KI in der Waldökosystemüberwachung voranzutreiben. Gefördert vom Deutschen Zentrum für Luft- und Raumfahrt (DLR) mit Mitteln des Bundesministeriums für Wirtschaft und Klimaschutz erfüllt Wald5Dplus die hohen Datenanforderungen von KI-Anwendungen in den Umweltwissenschaften und bietet eine zuverlässige Grundlage für skalierbare, hochauflösende Waldüberwachung in verschiedenen ökologischen Regionen [1]. Dabei werden sogenannte Analysis Ready Data (ARD)-Cubes verwendet, die bis zu 512 Datenebenen pro Jahr in einem GIS-kompatiblen Format organisieren. Diese Cubes minimieren den Verarbeitungsaufwand und erleichtern den Zugang zu spektralen, strukturellen und zeitlichen Informationen für KI-gestützte Anwendungen. Dies ermöglicht eine schnelle Implementierung von ML-Modellen für die Überwachung von Phänologie, Kronenstruktur und Biomasse. Sentinel-1- (Radar) und Sentinel-2-Zeitreihen (optisch) mit Labels aus luftgestützten LiDAR- und Multispektraldaten kombiniert. Die Multi-Sensor-Fusion fußt auf hyperkomplexen Basen (HCBs) [3], einer fortschrittlichen Technik, die spektrale und strukturelle Informationen auf Pixelebene integriert. HCBs bewahren die Einzigartigkeit der Beiträge jedes Sensors, sind vollständig reversibel und gewährleisten somit die Datenqualität für wiederholte Analysen. Das Ergebnis ist ein konsistenter Datensatz, der präzise Zeitreihenanalysen von Waldattributen ermöglicht.

Wald5Dplus ist über Zenodo [2] öffentlich zugänglich und bietet:

- Fusionierte Sentinel-1- und Sentinel-2-Daten lückenlos über zwei volle Jahre.
- Hochauflösende aggregierte Referenzdaten aus LiDAR- und Multispectral-Scans.
- Vortrainierte ML-Modelle für die flächendeckende Ableitung von Waldattributen.

Die technische Innovation von Wald5Dplus liegt in der Multi-Sensor-Fusion, die Radar- und optische Daten in einem gemeinsamen radiometrischen Rahmen kombiniert. Durch die Transformation in den Kennaugh-Raum werden spektrale und zeitliche Informationen verlustfrei integriert. Diese Struktur ermöglicht es, operationelle Arbeitsabläufe zu vereinfachen und große Datenmengen direkt zugänglich zu machen. Die ARD-Cubes minimieren den Aufwand für Datenvorbereitung und erleichtern so die Integration in forstliche Prozesse. Sentinel-1-SAR- und Sentinel-2-optische Daten werden durch eine spezialisierte Multi-Sensor-Fusion unter Verwendung hyperkomplexer Basen (HCBs) [3] kombiniert. Dieser Prozess etabliert einen gemeinsamen radiometrischen Rahmen, der eine Fusion auf Pixelebene ermöglicht und Radar- und multispektrale optische Daten mit minimalen Verzerrungen kombiniert. Durch die Transformation der Sentinel-2-Reflexionen in „Kennaugh-ähnliche“ Elemente zur Angleichung an die Kennaugh-Elemente von Sentinel-1 und einer linearen Transformation in den Kennaugh-Raum entsteht eine HCB-Struktur, die sowohl eine mittlere jährliche Reflexionsebene als auch 63 zeitliche Ebenen umfasst, welche saisonale Veränderungen dokumentieren. Diese Daten werden als UInt8 gespeichert, was eine effiziente, verlustfreie Speicherung ermöglicht und den Zugriff über GIS- und Bildverarbeitungsplattformen erleichtert [3]. Die Fusion erzeugt einen fortschrittlichen ARD-Cube mit bis zu 512 Schichten pro Jahr, der aus 64-mal-8-fusionierten Kennaugh-ähnlichen Elementen besteht [1].

Der Wald5Dplus-Benchmark-Datensatz deckt dabei drei ökologisch und strukturell unterschiedliche Gebiete in Süddeutschland ab:

- Nationalpark Bayerischer Wald: Umfasst eine Mischung aus alten und regenerierenden Beständen in höheren Lagen und erfasst Variabilitäten in Baumhöhe und Biomassendichte.
- Steigerwald: Von Laubbaumarten in niedrigeren Lagen dominiert, erfasst er saisonale Variationen in Blattabdeckung und Artenzusammensetzung, ideal für Phänologiestudien.
- Kranzberger Forst: Als kontrolliertes Versuchsgelände ermöglicht er die Analyse von Bewirtschaftungseffekten unter konstanten Bedingungen.

Zusammen bilden diese Standorte einen umfassenden Datensatz, der ein breites Spektrum an Waldtypen umfasst und die Entwicklung robuster KI-Modelle unterstützt, die an verschiedene ökologische Kontexte anpassbar sind. Die Referenzdaten von Wald5Dplus, die aus hochauflösenden LiDAR- und multispektralen Quellen stammen [4, 5, 6], liefern umfassende, präzise Metriken für jedes UTM-33/32N-10m-Pixel [1].

Um die Verbindung zwischen den Earth Observation (EO)-Daten und den Referenzdaten zu überprüfen, wurden verschiedene ML-Modelle getestet. Besonders das Random Forest (RF)-Ensemble-Modell erwies sich aufgrund seiner Genauigkeit und Robustheit als optimal. RF-Modelle sind dafür bekannt, Nichtlinearitäten in den Daten effektiv zu handhaben, robuste Vorhersagen zu liefern und Overfitting zu vermeiden. Diese Eigenschaften machen sie besonders geeignet für die Modellierung komplexer Waldattribute wie Baumhöhe, Kronenvolumen und Artenklassifikation. Das RF-Ensemble kombiniert Vorhersagen mehrerer RF-Modelle, die auf unterschiedlichen Teilmengen der Daten trainiert wurden (Data Tiling). Dieser Ansatz sorgt für eine höhere Stabilität und Genauigkeit der Modelle. Die gestapelten Vorhersagen nutzen die Stärken einzelner Modelle und minimieren gleichzeitig deren Schwächen. So entsteht ein robustes Gesamtmodell, das verlässliche Vorhersagen über verschiedene Waldregionen hinweg ermöglicht.

Wald5Dplus richtet sich an ein breites Spektrum von Stakeholdern. Forschende profitieren von einer umfassenden Datengrundlage für die Entwicklung und Validierung von KI-Modellen. Forstpraktiker:innen können KI-basierte Lösungen in operationelle Arbeitsabläufe integrieren und so Effizienzsteigerungen erzielen. Politische Entscheidungsträger:innen erhalten eine solide Basis für datengetriebene Entscheidungen in der Forstwirtschaft und im Klimaschutz, während Softwareentwickler:innen neue Anwendungen zur Waldüberwachung entwickeln können.

Für die Teilnehmer:innen der Konferenz bietet Wald5Dplus einen erheblichen Mehrwert: Sie erhalten praktische Einblicke in Multi-Sensor-Datenfusion und KI-Anwendungen, begleitet von Anleitungen zur effizienten Nutzung großer, heterogener Datensätze. Wald5Dplus inspiriert zu datengetriebenen Ansätzen für eine nachhaltige Waldüberwachung.

Von der Konferenz erwartet das Wald5Dplus-Team wertvolle Rückmeldungen zur Übertragbarkeit auf andere Waldökosysteme sowie Anregungen zur Erweiterung der Datensatzstruktur für spezifische Anwendungen. Die Diskussion über datengetriebene Innovationen soll zudem neue Perspektiven eröffnen, wie KI-Technologien zur nachhaltigen Bewirtschaftung von Waldökosystemen beitragen können.

References:

- [1] Hauser, S., Ruhhammer, M., Schmitt, A., & Krzystek, P. An Open Benchmark Dataset for Forest Characterization from Sentinel-1 and -2 Time Series. *Remote Sens.* 2024, 16, 488.
- [2] Hauser, S., Schmitt, A., Krzystek, P., & Ruhhammer, M. (2024). Wald5Dplus (1.0.0) [Data set]. Zenodo. <https://doi.org/10.5281/zenodo.10848838>
- [3] Schmitt, A.; Wendleder, A.; Kleynmans, R.; Hell, M.; Roth, A.; Hinz, S. Multi-Source and Multi-Temporal Image Fusion on Hypercomplex Bases. *Remote Sens.* 2020, 12, 943.
- [4] Latifi, H., et al. (2021). "A laboratory for conceiving Essential Biodiversity Variables (EBVs)—The 'Data pool initiative for the Bohemian Forest Ecosystem.'" *Methods in Ecology and Evolution*, 12(11), 2073-2083. doi: 10.1111/2041-210X.13695
- [5] Dersch, S., Schöttl, A., Krzystek, P., & Heurich, M. (2023). "Towards complete tree crown delineation by instance segmentation with Mask R-CNN and DETR using UAV-based multispectral imagery and lidar data." *ISPRS Open Journal of Photogrammetry and Remote Sensing*, 8, 100037. doi: 10.1016/j.ophoto.2023.100037.
- [6] Heidrich, L., et al. (2020). "Heterogeneity-diversity relationships differ between and within trophic levels in temperate forests." *Nature Ecology & Evolution* 4:1431-1431.