Vogelwarte 63, 2025: 191–215

Hohe Sensitivität automatisierter Detektion von Seevögeln auf See auf digitalen Luftbildaufnahmen

Tim Schmoll, Guruprasad Hegde, Monika Dorsch & Georg Nehls

Schmoll T, Hegde G, Dorsch M & Nehls G 2025: High sensitivity of automated detection of seabirds at sea on digital aerial survey footage. Vogelwarte 63: 191–215.

The reliable estimation of the abundance of seabirds at sea is an important basis for conservation and environmental impact assessments. The use of artificial intelligence (AI) for automated processing of digital aerial images promises a faster, more cost-effective, and better reproducible analysis compared to manual processing. It is unclear, however, whether an AI-supported workflow can achieve comparable quality to that of specifically trained observers, a prerequisite for establishing it as a standard in maritime environmental planning.

Here we describe the architecture, training, and testing of the object detection model HiDeFIND 1.0, a convolutional neural network with more than 400 layers and more than 86 million parameters. HiDeFIND was trained on more than 138,000 annotated still images of birds and marine mammals from digital aerial video footage and then confronted with images of an independent test image set featuring more than 111,000 verified biological objects. Objects of both sets had previously been detected and identified at the species or species group level by trained observers.

Although the test image set with more than 120 species/species groups had nearly twice as many taxa as the training image set, HiDeFIND found 96.5% of the manually detected objects overall. Accounting for the test data set's hierarchical dependency structure in a mixed effects model analysis, it achieved a high mean sensitivity („recall" in machine learning) of >99%. This included the detection of many key species of maritime environmental planning such as Red-throated Diver *Gavia stellata*, Common Guillemot *Uria aalge* and Black-legged Kittiwake *Rissa tridactyla*, as well as Harbour Porpoise *Phocoena phocoena* among marine mammals (all with >99% mean sensitivity). The achieved sensitivity was independent of the seasons and largely independent of detection-relevant environmental variation.

The overall high sensitivity came with a high rate of false positive detections, especially under glare. As a consequence, manual removal of false positive detections is required. Currently this reduces the efficiency of an AI-supported workflow and thus time savings, albeit not the high sensitivity of HiDeFIND. Further development of HiDeFIND will specifically focus on reducing the rate of false positive detections without meaningfully sacrificing sensitivity.

For the analysis of offshore digital aerial survey footage in environmental planning, monitoring and research, the use of HiDeFIND represents a forward-looking alternative to exclusively manual object detection. HiDeFIND operates here as part of an integrated "human-in-the-loop" work process, in which automated initial detection is flanked by manual supervision.

✉ TS: BioConsult SH GmbH & Co. KG, AG Fernerkundung, Schobüller Str. 36, 25813 Husum.
E-Mail: t.schmoll@bioconsult-sh.de. ORCID: 0000-0003-3234-7335
MD: BioConsult SH GmbH & Co. KG, AG Fernerkundung, Schobüller Str. 36, 25813 Husum.
E-Mail: m.dorsch@bioconsult-sh.de
GH: BioConsult SH GmbH & Co. KG, AG Fernerkundung, Schobüller Str. 36, 25813 Husum.
E-Mail: g.hegde@bioconsult-sh.de
GN: BioConsult SH GmbH & Co. KG, Schobüller Str. 36, 25813 Husum.
E-Mail: g.nehls@bioconsult-sh.de. ORCID: 0009-0000-6424-1989

1 Einleitung

Die verlässliche Erfassung von Seevögeln auf See ist von großer Bedeutung für die maritime Raum- und Umweltplanung sowie für die Bewertung der Bestandsdynamik von Seevögeln im Rahmen von nationalen und internationalen Monitoring-Programmen. Seit der Jahrtausendwende sind flugzeug-gestützte Erfassungen und seit mehr als zehn Jahren digitale Transekt-Erfassungsflüge (Buckland et al. 2012) in vielen Ländern der Standard für Seevogel- und Meeressäugererfassungen. In Deutschland ist dies im "Standard Untersuchung der Auswirkungen von Offshore-Windenergieanlagen auf die Meeresumwelt StUK 4" festgelegt (BSH 2013). Eine bewährte Methode ist die digitale HiDef-Videoerfassung (Weiß et al. 2016, Žydelis et al. 2019), die international seit über zehn Jahren bei mehr als 3000 Erfassungsflügen zum Einsatz gekommen ist (Dorsch et al. 2024).

Künstliche Intelligenz (KI) in Form von maschinellem Lernen und insbesondere von „Deep Learning" hat in den meisten Bereichen der Biologie jüngst sehr stark an Bedeutung gewonnen (Übersicht in Greener et al. 2022). Dies gilt auch für die Ökologie im Allgemeinen (Übersichten in Borowiec et al. 2022, Ditria et al. 2022) und das Monitoring von Tierbeständen im

Besonderen (Übersichten in Tuia et al. 2022, Nakagawa et al. 2023, Xu et al. 2024). In Bezug auf Letzteres hat sich automatisierte Bilderkennung sowohl in terrestrischen (z. B. Tabak et al. 2019) als auch in aquatischen Umwelten (z. B. Li et al. 2023) als sehr nützlich erwiesen. Maschinelles Lernen wurde zum Beispiel erfolgreich eingesetzt, um Quallenblüten zu erfassen (Mcilwaine & Rivas Casado 2021), und um die Abundanz von Landsäugetieren (Torney et al. 2019, Lenzi et al. 2023) oder die Abundanz von (brütenden) Wasservögeln und Seevögeln auf Drohnenaufnahmen abzuschätzen (Dujon et al. 2021, Kellenberger et al. 2021, Marchowski 2021). Darüber hinaus unterstützte maschinelles Lernen z. B. auch die Schätzung von Wal- und Delfinbeständen auf flugzeuggestützten Luftbildern (Boulent et al. 2023), auf Satellitenbildern (Borowicz et al. 2019, Guirado et al. 2019) oder mittels akustischer Erfassung unter Wasser (Frainer et al. 2023).

Während maschinelles Lernen in Kombination mit Drohnen oder Satelliten bereits erfolgreich in ganz unterschiedlichen Kontexten eingesetzt wurde, gibt es bislang kaum veröffentlichte Arbeiten zur Anwendung von maschinellem Lernen auf flugzeuggestützte Bestandserfassungen von Seevögeln auf See. Eine Studie von Kuru et al. (2023) beschreibt eine Methode zur automatisierten Erkennung von Basstölpeln *Morus bassanus* auf Luftbildaufnahmen. Dieser Ansatz basierte allerdings nicht auf den jüngsten Techniken des „Deep Learning". Darüber hinaus beschreiben Ke et al. (2024) ein „*Deep Learning*"-Detektionsmodell im Rahmen eines vollautomatisierten Arbeitsprozesses, der in der Zukunft Bestandsschätzungen von Seevögeln auf See während des Fluges und nahezu in Echtzeit erlauben soll.

Angesichts der Erfolgsbilanz maschinellen Lernens zur Bestandsschätzung von Tieren ist eine Anwendung auf die Detektion von Vögeln und Meeressäugern auf HiDef-Videomaterial vielversprechend. Mindestens vier Vorteile im Vergleich zur manuellen Objektdetektion sind denkbar. Erstens ist zu erwarten, dass die automatisierte Objekterkennung schneller als der manuelle Prozess ist. Ergebnisse könnten dann zeitnaher z. B. in Planungsentscheidungen einfließen. Zweitens könnte eine daraus resultierende höhere Kosteneffektivität eine zeitlich wie räumlich engmaschigere Erfassung der Meeresumwelt durch häufigere und/oder ausgedehntere Erhebungen erlauben. Drittens ist zu erwarten, dass die Ergebnisse automatisierter Objekterkennung besser reproduzierbar sind als die Ergebnisse manueller Objekterkennung, was z. B. die Transparenz daraus abgeleiteter Planungsentscheidungen erhöhen würde. Eine automatisierte Objekterkennung bietet viertens auch das Potenzial, die Qualität von Bestandsschätzungen von Seevögeln auf See und Meeressäugern zu steigern. Das Zusammenspiel von einerseits ausgeprägter biologischer Variation (z. B. Arten, Geschlechtsdimorphismen, Altersklassen, Verhaltensweisen, Körperhaltungen) und andrerseits ausgeprägter Umweltvariation (z. B. Lichtverhältnisse, Seegang, Luft- und Wassertrübung) macht die zuverlässige automatisierte Objekterkennung in digitalen Luftaufnahmen allerdings zu einer Herausforderung (Miao et al. 2023, Xu et al. 2024). Ob ein KI-unterstützter Ansatz geschulten Beobachter:innen ebenbürtig oder womöglich überlegen ist, erfordert daher eine Evaluation der Leistung jedes einzelnen Objekterkennungs-Modells. Solange Aufsichtsbehörden keinen Nachweis für das Versprechen hoher Ergebnisqualität verlangen und keine allgemeinen Qualitätsstandards für den Einsatz von KI bei Bestandsschätzungen mittels Fernerkundung etabliert sind (Converse et al. 2024), besteht sonst das Risiko, Planungsentscheidungen oder Bewertungen von Bestandsdynamiken auf unzuverlässiger Datengrundlage zu treffen.

In diesem Beitrag beschreiben wir zunächst das von uns entwickelte Objekterkennungsmodell HiDeFIND, ein auf „Deep Learning" basierendes künstliches neuronales Netzwerk zur Detektion von Seevögeln und Meeressäugern auf digitalem HiDef-Videomaterial. Wir dokumentieren Architektur, Training und Testen des Modells sowie die umfangreichen HiDef-Bildsätze, die zum Trainieren und Testen von HiDeFIND verwendet wurden. Anschließend analysieren wir die Modellleistung im Detail und untersuchen ihre Abhängigkeit z. B. von Artzugehörigkeit, Jahreszeit und detektionsrelevanten Umweltbedingungen. Wir zeigen, dass HiDeFIND Vögel auf See sowie Meeressäuger ganzjährig und unter einer Vielzahl detektionsrelevanter Umweltbedingungen fast genauso gut detektiert wie geschulte Beobachter:innen und diskutieren die Eignung des Systems für den Einsatz in der maritimen Raum- und Umweltplanung und im Seevogel-Monitoring.

2 Material und Methoden

Die Entwicklung des Objekterkennungsmodells HiDeFIND (Version 1.0) basierte auf digitalen Videoaufnahmen, die mit der HiDef-Methode gewonnen wurden. Objekterkennungen des etablierten, manuellen Arbeitsprozesses stellten dabei den Maßstab dar, mit dem wir die Leistung von HiDeFIND verglichen haben. Demnach repräsentierten diese von menschlichen Beobachter:innen erzielten Objekterkennungen die sogenannte „ground truth" und nicht etwa die tatsächliche Anzahl potenziell detektierbarer Objekte, die nur im Feldvergleich ermittelt werden könnte. Wir beschreiben daher zunächst knapp den HiDef-Standard-Arbeitsprozess (für Details siehe Weiß et al. 2016, Žydelis et al. 2019). Anschließend dokumentieren wir Architektur, Training und Testen von HiDeFIND selbst.

2.1 Der HiDef-Standard-Arbeitsprozess

2.1.1 Datenerhebung

Für die digitalen Erfassungsflüge waren zweimotorige Hochdecker-Propellerflugzeuge (z. B. Vulcanair P 68) in einer Flughöhe von ca. 500 m im Einsatz. Die Flugzeuge waren mit Sensoren-Systemen ausgestattet, die aus vier hochauflösenden digitalen Videokameras bestanden (Abbildung 1a). An der Meeresoberfläche erreichten diese eine mittlere Bodenauflösung („ground sampling distance") von ca. 2 cm bei einer

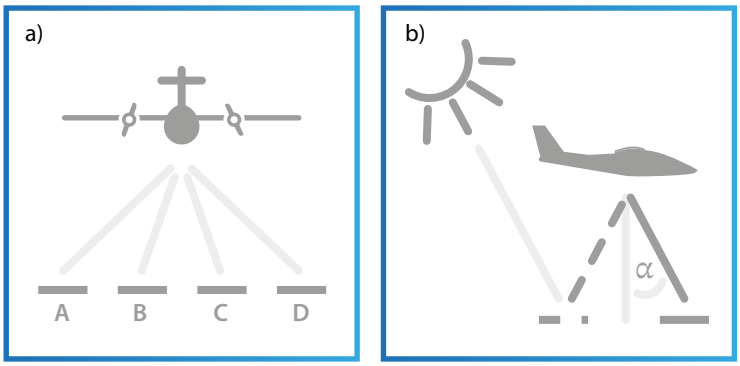


Abb. 1: Schematische Darstellung des HiDef-Sensoren-Systems in a) Frontalansicht und b) Seitenansicht. – *Schematic depiction of the HiDef sensor system in a) frontal view and b) lateral view.*

Bildrate von sieben Bildern pro Sekunde. Aufgrund einer etwas größeren Distanz zwischen Objektiv und Meeresoberfläche hatten die inneren im Vergleich zu den äußeren Kameras eine geringfügig höhere Bodenauflösung. Das Sensoren-System war dabei nicht starr lotrecht zur Meeresoberfläche ausgerichtet, sondern war je nach Kurs und Sonnenstand um 30° in oder gegen die Flugrichtung schwenkbar (Abbildung 1b). Dies diente der Vermeidung störender Sonnenreflexion auf der Meeresoberfläche, die Erkennung und Bestimmung von Zielobjekten erschweren kann. Die äußeren Sensoren A und D deckten einen Streifen von jeweils 143 m ab, die inneren Sensoren B und C einen Streifen von jeweils 129 m Breite. Um Doppelzählungen zu vermeiden, wurde zwischen den Streifen ein Abstand von ca. 20 m gehalten. Daraus ergab sich eine effektive Schwadbreite von 544 m, verteilt auf eine Gesamtstreifenbreite von ca. 604 m.

Die Flugzeuge flogen mit einer mittleren Geschwindigkeit von etwa 220 km/h (120 Knoten). Ein GPS-Gerät zeichnete die Position im Sekundentakt auf, was die Georeferenzierung der erfassten Objekte ermöglichte. Die gesammelten Daten wurden zur späteren Analyse auf mobilen Festplatten gespeichert.

2.1.2 Datenauswertung

Die Videodateien wurden mit der Bilderfassungs- und Bildverwaltungssoftware StreamPix (NorPix, Montreal, Kanada) weiterverarbeitet und die von einer Kamera aufgezeichneten Dateien zur leichteren Bearbeitung längsmittig geteilt. Die vier Kameras beherbergen somit insgesamt acht virtuelle Kameras, die acht Videosequenzen pro Transekt liefern. Für die Datenauswertung wurden in einem ersten Schritt Standbilder von geschulten Mitarbeitenden begutachtet, und alle erkannten Objekte wurden in StreamPix digital punktmarkiert sowie für die spätere Objektbestimmung vorklassifiziert (z. B. Verdacht auf Vogel, Säugetier, Tier oder menschengemachtes Objekt). Videosequenzen oder deren Teile, die aufgrund zu starker Sonnenreflexion oder Wolken nicht verlässlich ausgewertet werden konnten, wurden markiert und in nachfolgende Auswertungsschritte nicht einbezogen (in zukünftigen Arbeitsprozessen wird zunächst eine Analyse der Erfassungsbedingungen erfolgen, bevor gesichert auswertbare Teile des Videomaterials der KI-unterstützen Objekterkennung zugewiesen werden). Um eine gleichbleibend hohe Qualität zu gewährleisten, wurde ein zufällig ausgewählter Anteil von 20 % des Videomaterials unabhängig von zwei Beobachter:innen bearbeitet (ohne Kenntnis der Ergebnisse der anderen). Erfasst wurden überdies die detektionsrelevanten Umweltbedingungen Sonnenreflexion, Seegang sowie Luft- und Wassertrübung (siehe Anhang 1). Für Seegang, Luft- und Wassertrübung galt die Annahme, dass diese im Moment der Aufnahme identisch waren für den gesamten von den vier Kameras abgedeckten Erfassungsstreifen und somit auch für die durch Zweiteilung resultierenden acht Videosequenzen eines jeden Transekts. Diese Umweltbedingungen wurden daher einzelbildscharf auf nur einer einzigen Videosequenz pro Transekt bewertet. Die Sonnenreflexion kann jedoch in Abhängigkeit von Sonnenstand und Kurs auch zwischen Videosequenzen eines Transekts variieren. Sie wurde daher einzelbildscharf für jede der acht Videosequenzen eines jeden Transekts bewertet.

In einem zweiten Schritt wurden markierte Objekte von erfahrenen, ornithologisch und in der Meeressäuger-Bestimmung versierten Mitarbeitenden auf der genauesten möglichen taxonomischen Ebene bestimmt, in der Regel auf Artebene. War dies etwa aufgrund der Verwechslungsgefahr sehr ähnlicher Arten (z. B. Flussseeschwalbe *Sterna hirundo* und Küstenseeschwalbe *S. paradisaea*) nicht möglich, wurden diese Objekte Gruppen ähnlicher Arten zugeordnet (z. B. Artengruppe Fluss-/Küstenseeschwalbe). Zusätzlich wurden, falls möglich, Geschlecht und Alter sowie das Verhalten (z. B. schwimmend oder fliegend), Assoziation (z. B. mit Individuen eigener oder anderer Arten) und gegebenenfalls die Flugrichtung erfasst. Desweiteren wurden zur Qualitätskontrolle 20 % der markierten Objekte unabhängig von einer zweiten Person bestimmt (ohne Kenntnis der Ergebnisse der anderen). Eventuelle Diskrepanzen zwischen dem ersten und dem zweiten Identifizierungsprozess wurden von einer dritten Person erneut geprüft und bei Bedarf korrigiert. Nur wenn eine Übereinstimmung von mindestens 90 % zwischen den beiden Identifizierungsprozessen bestand, wurden die Daten zur weiteren Analyse freigegeben. Falls die Übereinstimmung weniger als 90 % betrug, wurden systematische Fehler wie z. B. gehäufte Probleme innerhalb bestimmter Artengruppen diskutiert und alle Objekte auf dem betroffenen Filmmaterial erneut bestimmt.

2.2 Entwicklung eines neuronalen Netzwerks für die Objekterkennung: HiDeFIND

Die Entwicklung eines künstlichen neuronalen Netzwerkmodells für die HiDef-Objekterkennung umfasste die folgenden fünf hauptsächlichen Schritte:

1. **Basismodell: Auswahl eines Basismodells**, dessen Architektur unseren spezifischen Anforderungen am besten entsprach.
2. **Datenannotation:** Auswahl und Annotation eines möglichst umfangreichen und mannigfaltigen Trainingsbildsatzes.
3. **Training:** Im Trainingsprozess haben wir das Modell verschiedenen Hyperparameter-Einstellungen ausgesetzt und es dadurch in einem rekursiven Prozess mit den annotierten Trainingsbildern trainiert und optimiert.

4. **Validierung:** Anschließend prüften wir das Modell an einem kleinen Validierungsbildsatz und bewerteten über den Trainingsprozess hinweg wiederholt seine aktuelle Leistung.
5. **Testen:** Zum Schluss führten wir einen umfangreichen Leistungstest durch, bei dem ein unabhängiger, großer, heterogener Testbildsatz Verwendung fand.

2.2.1 Auswahl und Spezifikation eines Modells

Nach Abwägung unterschiedlicher grundsätzlich geeigneter Modellarchitekturen wählten wir als Basismodell einen einstufigen Objekterkennungsalgorithmus aus der **Y**ou **O**nly **L**ook **O**nce-Familie (YOLO), die weitverbreitet Anwendung auf Probleme des Computersehens findet. Ergänzt um benutzerdefinierte Netzwerk-Schichten, besteht das resultierende künstliche neuronale Netzwerk HiDeFIND (ein „Convolutional Neural Network, CNN") aus mehr als 400 Schichten und mehr als 86 Millionen Parametern. Die Eingabeschicht verarbeitet Digitalfotos als Vektor der RGB-Werte aller ihrer Pixel und die Ausgabeschicht liefert von HiDeFIND generierte Bounding Boxen, die mit vorgegebener Wahrscheinlichkeit Zielobjekte beinhalten. HiDeFIND nimmt dabei eine Klassifikation nur entlang einer einzigen Kategorie vor (semantisiert): „Ist biologisches Objekt? Ja/Nein". Das Hauptziel der Entwicklung bestand darin, eine Erfassungsgenauigkeit zu erzielen, die mindestens so gut wie die des etablierten manuellen Prozesses ist, um höchste Standards bei der Ergebnisqualität zu gewährleisten. Daher wurde HiDeFIND so konfiguriert, dass es im Zweifelsfall visuelle Muster als Objekt von Interesse markiert, anstatt sie zu verwerfen, also grundsätzlich der Sensitivität („recall" im maschinellen Lernen) zu Lasten der Präzision Priorität einräumt (s. 2.2.5.3).

2.2.2 Objektverfolgung über Standbilder

HiDef-Videomaterial produziert zeitlich orientierte, räumlich überlappende Sequenzen von Einzelbildern. Jedes Objekt auf HiDef-Videomaterial erscheint daher in der Regel auf mehr als einem Standbild (je nach Position im Probestreifen und der Flughöhe bei Vögeln auf bis zu acht Standbildern). Dementsprechend wird von HiDeFIND jedes Objekt in der Regel auch mehr als einmal angesprochen (dies gilt insbesondere für richtig positive Detektionen). Um unerwünschte Mehrfach-Detektionen desselben biologischen Objekts kontrollieren und automatisiert herausfiltern zu können, haben wir daher einen Hilfsalgorithmus zur Objektverfolgung entwickelt, der auf dem Kuhn-Munkres-Algorithmus basiert („Hungarian matching algorithm", Kuhn 1955). Dies stellte sicher, dass die Detektion(en) eines nachverfolgten biologischen Objekts nur auf einem einzigen Standbild als richtig positiv gewertet wurden (auf dem Standbild, auf dem menschliche Beobachter:innen das Objekt markiert hatten). Weitere Detektionen desselben nachverfolgten biologischen Objekts auf anderen Standbildern wurden dagegen als falsch positive Detektionen gewertet (siehe 2.2.5.2).

2.2.3 Herkunft des Bildmaterials

Für die Erstellung des Trainings- und des Testbildsatzes haben wir auf das gemeinsame Archiv digitalen HiDef-Videomaterials der Firmen BioConsult SH GmbH & Co KG (Husum, Deutschland, https://www.bioconsult-sh.de), HiDef Aerial Surveying Ltd. (Workington, Vereinigtes Königreich, https://www.hidefsurveying.co.uk) und Biotope (Mèze, Frankreich, https://www.biotope.fr) zurückgegriffen. Eine detaillierte Charakterisierung der genutzten Bildsätze erfolgt unten.

2.2.4 Training

2.2.4.1 Annotation des Trainingsbildsatzes

Entscheidend für das erfolgreiche Training eines Objekterkennungsmodells ist die Verfügbarkeit eines großen und hinreichend diversen Trainingsbildsatzes, der geeignete Bilder samt Annotation der enthaltenen Trainingsobjekte in Form von Bounding Boxen umfasst. Im genutzten Archivmaterial hatten Beobachter:innen die von ihnen entdeckten Objekte verortet, indem sie eine digitale Punktmarkierung im Objektzentrum angebracht hatten. Um die vorhandenen digitalen Punktmarkierungen nachnutzen zu können, haben wir ein maßgeschneidertes Digitalwerkzeug entwickelt, mit dessen Hilfe Benutzer:innen vorhandene Punktmarkierungen auf HiDef-Material importieren und als Grundlage für eine manuelle Annotation mit Bounding Boxen verwenden konnten.

2.2.4.2 Trainingsprozess

Zum Training verarbeitete das Modell die annotierten Trainingsbilder während Dutzender sogenannter Epochen. In jeder dieser Epochen wurde das Modell mit dem gesamten Trainingsmaterial konfrontiert und seine Vorhersagen durch modell-generierte Bounding Boxen repräsentiert. Deren Mittelpunktskoordinaten sowie ihre Höhe und Breite wurden genutzt, um im Gradientenverfahren iterativ Diskrepanzen zwischen der tatsächlichen Verortung der Objekte (hinterlegt als manuell gesetzte Bounding Boxen im Trainingsmaterial) und der vom Modell vorhergesagten Verortung (als Bounding Boxen der Modellausgabe) durch eine systematische Anpassung der Modell-Parameter zu minimieren. Je Epoche haben wir den Lernfortschritt des Modells mit einem Validierungsbildsatz (knapp 35 000 Standbilder) nachverfolgt, der keine Überlappung mit dem Trainingsbildsatz (gut 138 000 Standbilder, siehe 2.2.4.3) oder dem Testbildsatz (gut 111 000 Standbilder, siehe 2.2.5.1) aufwies.

2.2.4.3 Trainingsbildsatz

Wir trainierten das Modell mit 138 681 annotierten Objekten aus 21 Erfassungsflügen zu unterschiedlichen Jahreszeiten in vier Projektgebieten in zwei Meeresgebieten (Nordsee, Ostsee) (Abbildung 2a, Anhang 2). Im Trainingsbildsatz können Individuen mit mehr als einem Bild repräsentiert sein, da auf HiDef-Videomaterial dieselben Objekte in der Regel auf mehreren aufeinanderfolgenden Standbildern erfasst wurden (siehe 2.2.2), oft in unterschiedlicher Exposition oder Flügelhaltung, was dem Training förderlich war (im Testbildsatz entsprach die Anzahl der manuell detektierten Objekte der Anzahl manuell detektierter Individuen). Der Trainingsbildsatz enthielt 66 Arten/Artengruppen, darunter 57 Vogeltaxa (siehe Anhang 2 für Details). Trottellummen *Uria aalge* und Tordalken *Alca torda* machten zusammen mit der Artengruppe Trottellumme/Tordalk 31 % der Gesamtzahl der Objekte aus. Schweinswale *Phocoena phocoena* repräsentierten 85 % der Meeressäugerobjekte. Abbildung 3a zeigt die Häufigkeit der 50 häufigsten Taxa im Trainingsbildsatz.

Abb. 2: Erfassungsflüge nach Jahreszeit für a) Trainingsbildsatz und b) Testbildsatz (Jahreszeiten meteorologisch). Kompartimente visualisieren Anteil unterschiedlicher Projektgebiete. – *Aerial surveys by season for a) training image set and b) test image set (seasons meteorological). Stacks visualize contributions of different project sites.*

2.2.5 Evaluierung der Modellleistung

Wir bewerteten die Leistung von HiDeFIND anhand eines Testbildsatzes, der keine Überlappung mit dem Trainingsbildsatz oder dem Validierungsbildsatz aufwies.

2.2.5.1 Testbildsatz

Für den Testbildsatz wählten wir Erfassungsflüge aus, die unterschiedliche Artenzusammensetzungen und ein Spektrum an detektionsrelevanter Umweltvariation umfassten, so dass wir das Modell unter einer Vielzahl realistischer Bedingungen testen konnten. Der Testbildsatz umfasste 111.666 verifizierte biologische Objekte aus 13 Erfassungsflügen zu unterschiedlichen Jahreszeiten in sechs Projektgebieten in drei unterschiedlichen Meeresgebieten (Nordsee, Ostsee, Englischer Kanal) (Abbildung 2b, Anhang 2). Der Testbildsatz enthielt 124 Arten/Artengruppen, darunter 109 Vogeltaxa (siehe Anhang 2 für Details). Eiderenten *Somateria mollissima* und Trauerenten *Melanitta nigra* waren am häufigsten und machten 23 % beziehungsweise 22 % der Gesamtzahl aller Objekte aus, Schweinswale 58 % der Meeressäugetiere. Abbildung 3b zeigt die Häufigkeit der 50 häufigsten Taxa im Testbildsatz. Die Objekte waren zuvor bei manuellen Standardanalysen von HiDef-Videomaterial entdeckt und auf der Ebene der Art oder Artengruppe bestimmt worden (siehe 2.1.2). Die zugehörigen Standardanalysen waren vor Entwurf des Testdesigns abgeschlossen, die beteiligten Beobachter:innen konnten daher nicht wissen, dass ihre Leistung den Maßstab für die Leistungsbewertung von HiDeFIND darstellen würde.

2.2.5.2 Leistungsbewertung

Die Bewertung der Modellleistung bei Klassifikationsaufgaben im Allgemeinen und bei der Objektdetektion im Besonderen basiert zumeist auf einer Wahrheitsmatrix („confusion matrix"), in der tatsächliche sowie vom Modell vorhergesagte Ereignisse einander gegenüber gestellt werden (Sokolova & Lapalme 2009). Die Wahrheitsmatrix für die HiDeFIND-Leistungsbewertung kann wie in Tabelle 1 spezifiziert werden. Neben Vögeln und Meeressäugern haben wir dabei einige wenige weitere Vertreter mariner Megafauna berücksichtigt, die regelmäßig bei HiDef-Flugerfassungen registriert werden (z. B. Roter Thun *Thunnus thynnus* oder Mondfisch *Mola mola*).

Die Evaluierung eines Objektdetektionsmodells erfordert normalerweise aus dem Feldvergleich abgeleitete Annotationen in Form von Bounding Boxen, die dann mit den vom Modell vorgesagten Bounding Boxen abgeglichen werden, um die sogenannte „Intersection over Union (IoU)" zu be-

Tab. 1: Wahrheitsmatrix, die mögliche Ergebnisse der HiDeFIND-Leistungsbewertung abbildet. – *Confusion matrix mapping potential outcomes of the HiDeFIND performance evaluation.*

		Maßstab (geschultes Personal) – *Benchmark (trained staff)*	
		Ist Vogel/Säuger *– Is bird/mammal*	Ist nicht Vogel/Säuger *– Is not bird/mammal*
Modellvorhersage ***– Model prediction***	Ist Vogel/Säuger *– Is bird/mammal*	Richtig positiv *– True positive*	Falsch positiv *– False positive*
	Ist nicht Vogel/Säuger *– Is not bird/mammal*	Falsch negativ *– False negative*	Nicht definiert *– Not defined*

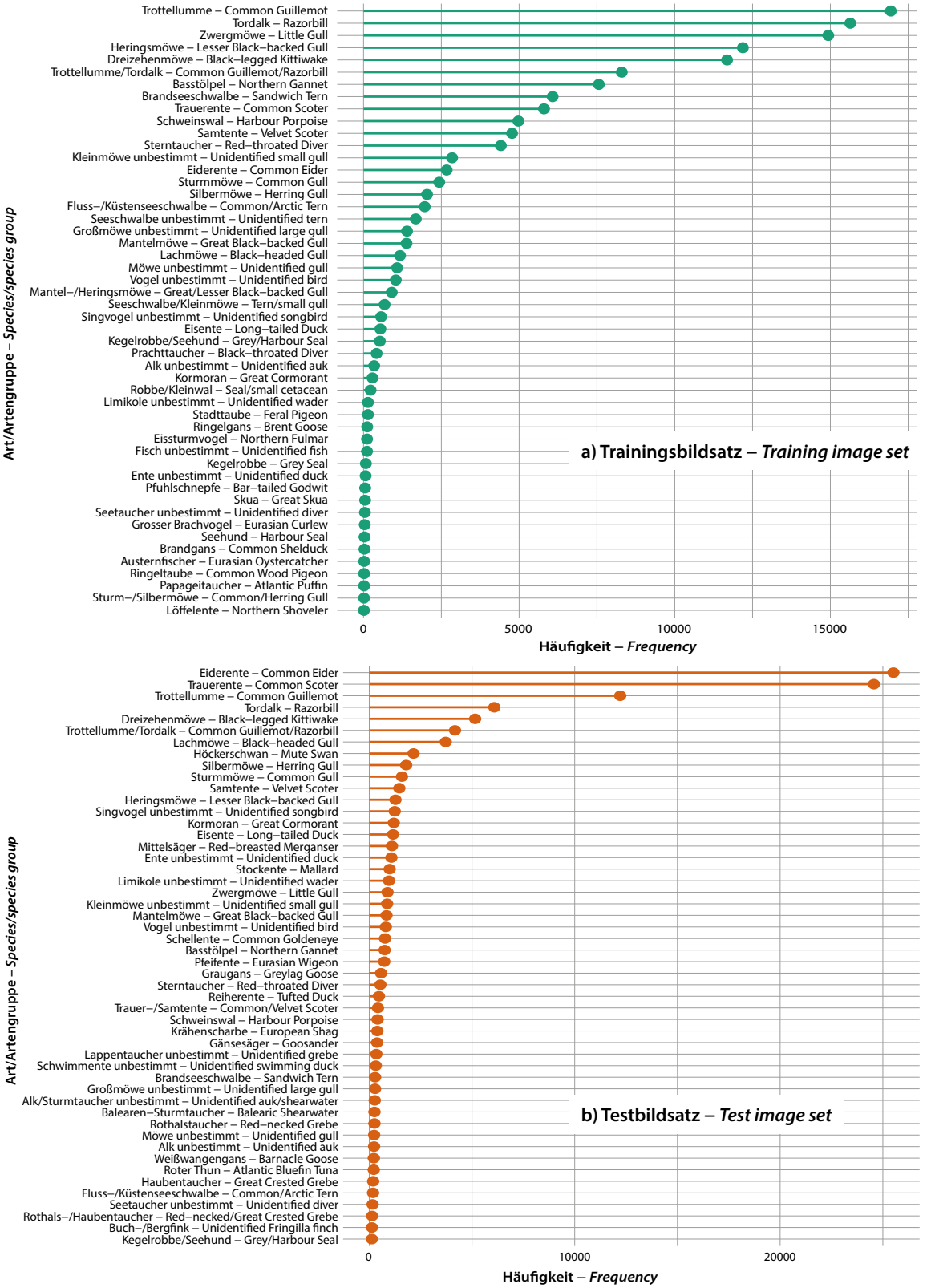


Abb. 3: Häufigkeit der jeweils 50 häufigsten Taxa im a) Trainingsbildsatz und b) Testbildsatz. – *Frequency of the 50 most common taxa in a) the training image set and b) the test image set.*

rechnen. In unserem Testdesign stellten jedoch von Beobachter:innen auf dem Videomaterial gemachte Objekterkennungen den Maßstab dar, mit dem wir die Leistung von HiDeFIND verglichen haben, und nicht die tatsächliche Anzahl der im Feldvergleich potenziell detektierbaren Objekte (siehe oben). Die von Beobachter:innen erzielten Detektionen waren durch Punktmarkierungen verortet (siehe 2.2.4.1). Wir haben daher Modellvorhersagen dann als richtig positiv bewertet, wenn vom Modell vorhergesagte Bounding Boxen die manuell gesetzten Punktmarkierungen einschlossen. Wir bewerteten Modellvorhersagen als falsch negativ, wenn manuelle Punktmarkierungen nicht von modell-generierten Bounding Boxen umfasst wurden. Alle weiteren Modellvorhersagen werteten wir als falsch positive Detektionen. Aus technischen Gründen schlossen falsch positive Detektionen daher auch Detektionen von nachverfolgten biologischen Objekten ein, die nicht auf dem Standbild erfolgten, auf dem die Beobachter:innen das Objekt punktmarkiert hatten (siehe auch 2.2.2). Richtig negative Modellvorhersagen waren in unserem Testdesign nicht definiert.

2.2.5.3 Leistungskennzahlen

Wir haben die folgenden Kennzahlen verwendet, die als Standard zur Bewertung der Modellleistung für Klassifizierungsaufgaben verwendet werden:

$$\text{Präzision} = \frac{\text{Richtig Positive}}{\text{Richtig Positive + Falsch Positive}}$$

$$\text{Sensitivität} = \frac{\text{Richtig Positive}}{\text{Richtig Positive + Falsch Negative}}$$

Es wird empfohlen, die Bewertung von KI-Modellen in stratifizierter Form zu berichten (Burnell et al. 2023). Um zu analysieren, wie stark die HiDeFIND-Leistung je nach Art/Artengruppe variiert hat, berichten wir daher Sensitivitäten nicht nur global (alle Arten/Artengruppen umfassend), sondern auch getrennt nach Arten/Artengruppen, sofern deren Häufigkeit im Testbildsatz dies sinnvoll erscheinen ließ. Da in unserem Testdesign falsch positive Detektionen nicht Arten/Artengruppen zugeordnet werden können, unterblieb dies für die Präzision. Darüber hinaus analysierten wir die Abhängigkeit der Leistungskennzahlen von der Position der Sensoren, der Jahreszeit sowie potenziell detektionsrelevanten Umweltbedingungen (siehe Anhang 1 sowie 2.3.3 und 2.3.4).

2.3 Statistische Analyse

2.3.1 Hintergrund

Im Gegensatz zu Standard-Testdesigns im Computersehen konnten wir für den HiDeFIND-Leistungstest keinen etablierten Satz verifizierter Bilder mit relevanten Objekten nutzen (solche Bildsätze sind für Trainings-, Validierungs- und Testzwecke für viele Standard-Anwendungen in der Objektklassifikation im Internet verfügbar). Für den Testbildsatz wurde vielmehr eine Stichprobe aus abgeschlossenen Flugerfassungen gezogen (siehe 2.2.5.1). Die resultierenden Datensätze unterlagen überdies einer hierarchischen Abhängigkeitsstruktur (siehe 2.3.2 und Anhang 3). Zusätzlich zur detaillierten Beschreibung der Modellleistung verwendeten wir daher statistische Ansätze, um erwartungstreue Konfidenzintervalle um die Punktschätzer von Leistungsmetriken berechnen und so den Stichprobenfehler berücksichtigen zu können. In einem weiteren Schritt untersuchten wir Ursachen für Variation in der Modellleistung, um so Bedingungen zu identifizieren, unter denen HiDeFIND noch keine optimale Leistung gezeigt hatte. Die Ergebnisse dieser exploratorischen Analysen sollen zum Beispiel die weitere Optimierung der Modellarchitektur oder die Komposition zukünftiger Trainingsbildsätze leiten.

2.3.2 Hierarchische Abhängigkeitsstruktur von HiDef-Daten

Die Detektionswahrscheinlichkeiten individueller Objekte auf HiDef-Videomaterial sind nicht unabhängig voneinander, sondern unterliegen einer hierarchischen Abhängigkeitsstruktur (für Details siehe Anhang 3). Dies gilt grundsätzlich sowohl für manuelle als auch maschinelle Objektdetektion. Um Pseudoreplikation (Hurlbert 1984) zu kontrollieren und damit Typ-1-Fehler-Inflation zu vermeiden (Forstmeier et al. 2017), muss diese Abhängigkeitsstruktur in statistischen Modellen der Modellleistung berücksichtigt werden. Dies wird durch Anpassung von statistischen Modellen erreicht, die simultan sowohl feste Effekte als auch Zufallseffekte schätzen (Modelle mit gemischten Effekten, „mixed effects models"). Eine geschachtelte Struktur relevanter Zufallseffekte bildet dabei die maßgeblichen Hierarchieebenen als Gruppierungsfaktoren im Modell ab (Gelman & Hill 2006).

2.3.3 Sensitivität

Um Konfidenzintervalle für globale und artspezifische Sensitivitäten zu schätzen, verwendeten wir verallgemeinerte lineare gemischte Modelle („Generalized Linear Mixed Models", GLMMs) mit binomialer Fehlerstruktur und Logit-Linkfunktion sowie Objekt/Individuum als statistischer Einheit. Wir passten den Gesamtmittelwert als einzigen festen Effekt an (Nullmodell) und berücksichtigten zusätzlich die Identität von Projektgebieten, Transekten und Videosequenzen als geschachtelte Zufallseffekte (siehe auch Anhang 3). Wir berechneten zugehörige 95%-Konfidenzintervalle, indem wir den Standardfehler der Schätzer auf der Linkskala mit 1,96 multiplizierten und stellen Detektionswahrscheinlichkeiten der besseren Verständlichkeit halber in Prozent inklusive retransformierter 95%-Konfidenzintervalle dar (letztere sind daher asymmetrisch). Bei der Berechnung von artspezifischen Sensitivitäten traten auch für die häufigeren Arten im Testdatensatz regelmäßig Konvergenzprobleme auf, weil die Modelle aufgrund der kleineren Stichprobenumfänge sehr kleine Zufallseffekte nicht von null abgrenzen konnten (die betreffenden Varianzen waren effektiv null). In solchen Fällen haben wir artspezifisch die Zufallseffekt-Struktur vereinfacht, bis die Modelle konvergierten, wobei Zufallseffekte schrittweise absteigend entfernt wurden. Um Effekte detektionsrelevanter Umweltbedingungen auf die Sensitivität zu prüfen, nutzten wir GLMMs mit binomialer Fehlerstruktur und Logit-Linkfunktion. Umweltbedingungen wurden in je separaten Modellen als feste Effekte modelliert (siehe auch Anhang 1, Wassertrübung nur für Meeressäuger). Zusätzlich berücksichtigten wir die Identität von Projektgebieten, Transekten und Videosequenzen als geschachtelte Zufallseffekte (siehe auch Anhang 3). Wir prüften die statistische Signifikanz von festen Effekten, indem wir mit Hilfe eines Likelihood-Quotienten-Tests (R-Funktion „anova") ein gegebenes Modell mit dem Nullmodell verglichen.

2.3.4 Präzision und Anzahl falsch positiver Detektionen

In unserem Testdesign können auf jedem der insgesamt über zwei Millionen Standbilder des Testbildsatzes falsch positive Detektionen auftreten, nicht nur auf den gut 51 000 Bildern, auf denen menschliche Beobachter:innen zuvor ein relevantes Objekt markiert hatten. Die Präzision in unserem Testdesign ist daher nicht mit anderen Leistungstests im Computersehen vergleichbar. Andere Studien zeigen in der Regel eine deutlich höhere Präzision, weil sie im Rahmen des Tests eine um Größenordnungen geringere Anzahl von ausgewiesenen Nicht-Objekt-Testbildern anbieten (in der Regel in ähnlicher Anzahl wie Testbilder, die ein relevantes Objekt enthalten). Wir berichten daher der Vollständigkeit halber nur die Gesamt-Präzision und analysieren stattdessen im Detail die Anzahl falsch positiver Detektionen pro Standbild in Abhängigkeit potenziell detektionsrelevanter Umweltbedingungen. Dies erlaubt wertvolle Rückschlüsse auf Bedingungen, unter denen HiDeFIND noch keine optimale Präzision liefern konnte. Wir modellierten die Anzahl falsch positiver Detektionen pro Standbild mit linearen gemischten Modellen („linear mixed effects models", LME) mit normaler Fehlerverteilung nach Logarithmus-Transformation (log10). Die maximal zugelassene Anzahl an Detektionen pro Standbild (Summe richtig positiver und falsch positiver Detektionen) und damit auch die maximal mögliche Anzahl falsch positiver Detektionen war auf 1000 begrenzt (lediglich 49 der über zwei Millionen Standbilder wiesen >1000 falsch positive Detektionen auf). Umweltbedingungen wurden in je separaten Modellen als feste Effekte modelliert (siehe Anhang 1, Wassertrübung nur für Meeressäuger) und wir berücksichtigten die Identität von Projektgebieten, Transekten und Videosequenzen als geschachtelte Zufallseffekte (siehe auch Anhang 3). Wir prüften die statistische Signifikanz von festen Effekten, indem wir mit Hilfe eines Likelihood-Quotienten-Tests (R-Funktion „anova") ein gegebenes Modell mit dem Nullmodell verglichen.

2.3.5 Diversitätsindizes

Jeweils getrennt für beide Datensätze nutzten wir i) die Anzahl der Arten/Artengruppen, um den Arten- oder Artengruppenreichtum zu quantifizieren; ii) den Shannon-Diversitätsindex $H' = -\sum_{i=1}^{S} p_i (\log 2\ p_i)$

mit S = Anzahl der Arten und p_i = relative Abundanz der Art i im jeweiligen Datensatz zur Quantifizierung der Arten- oder Artengruppendiversität; iii) den Shannon-Diversitätsindex dividiert durch seinen maximal möglichen Wert für den gegebenen Artenreichtum des Datensatzes zur Quantifizierung der Ausgeglichenheit („evenness") von Arten oder Artengruppen:

$$E = \frac{H'}{H_{max}} \text{ mit } H_{max} = \log_2 S.$$

3 Ergebnisse

3.1 Globale Sensitivität (über alle Arten/ Artengruppen)

Von insgesamt 111 666 Objekten im Testbildsatz entdeckte HiDeFIND 107 778, die globale Gesamt-Sensitivität betrug damit 96.5 %. Unter Berücksichtigung der hierarchischen Abhängigkeitsstruktur der Daten betrug die globale gewichtete mittlere Sensitivität 99.4 % (Tabelle 2). Knapp zwei Drittel der Varianz in der Detektionswahrscheinlichkeit im Testbildsatz wurden von Unterschieden zwischen Videosequenzen erklärt, etwa ein Viertel von Unterschieden zwischen Transekten und gut ein Zehntel von Unterschieden zwischen Erfassungsflügen (Tabelle 2).

Die Detektionswahrscheinlichkeit war unabhängig von der Position der Sensoren (innen versus außen), von der Jahreszeit und von potenziell detektionsrelevanten Umweltbedingungen mit Ausnahme der Sonnenreflexion (Tabelle 3). Mit zunehmender Sonnenreflexion fiel die Detektionswahrscheinlichkeit signifikant ab.

Abbildung 4 kontrastiert exemplarisch falsch negative mit richtig positiven Detektionen des Modells für eine Auswahl an Arten mit hoher Relevanz für die maritime Umweltplanung (siehe auch Diskussion).

3.2 Art- und geschlechtsspezifische Sensitivitäten

Wie zu erwarten waren auch artspezifische Sensitivitäten generell hoch: Tabelle 4 weist neben den artspezifischen Gesamt-Sensitivitäten die gewichteten mittleren Detektionswahrscheinlichkeiten für die zwölf häufigsten Taxa im Testbildsatz sowie für sechs weitere wichtige Zielarten der maritimen Umweltplanung aus (inklusive des Schweinswals). Nur für eine von sechs geschlechtsdimorphen Arten (alles Enten) konnte eine geschlechtsspezifische Sensitivität nachgewiesen werden: Männchen der Trauerente hatten eine etwas niedrigere Detektionswahrscheinlichkeit als Weibchen (Tabelle 4).

3.3 Gesamt-Präzision und Anzahl falsch positiver Modellvorhersagen

Von insgesamt 6 443 717 Modellvorhersagen waren 107 778 richtig positiv, die Gesamt-Präzision betrug damit 1.7 % (die hier berechnete Präzision ist nicht mit der aus anderen Leistungstests im Computersehen vergleichbar, siehe 2.3.4). Falsch positive Detektionen traten auf 99.3 % der insgesamt 2 096 554 Standbilder und auf allen Videosequenzen auf. Die Anzahl falsch positiver Detektionen pro Standbild rangierte von 0 bis 1000 (Deckelung bei 1000, siehe 2.3.4). Der Median betrug 1.

Die Anzahl falsch positiver Modellvorhersagen war unabhängig von der Position der Sensoren (innen *versus* außen) und von der Jahreszeit (Tabelle 5, Abbildungen 5a und 5b). Mit zunehmender Sonnenreflexion nahm jedoch die Anzahl falsch positiver Detektionen zu, und dies war besonders ausgeprägt bei starker Sonnenreflexion der Fall (Tabelle 5, Abbildung 5c). Für Seegang sowie Luft- und Wassertrübung waren signifikante Effekte nachweisbar (Tabelle 5), die Effektgrößen waren jedoch klein (Abbildungen 5d bis 5f).

4 Diskussion

4.1 Sensitivität

Globale Sensitivität

Die globale mittlere Sensitivität von HiDeFIND über alle 124 im Testbildsatz vertretenen Arten/Artengruppen

Tab. 2: Globale gewichtete mittlere Sensitivität des künstlichen neuronalen Netzwerks HiDeFIND über alle 124 Arten/Artengruppen des Testbildsatzes. Ergebnisse eines generalisierten linearen gemischten Modells mit binomialer Fehlerstruktur und Logit-Linkfunktion. – *Global weighted mean sensitivity of the artificial neural network HiDeFIND across all 124 species/species groups of the test image set. Results of a generalized linear mixed model with binomial error structure and logit link function.*

				Feste Effekte – *Fixed effects*	Zufallseffekte – *Random effects*		
Modell – *Model*	**N (Objekte)** – *N (objects)*	**Richtig positiv** – *True positive*	**Falsch negativ** – *False negative*	**Y-Achsenabschnitt (95 % Konfidenz)** – *Intercept (95 % confidence)*[1]	**Erfassungsflüge** – *Surveys*	**Transekte** – *Transects*	**Videosequenzen** – *Reels*
Alle Arten/Artengruppen – *All species/species groups*	111 666	107 778	3888	99.41 (99.15, 99.59)	0.35	0.61	1.60

[1]Gewichtete mittlere Sensitivität. – *Weighted mean sensitivity*

Tab. 3: Globale gewichtete mittlere Sensitivität des künstlichen neuronalen Netzwerks HiDeFIND über alle 124 Arten/Artengruppen des Testbildsatzes in Abhängigkeit potenziell detektionsrelevanter Umweltbedingungen. Ergebnisse generalisierter linearer gemischter Modelle mit binomialer Fehlerstruktur und Logit-Linkfunktion. – *Global weighted mean sensitivity of the artificial neural network HiDeFIND across all 124 species/species groups of the test image set in relation to potentially detection-relevant environmental variation. Results of generalized linear mixed models with binomial error structure and logit link function.*

		Feste Effekte – *Fixed effects*				Zufallseffekte – *Random effects*		
Modell – *Model*	**N (Objekte)** – *N (objects)*		**Chisq**	**FG** – *DF*	**p**[1]	**Erfassungsflüge** – *Surveys*	**Transekte** – *Transects*	**Videosequenzen** – *Reels*
Innere/äußere Sensoren – *Inner/outer sensors*	111 666	Faktor mit zwei Stufen – *Two-level factor*	1.01	1	0.32	0.35	0.61	1.60
Jahreszeit – *Season*	111 666	Faktor mit vier Stufen – *Four-level factor*	2.59	3	0.46	0.27	0.62	1.63
Sonnenreflexion – *Glare*	111 666	Faktor mit vier Stufen – *Four-level factor*	26.10	3	<0.001[2]	0.33	0.61	1.55
Seegang – *Sea state*	111 564[3]	Faktor mit sechs Stufen – *Six-level factor*	3.20	5	0.67	0.29	0.64	1.66
Lufttrübung – *Air turbidity*	111 564[4]	Faktor mit zwei Stufen – *Two-level factor*	0.81	1	0.37	0.35	0.60	1.60
Wassertrübung (Meeressäuger) – *Water turbidity (marine mammals)*	702[5]	Faktor mit zwei Stufen – *Two-level factor*	0.03	1	0.85	Effektiv null – *Effectively zero*[6]	9.43	77.70

[1]Für Faktoren mit mehr als zwei Stufen p-Wert für Omnibus-Test. – *For factors with more than two levels p value for omnibus test.*
[2]Sensitivität sinkt unter mäßiger und insbesondere starker Sonnenreflexion. – *Sensitivity drops under moderate and in particular strong glare.*
[3]Information für Seegang fehlte für 102 Beobachtungen. – *Information on sea state missing for 102 observations.*
[4]Information für Lufttrübung fehlte für 102 Beobachtungen. – *Information on air turbidity missing for 102 observations.*
[5]Information für Wassertrübung fehlte für eine Beobachtung. – *Information on water turbidity missing for one observation.*
[6]Singularität (Varianz nicht von null abgrenzbar). – *Singularity (variance indistinguishable from zero).*

Tab. 4: Artspezifische gewichtete mittlere Sensitivitäten des künstlichen neuronalen Netzwerks HiDeFIND für die zwölf häufigsten Taxa im Testbildsatz sowie für sechs weitere ausgewählte Arten mit hoher Relevanz für die maritime Umweltplanung. Ergebnisse generalisierter linearer gemischter Modelle mit binomialer Fehlerstruktur und Logit-Linkfunktion. – *Species-specific weighted mean sensitivity of the artificial neural network HiDeFIND for the twelve most common taxa in the test image set plus six further selected species with high relevance for maritime environmental planning. Results of generalized linear mixed models with binomial error structure and logit link function.*

				Feste Effekte *– Fixed effects*	**Zufallseffekte** *– Random effects*		
Modell *– Model*	**N (Objekte)** *– N (objects)*	**Richtig positiv** *– True positive*	**Falsch negativ** *– False negative*	**Y-Achsenabschnitt (95 % Konfidenz)** *– Intercept (95 % confidence)*[1]	**Erfassungsflüge** *– Surveys*	**Transekte** *– Transects*	**Videosequenzen** *– Reels*
Eiderente *– Common Eider*	25513	25322	191	99.73 (99.57, 99.83)[2]	Effektiv null *– Effectively zero*[3]	0.37	1.63
Trauerente *– Common Scoter*	24567	23116	1451	98.81 (97.87, 99.34)[4]	Effektiv null *– Effectively zero*[3]	1.77	1.48
Trottellumme *– Common Guillemot*	12215	12146	69	99.99 (99.98, 100)	Effektiv null *– Effectively zero*[3]	Effektiv null *– Effectively zero*[3]	36.40
Tordalk *– Razorbill*	6087	6027	60	99.99 (99.98, 100)	Effektiv null *– Effectively zero*[3]	Effektiv null *– Effectively zero*[3]	52.76
Dreizehenmöwe *– Black-legged Kittiwake*	5154	5124	30	99.99 (99.90, 100)	Effektiv null *– Effectively zero*[3]	Effektiv null *– Effectively zero*[3]	24.09
Trottellumme/Tordalk *– Guillemot/Razorbill*	4174	4127	47	99.99 (99.95, 99.99)	Effektiv null *– Effectively zero*[3]	Effektiv null *– Effectively zero*[3]	46.07
Lachmöwe *– Black-headed Gull*	3722	3673	49	99.96 (99.59, 100)	Effektiv null *– Effectively zero*[3]	Effektiv null *– Effectively zero*[3]	19.04
Höckerschwan *– Mute Swan*	2153	2151	2	100 (99.92, 100)	Effektiv null *– Effectively zero*[3]	Effektiv null *– Effectively zero*[3]	84.12
Silbermöwe *– Herring Gull*	1795	1728	67	100 (99.98, 100)	Effektiv null *– Effectively zero*[3]	Effektiv null *– Effectively zero*[3]	135.70
Sturmmöwe *– Common Gull*	1590	1575	15	100 (99.96, 100)	Effektiv null *– Effectively zero*[3]	Effektiv null *– Effectively zero*[3]	76.95
Samtente *– Velvet Scoter*	1462	1432	30	100 (99.34, 100)[5]	Effektiv null *– Effectively zero*[3]	Effektiv null *– Effectively zero*[3]	24.57

Tab. 4: Fortsetzung

				Feste Effekte – *Fixed effects*	Zufallseffekte – *Random effects*		
Modell – *Model*	**N (Objekte)** – *N (objects)*	**Richtig positiv** – *True positive*	**Falsch negativ** – *False negative*	**Y-Achsenabschnitt (95 % Konfidenz)** – *Intercept (95 % confidence)*[1]	**Erfassungsflüge** – *Surveys*	**Transekte** – *Transects*	**Videosequenzen** – *Reels*
Heringsmöwe – *Lesser Black-backed Gull*	1275	1275	0	Nicht schätzbar – *Not estimable* [6]	–	–	–
Kormoran – *Great Cormorant*	1193	1048	145	99.82 (98.37, 99.98)	Effektiv null – *Effectively zero*[3]	0.88	22.63
Eisente – *Long-tailed Duck*	1157	1143	14	99.93 (80.07[7], 100)[8]	Effektiv null – *Effectively zero*[3]	Effektiv null – *Effectively zero*[3]	14.72
Basstölpel – *Northern Gannet*	746	744	2	100 (99.48, 100)	24.12	Effektiv null – *Effectively zero*[3]	332.78
Pfeifente – *European Wigeon*	728	699	29	98.63 (94.90, 99.64)[9]	Effektiv null – *Effectively zero*[3]	Effektiv null – *Effectively zero*[3]	4.33
Sterntaucher – *Red-throated Diver*	540	538	2	99.63 (98.86, 99.94)	Effektiv null – *Effectively zero*[3]	Effektiv null – *Effectively zero*[3]	Effektiv null – *Effectively zero*[3]
Schweinswal – *Harbour Porpoise*	406	401	5	99.61 (91.88, 99.98)	2.72	Effektiv null – *Effectively zero*[3]	Effektiv null – *Effectively zero*[3]

[1] Gewichtete mittlere Sensitivität. – *Weighted mean sensitivity.*

[2] Geschlechtsspezifität: p=0.51, N=7797 geschlechtsbestimmte Individuen. – *Sex specificity: p=0.51, N=7797 sexed individuals.*

[3] Singularität (Varianz nicht von null abgrenzbar). – *Singularity (variance indistinguishable from zero).*

[4] Geschlechtsspezifität: p<0.001, N=16036 geschlechtsbestimmte Individuen; Männchen mit 0.7 % niedrigerer Sensitivität. – *Sex specificity: p<0.001, N=16036 sexed individuals; males with 0.7 % lower sensitivity.*

[5] Geschlechtsspezifität: p=0.29, N=1029 geschlechtsbestimmte Individuen. – *Sex specificity: p=0.29, N=1029 sexed individuals.*

[6] Komplette Separation: Nur richtig positive Vorhersagen. – *Complete separation: Only true positive predictions.*

[7] Eine von 109 Videosequenzen mit Eisenten steuerte mit 120 Individuen >10 % der Stichprobe bei (alles richtig positive Vorhersagen). – *One of 109 reels with Long-tailed Ducks contributed with 120 individuals >10 % of sample (all true positive predictions).*

[8] Geschlechtsspezifität: Nicht schätzbar, komplette Separation (alle 13 Weibchen detektiert). – *Sex specificity: Not estimable, complete separation (all 13 females detected).*

[9] Geschlechtsspezifität: p=0.73, N=191 geschlechtsbestimmte Individuen. – *Sex specificity: p=0.73, N=191 sexed individuals.*

Tab. 5: Anzahl falsch positiver Vorhersagen pro Standbild des künstlichen neuronalen Netzwerks HiDeFIND in Abhängigkeit potenziell detektionsrelevanter Umweltbedingungen. Ergebnisse log-linearer gemischter Modelle mit normaler Fehlerstruktur. – *Number of false positive predictions per frame of the artificial neural network HiDeFIND in relation to potentially detection-relevant environmental conditions. Results of log-linear mixed models with Gaussian error structure.*

		Feste Effekte *– Fixed effects*				**Zufallseffekte** *– Random effects*			
Modell *– Model*	**N (Standbilder)** *– N (frames)*		**Chisq**	**FG** *– DF*	**p**[1]	**Erfassungsflüge** *– Surveys*	**Transekte** *– Transects*	**Videosequenzen** *– Reels*	**Unerklärt** *– Residual*
Innere/äußere Sensoren *– Inner/outer sensors*	2 096 554	Faktor mit zwei Stufen *– Two-level factor*	1.86	1	0.17	Nicht konvergiert *– Not converged*	Nicht konvergiert *– Not converged*	0.022	0.090
Jahreszeit *– Season*	2 096 554	Faktor mit vier Stufen *– Four-level factor*	5.45	3	0.14	0.003	0.011	0.005	0.090
Sonnenreflexion *– Glare*	2 096 554	Faktor mit vier Stufen *– Four-level factor*	451.35	3	<0.001[2]	0.005	0.011	0.005	0.090
Seegang *– Sea state*	2 090 829[3]	Faktor mit sechs Stufen *– Six-level factor*	454.20	5	<0.001[4]	0.005	0.011	0.005	0.090
Lufttrübung *– Air turbidity*	2 090 829[5]	Faktor mit zwei Stufen *– Two-level factor*	5.73	1	0.02[6]	0.090	0.011	0.005	0.090
Wassertrübung *– Water turbidity*	2 090 829[7]	Faktor mit zwei Stufen *– Two-level factor*	64.40	1	<0.001[8]	0.090	0.017	0.005	0.090

[1] Für Faktoren mit mehr als zwei Stufen p-Wert für Omnibus-Test. – *For factors with more than two levels p value for omnibus test.*
[2] Falsch Positive nehmen mit zunehmender Sonnenreflexion zu (siehe Abb. 5). – *False positives increase with increasing glare (see figure 5).*
[3] Information für Seegang fehlte für 5725 Beobachtungen. – *Information on sea state missing for 5725 observations.*
[4] Fünf von 15 Kontrasten signifikant, aber kein Trend (siehe Abb. 5). – *Five out of 15 contrasts significant, but no trend (see figure 5).*
[5] Information für Lufttrübung fehlte für 5725 Beobachtungen. – *Information on air turbidity missing for 5725 observations.*
[6] Minimal höherer Median für Stufe keine *versus* etwas Lufttrübung (siehe Abb. 5). – *Minimally higher median for level no versus some air turbidity (see figure 5).*
[7] Information für Wassertrübung fehlte für 5725 Beobachtungen. – *Information on water turbidity missing for 5725 observations.*
[8] Höherer Median für Stufe etwas *versus* keine Wassertrübung (siehe Abb. 5). – *Higher median for level some versus no water turbidity (see figure 5).*

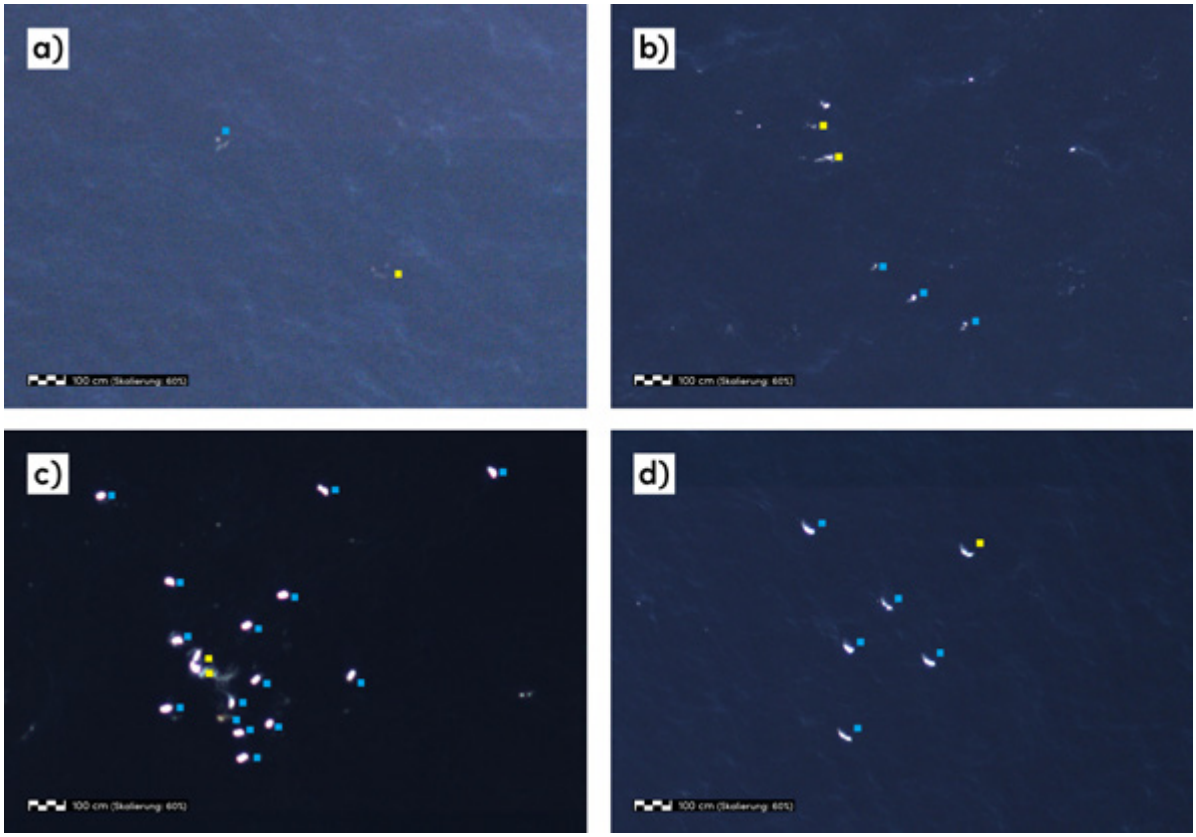


Abb. 4: Beispiele falsch negativer Modellvorhersagen. Die Abbildung zeigt exemplarisch HiDef-Standbilder, auf denen es für dieselbe Art sowohl richtig positive Detektionen (blau annotiert) als auch falsch negative Detektionen (gelb annotiert) gab. a) Trauerenten *Melanitta nigra*, b) Tordalken *Alca torda*, c) Eiderenten *Somateria mollissima*, d) Sterntaucher *Gavia stellata*. – *Examples of false negative model predictions. The figure shows in an exemplary manner HiDef frames featuring both true positive detections (blue annotations) and false negative detections (yellow annotations) for the same species. a) Common Scoters* Melanitta nigra*, b) Razorbills* Alca torda*, c) Common Eiders* Somateria mollissima*, d) Red-throated Divers* Gavia stellata.

(davon 109 Vogeltaxa) war mit über 99 % sehr hoch. HiDeFIND hat demnach Vögel auf See sowie Meeressäuger auf HiDef-Videomaterial praktisch genauso gut entdeckt wie speziell für diesen Zweck geschulte Beobachter:innen. Ein 95 % Konfidenzintervall von 99.2-99.6 % weist die Schätzung überdies als sehr präzise aus und lässt erwarten, dass das Modell für HiDef-Flugerfassungen mit ähnlichen Artenspektren und Wetterbedingungen eine vergleichbar hohe Sensitivität erzielen können wird. Wir fanden außerdem keinen nachweisbaren Unterschied in der mittleren Sensitivität zwischen den inneren und äußeren Sensoren (Tabelle 3). HiDeFIND hat demnach auf der gesamten Streifenbreite von effektiv 544 m eine vergleichbar gute Leistung erzielt.

Die mittels gemischten Modells geschätzte globale mittlere Sensitivität fiel etwas höher aus als die globale Gesamt-Sensitivität. Erstere berücksichtigte die hierarchische Abhängigkeitsstruktur der HiDef-Daten und gewichtete die Beiträge von Gruppen abhängiger Daten zum Gesamteffekt. Diskrepanzen zwischen beiden Maßzahlen können sich zum Beispiel ergeben, wenn sich falsch negative Detektionen auf relativ wenige Standbilder und/oder Videosequenzen konzentrieren. In diesem Fall zeigt die Gesamt-Sensitivität dann eine Tendenz, die tatsächliche Sensitivität zu unterschätzen. Eine derartige Klumpung war im Testdatensatz in der Tat nachweisbar: Zum Beispiel waren 947 von insgesamt 3888 (24.4 %) falsch negativen Detektionen gerade einmal zehn der insgesamt 51 218 (0.02 %) Standbildern zuzuordnen, die mindestens ein biologisches Objekt enthielten.

Die im Rahmen des etablierten Arbeitsprozesses manuell detektierten Objekte waren der Maßstab für die Beurteilung der Sensitivität. Wenn im manuellen Prozess Objekte übersehen wurden, ist anzunehmen, dass Teile der als falsch positiv gewerteten Modellvorhersagen (siehe 4.2 unten) tatsächlich richtig positive Modellvorhersagen (falsch negative manuelle Detektionen) waren. Unser Testdesign konnte diese allerdings nicht als solche erkennen. Die sehr hohe globale Sensitivität lässt vermuten, dass HiDeFIND auch einen Teil der im manuellen Arbeitsprozess übersehenen Objekte finden könnte. Dieses vielversprechende zusätzliche Potenzial

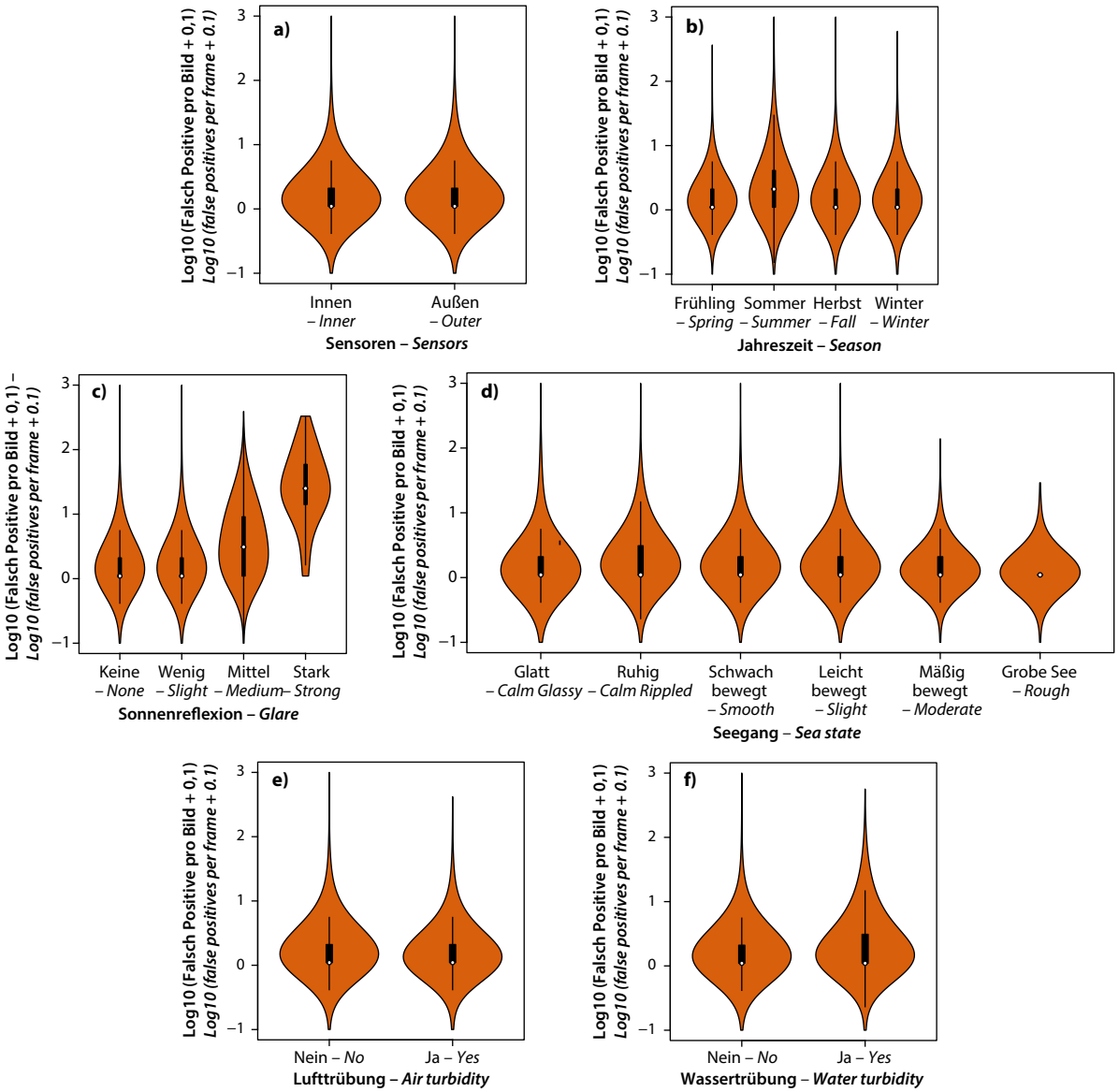


Abb. 5: Anzahl falsch positiver Modellvorhersagen pro Standbild in Abhängigkeit von a) Position der Sensoren, b) Jahreszeit, c) Sonnenreflexion, d) Seegang (Petersen-Skala), e) Lufttrübung und f) Wassertrübung (siehe auch Anhang 1). – *Number of false positive model predictions per frame in relation to a) positions of the sensors, b) seasons, c) glare, d) sea state (Petersen scale), e) air turbidity and f) water turbidity (see also electronic supplementary material 1).*

soll an unabhängigem Testbildmaterial in zukünftigen Tests mit geeignetem Design quantifiziert werden.

Generalisierbarkeit

Obwohl HiDeFIND mit annotierten Objekten aus 66 Arten/Artengruppen und zwei Meeresgebieten trainiert wurde, erzielte das Modell seine Leistung auf einem Testbildsatz, der weit über hundert Arten/Artengruppen aus drei Meeresgebieten umfasste (Anhang 2). Dies unterstreicht, dass HiDeFIND gelernt hat zu generalisieren. Für künstliche neuronale Netzwerke ist genau das erwünscht und der Befund lässt erwarten, dass HiDeFIND hohe Sensitivitäten auch dann wird erzielen können, wenn es mit erweiterten oder völlig neuen Artenspektren konfrontiert wird. Dies könnte beim Einsatz in neuen Meeresgebieten sowie bei Verschiebungen von Verbreitungsarealen z. B. als Folge anthropogenen Umweltwandels von großem Vorteil sein.

Artspezifische Sensitivitäten

Wie zu erwarten waren die artspezifischen Sensitivitäten für die meisten Arten ebenfalls sehr hoch (Tabelle 4). Dies schloss die Detektion von Arten mit herausragender Bedeutung für die maritime Raum- und Um-

weltplanung ein, wie z. B. Sterntaucher, Trottellumme und Dreizehenmöwe (Furness et al. 2013, Fliessbach et al. 2019, Dierschke et al. 2024) sowie unter den Meeressäugern den Schweinswal. Artspezifische Sensitivitäten unterschieden sich außerdem generell nur wenig (Tabelle 4). Dies legt nahe, zukünftigen Entwicklungsaufwand auf die weitere Optimierung eines generellen HiDeFIND-Modells zu konzentrieren, statt mehrere art- oder artgruppenspezifische HiDeFIND-Spezialmodule zu entwickeln (wie z. B. ein Seetaucher-Modul).

Geschlechtsspezifische Sensitivitäten
Bei ausgeprägt geschlechtsdimorphen Arten könnte die Sensitivität des Modells geschlechtsspezifisch sein. Dies konnten wir für einige Entenarten prüfen, für die wir aufgrund ihrer Häufigkeit im Testbildsatz eine adäquate Teststärke voraussetzen durften. Während für Eiderenten, Samtenten *Melanitta fusca* und Pfeifenten *Anas penelope* die Sensitivitäten für die Geschlechter statistisch nicht unterscheidbar waren, wies HiDeFIND für Männchen der Trauerente eine etwas niedrigere Sensitivität auf als für die Weibchen (Tabelle 4, siehe auch Abbildung 4a). Vor dem Hintergrund einer artspezifischen Sensitivität von >98 % war dieser Effekt mit einer Differenz von 0.7 % jedoch sehr klein, so dass er für Bestandsschätzungen vernachlässigt werden kann.

Umweltvariation
Die globale mittlere Sensitivität nahm mit zunehmender Sonnenreflexion signifikant ab (Tabelle 3). Dies ist wenig überraschend, da mit zunehmender Sonnenreflexion Objekte auf der Meeresoberfläche leicht überstrahlt und so maskiert werden können. Im Vergleich zum Trainingsbildsatz war im Testbildsatz außerdem im Mittel weniger intensive Sonnenreflexion zu verzeichnen (alle anderen detektionsrelevanten Umweltbedingungen waren ähnlich; siehe Anhang 2). Dies könnte die hohe globale Sensitivität im Testdatensatz begünstigt haben. Die Beurteilung der Modellsensitivität würde dies jedoch nur dann betreffen, wenn die mittlere Sonnenreflexion im Testbildsatz außergewöhnlich niedrig und damit für zukünftige Bildsätze nicht repräsentativ gewesen sein sollte. Auf die Gesamtleistung des Modells wirkt sich dies ohnehin nur sehr begrenzt aus, da durch Wegschwenken des HiDef-Sensorensystems von der Sonne stärkere Sonnenreflexion meist effektiv vermieden wurde (siehe 2.1.1). Entsprechend entfielen nur 3504 (0.03 %) beziehungsweise 158 (0.001 %) Objekte im Testdatensatz auf Videomaterial mit mäßiger beziehungsweise starker Sonnenreflexion. Anderen potenziell detektionsrelevanten Umweltbedingungen gegenüber zeigte sich HiDeFIND robust: Weder Seegang noch Lufttrübung oder Wassertrübung (bei Meeressäugern) hatten einen nachweisbaren Effekt auf die Sensitivität (Tabelle 3). Dasselbe galt für die Jahreszeiten (Tabelle 3). Dies lässt den Schluss zu, dass HiDeFIND grundsätzlich ganzjährig und unter vielen, auch suboptimalen Umweltbedingungen ohne Einschränkungen bei der Erfassungsqualität eingesetzt werden kann.

Ursachen falsch negativer Detektionen
Statistisch können Bedingungen eingegrenzt werden, mit denen die Wahrscheinlichkeit falsch negativer Detektionen variiert (siehe voriger Abschnitt). In jedem Einzelfall ist es jedoch für viele falsch negative Detektionen in der Praxis unmöglich, festzumachen, warum ausgerechnet die hier betroffenen Objekte vom Modell nicht erkannt wurden. Diese eingeschränkte Transparenz des maschinellen Entscheidungsprozesses ist kennzeichnend für auf „„Deep Learning““ basierende neuronale Netzwerke, deren sukzessive aufeinander aufbauenden Berechnungen über viele Netzwerk-Schichten in weiten Teilen sogenannten *Black-Box*-Operationen entsprechen. Dies führte zu quasi-stochastischem Auftreten falsch negativer Detektionen, über deren Ursachen nur spekuliert werden kann.

Der insgesamt sehr guten Generalisierbarkeit zum Trotz besteht die Möglichkeit, dass Arten oder Artengruppen, die nicht oder nur sehr selten im Trainingsbildsatz enthalten waren, im Testbildsatz schlechter erkannt wurden als im Trainingsbildsatz häufig vertretene Arten. Der Trainingsbildsatz enthielt zum Beispiel keine Blässrallen *Fulica atra* und das Modell hat nur zwei von fünf Blässrallen im Testbildsatz detektiert. Solche Fälle betreffen allerdings vor allem Arten, die in den beflogenen Projektgebieten selten und daher in aller Regel von untergeordneter Relevanz sind. Für häufigere Arten könnten falsch negative Detektionen resultieren, wenn sich Eigenschaften wie Körperhaltung oder Tauchstatus in ihrer Häufigkeit zwischen Trainings- und Testbildsatz unterscheiden. So betraf etwa eine falsch negative Detektion den Schnappschuss eines Schweinswals im Tauchvorgang, dessen Körper so zweigeteilt erschien. Abbildungen 5a-5c zeigen drei Beispiele falsch negativer Vorhersagen.

Wir haben eine Modellvorhersage dann als richtig positiv bewertet, wenn die vom Modell vorhergesagte Bounding Box die manuell gesetzte Punktmarkierung einschloss (siehe 2.2.5.2). Wenn eine Punktmarkierung nicht präzise im Zentrum des Objekts gesetzt und daher knapp außerhalb einer modell-generierten Bounding Box verortet war, konnte eine falsch negative Detektion begründet werden, obwohl das Modell das Objekt tatsächlich detektiert hatte („falsch falsch negative Detektionen“). Eine stichprobenartige Überprüfung ergab, dass weniger als 10 % aller falsch negativen Detektionen auf diese Unzulänglichkeit des Testdesigns zurückzuführen waren, unter ihnen Fälle, bei denen die Objekte auch für menschliche Beobachter:innen kaum zu übersehen waren (ein Beispiel dafür gibt ein nur scheinbar übersehener Sterntaucher in Abbildung 4d). Das Phänomen ist ein rein test-technisches Problem, führt allerdings nicht nur zu einer Unterschätzung der tatsächlichen Sensitivität des Modells, sondern gleich-

zeitig auch zu einer leichten Überschätzung der Falsch-Positiv-Rate, da die an unmittelbar benachbarter Stelle tatsächlich vom Modell detektierten Objekte als falsch positive Detektionen ausgewiesen wurden. Die von uns in diesem Beitrag vorgenommene HiDeFIND-Leistungsbewertung ist demnach als konservativ zu betrachten.

Alternative KI-Modelle mit ähnlicher Zielstellung

Während maschinelles Lernen in Kombination mit Drohnen oder Satelliten als Plattformen bereits sehr erfolgreich für das Monitoring von Tierbeständen auf See eingesetzt wurde (siehe Einleitung), gibt es bisher nur sehr wenige veröffentlichte Arbeiten zur Kombination von maschinellem Lernen mit flugzeuggestützten Erhebungen zur Bestandserfassung auf See. Ein direkter Vergleich von spezifischen Leistungskennzahlen konkurrierender Modelle im Computersehen ist generell nur eingeschränkt aussagekräftig, da sowohl dem Training als auch dem Testen meist ganz unterschiedliche Bildsätze zu Grunde liegen. Nach unserer Kenntnis weist bisher weniger als eine Handvoll begutachteter Studien ein Design auf, das dem unseren – Plattform: Flugzeug; Sensor: Digitales Kamera-System; Zielobjekt: Seevögel auf See – ähnlich genug ist, um überhaupt einen Vergleich ziehen zu können.

Kuru et al. (2023) beschreiben eine Methode zur halbautomatischen Detektion von Basstölpeln auf Luftbildaufnahmen, die in Art und Auflösung HiDef-Standbildern ähneln. Kuru et al. (2023) erzielten für ihr Modell eine ungewichtete Sensitivität von 97.1%[1], die der von HiDeFIND für den Basstölpel erreichten ungewichteten Sensitivität von 99.7 % vergleichbar ist (siehe auch Tabelle 4). Im Gegensatz zum breiten taxonomischen Anwendungsbereich unseres Ansatzes (Abbildung 3, Anhang 2) beschränkte sich die von Kuru et al. (2023) beschriebene Methode und/oder deren quantitative Leistungsanalyse jedoch auf den Basstölpel, die größte und auffälligste Seevogelart in europäischen Gewässern. Wie geeignet die von Kuru et al. (2023) beschriebene Methode für kleinere und unauffälligere Arten ist, blieb daher offen. Möglicherweise schloss die Analyse von Kuru et al. (2023) nur Individuen im besonders auffälligen Adultkleid ein (Abbildungen der Arbeit zeigten nur adulte Individuen im Flug). Dagegen enthielten sowohl unser Trainingsbildsatz als auch unser Testbildsatz neben (Sub-) Adulten auch Individuen in den weniger auffälligen Kleidern des 1. und 2. Kalenderjahres (160 im Trainingsbildsatz und 10 im Testbildsatz, letztere vollständig richtig positiv detektiert) sowie schwimmende Basstölpel (jeweils Hunderte in beiden unseren Bildsätzen). Darüber hinaus berichten Ke et al. (2024) Sensitivitäten von bis zu 65 % in einem „Deep Learning" -Ansatz, der Bestandsschätzungen von Seevögeln auf See während des Fluges und damit nahezu in Echtzeit erlauben soll. Aufgrund von sehr niedrigen Flughöhen von nur etwa 25 m bis 200 m standen diesem Modell Bilddaten mit einer sehr hohen Bodenauflösung von zwischen 0.14 cm und 1.47 cm zur Verfügung, was sich naturgemäß positiv auf Leistungstests von Modellen zur Objekterkennung auswirkt. Zudem waren die von Ke et al. (2024) angebotenen Objekte für Training und Testen auf nur zwei Probeflächen und weitgehend auf überwinternde Meeresenten beschränkt. HiDeFIND erzielte höhere Sensitivitäten, obwohl es mit Bilddaten konfrontiert war, die aus Gründen der Flugsicherheit aus deutlich größerer Höhe von etwa 500 m aufgenommen worden waren und dementsprechend schlechtere Bodenauflösungen von etwa 2 cm aufwiesen. Überdies deckte das HiDeFIND-Bildmaterial ein viel breiteres Artenspektrum ab (Abbildung 3, Anhang 2). Des Weiteren präsentierten Weiser et al. (2023) eine Methode zur automatisierten Erfassung von rastenden Meeresgänsen in geschütztem Flachwasser an einem Standort in Alaska. Obwohl das Modell mit gutem Erfolg Bilder mit von solchen ohne Meeresgänse abgrenzen konnte, hat die sich anschließende automatisierte Auszählung auf Basis von Individuen die Bestände unterschätzt (mangelnde Sensitivität), so dass aufwändige manuelle Nacharbeiten im Rahmen eines „human-in-the-loop" Arbeitsprozesses erforderlich waren.

4.2 Präzision und Anzahl falsch positiver Modellvorhersagen

Präzision

HiDeFIND generierte eine hohe Anzahl falsch positiver Detektionen, die sich in einer niedrigen ungewichteten Gesamt-Präzision widerspiegelten. Diese Tendenz zur Überdetektion war als Konstruktionsmerkmal ausdrücklich erwünscht, da wir im Zielkonflikt zwischen Sensitivität (möglichst vollständige Erfassung) und Präzision (möglichst hohe Effizienz) ersterer Priorität eingeräumt hatten, um bei der Erfassungsqualität im Vergleich zum etablierten manuellen Prozess keine Abstriche machen zu müssen. In einem teil-automatisierten „human-in-the-loop" Arbeitsprozess zieht niedrige Präzision allerdings hohen manuellen Aufwand nach sich, um falsch positive Detektionen vor Übergabe an die sich anschließende Artbestimmung abzuscheiden. Durch die Automatisierung der initialen Objektdetektion erzielbare Effizienzsteigerungen werden dadurch geschmälert. Ein vordringliches Ziel bei der Weiterentwicklung eines KI-unterstützten Arbeitsprozesses ist daher, die Präzision zu verbessern ohne substanzielle Abstriche bei der Sensitivität machen zu müssen.

[1] Errechnet aus Tabelle 3 in Kuru et al. (2023). Im Gegensatz zu unserer Analyse lagen deren Berechnung nicht Individuen als Einheit zu Grunde, sondern Luftbilder, die entweder mindestens einen oder aber keinen Basstölpel zeigten. Als richtig positive Detektion wurde gewertet, wenn die Methode ein Bild, das mindestens einen Basstölpel zeigte, als ein solches erkannte.

Position der Sensoren und Umweltvariation
Die Position der Sensoren (Tabelle 5, Abbildung 5a) sowie die Jahreszeiten (Tabelle 5, Abbildung 5b) übten keinen nachweisbaren Einfluss auf die Anzahl falsch positiver Detektionen aus. Mit zunehmender Sonnenreflexion löste das Modell allerdings vermehrt falsch positive Detektionen aus. Dies war besonders ausgeprägt für starke Sonnenreflexion (Tabelle 5, Abbildung 5c). Weitere potenziell detektionsrelevante Umweltbedingungen wie Seegang, Lufttrübung oder Wassertrübung zeigten zwar signifikante Unterschiede (Tabelle 5), die Effekte waren jedoch klein (Abbildungen 5d bis 5f). Aufgrund der großen Stichprobe von >2 Millionen Standbildern war die Teststärke unserer Analysen hoch und erlaubte den Nachweis auch kleiner Unterschiede zwischen einzelnen Faktorenstufen. Diese kleinen Effekte dürften in der Anwendungspraxis kaum Relevanz haben, auch weil Erfassungsflüge in aller Regel nur für Zeitfenster terminiert werden, die gute Erfassungsbedingungen wie klare Luft und wenig Seegang erwarten lassen. In zukünftigen Leistungstests von HiDeFIND selbst, aber auch von vergleichbaren alternativen Modellen zur Detektion von Seevögeln auf See sollten die Effekte detektionsrelevanter Umweltbedingungen nichtsdestotrotz genauer erforscht werden.

Minderung von Effekten der Sonnenreflexion
Verbesserungen der HiDeFIND-Präzision und damit Verbesserungen der Effizienz des gesamten KI-unterstützten Arbeitsprozesses könnten insbesondere an einer weiteren Minimierung der Effekte der Sonnenreflexion ansetzen, die über die etablierten technischen Maßnahmen hinausgeht (siehe 2.2.1). Eine Abmilderung der Effekte der Sonnenreflexion könnte erzielt werden, indem die Empfindlichkeit des Modells kontrolliert reduziert wird. Dies kann durch die Erhöhung von Schwellenwerten in der Postproduktion geschehen, die die Konfidenz definieren, mit der das Modell ein visuelles Muster angesprochen haben muss, um eine Detektion auszulösen. Im Sinne einer möglichst vollständigen Erfassung war für die getestete HiDeFIND-Version 1.0 kein solcher Schwellenwert definiert worden. Um durch starke Sonnenreflexion verursachte Spitzen falsch positiver Detektionen abzumildern, könnte alternativ auch die maximal zulässige Anzahl an Detektionen pro Standbild abgesenkt werden, indem die auf einem Standbild ausgelösten Modell-Detektionen im Nachgang absteigend nach Konfidenz sortiert werden. Unter der Annahme, dass richtig positive Detektionen im Mittel höhere Konfidenzen aufweisen als falsch positive, wären auf Standbildern mit einer Mischung aus richtig positiven und falsch positiven Detektionen dann in erster Linie falsch positive Detektionen betroffen. Für den hier vorgestellten Leistungstest war die Anzahl zugelassener Detektionen pro Standbild auf 1000 beschränkt (49 Standbilder waren von dieser Deckelung betroffen), die höchste Anzahl relevanter Objekte betrug jedoch in einem Einzelfall nur 811. Für die restlichen 51 217 Standbilder mit mindestens einem relevanten Objekt betrug die Anzahl relevanter Objekte durchgängig weniger als 100. Auf Basis einer Stichprobe zukünftiger Leistungstests sollten die Effekte einer Absenkung der maximalen Anzahl zugelassener Detektionen pro Standbild geprüft werden.

Zielkonflikt Präzision versus Sensitivität
Sowohl eine Erhöhung der Schwellenwerte für die Konfidenz von Detektionen als auch eine Absenkung maximal zulässiger Detektionen pro Standbild könnte mit einer Minderung der globalen Sensitivität einhergehen. Das vordringliche Ziel künftiger Weiterentwicklung wird daher sein, die Präzision bei hoher Sensitivität zu erhöhen, um Vorteile der Automatisierung wie Beschleunigung, erhöhte Reproduzierbarkeit und gesteigerte Kosteneffektivität voll ausnutzen zu können. In jedem Fall sollte eine Ergebnisqualität gewährleistet sein, die mindestens gleich hoch ist wie die des etablierten rein manuellen Arbeitsprozesses. Das Erreichen dieses übergeordneten Entwicklungsziels wird durch technische Innovation bei den Sensoren erleichtert, die mittlerweile bei vergleichbarer Flughöhe und Schwadbreite eine merklich bessere mittlere Bodenauflösung von deutlich unter 2 cm erreichen (BioConsult SH, eigene Daten).

4.3 Qualitätssicherung KI-unterstützter Bestandserfassungen auf See

KI-unterstützte Methoden bieten große Chancen für die Auswertung von digitalen Zählflugdaten in der maritimen Raum- und Umweltplanung, insbesondere auch bei der Auswertegeschwindigkeit. Für die Bewertung von Bestand und Verbreitung von Seevögeln auf See und damit die Rechtssicherheit von Planungsvorhaben sind Qualität, Kontinuität und Vergleichbarkeit der erhobenen Daten von elementarer Bedeutung. Bei der Transformation von manuellen zu (teil-) automatisierten Arbeitsprozessen sollte daher sichergestellt werden, dass Leistung und Praxistauglichkeit von KI-Modellen vor Anwendung umfassend evaluiert werden. Überdies sollte die praktische Anwendung von einem Qualitätssicherungskonzept begleitet werden, in dem definiert ist, wie die Ergebnisqualität bei automatisierter Auswertung von digitalen Erfassungsflügen sichergestellt ist.

4.4 Fazit

Das künstliche neuronale Netzwerk HiDeFIND detektierte Seevögel auf See sowie Meeressäugetiere auf digitalen Videoaufnahmen fast genauso gut wie speziell geschulte Beobachter:innen. Dies galt auch für Schlüsselarten der maritimen Raum- und Umweltplanung, unabhängig von der Jahreszeit sowie weitgehend unabhängig von detektionsrelevanten Umweltbedingungen. Dabei zeigte das Modell eine gute Generali-

sierbarkeit, die maßgeblich auf den sehr umfangreichen und diversen Trainingsbildsatz zurückzuführen sein dürfte und einen erfolgreichen Einsatz auch in anderen Meeresgebieten mit neuen Artenspektren verspricht. Die hohe globale Sensitivität lässt überdies erwarten, dass HiDeFIND zukünftig auch einen Teil der im manuellen Arbeitsprozess übersehenen Objekte zusätzlich finden könnte. Die hohe Rate falsch positiver Detektionen mindert derzeit noch die Effizienz eines KI-unterstützen Arbeitsprozesses und damit die erzielbare Zeitersparnis und Wirtschaftlichkeit seines Einsatzes. Dank hervorragender Sensitivität mindert dies jedoch in keiner Weise die Effektivität des Modells, also die hohe Qualität der Ergebnisse und der daraus ableitbaren Bestands- oder Dichteschätzungen. Für die Auswertung digitaler Flugerfassungsdaten in der maritimen Raum- und Umweltplanung, im Umweltmonitoring sowie in der Forschung bietet HiDeFIND damit eine zukunftsweisende Alternative zu rein manueller Objektdetektion. HiDeFIND operiert dabei im Rahmen eines integrierten „human-in-the-loop" Arbeitsprozesses, bei dem eine automatisierte Objektdetektion durch KI von einem Qualitätssicherungsprozess durch geschulte Beobachter:innen flankiert wird.

Dank

Wir danken Claudia Burger, Kelly Macleod und besonders Hanna Kreutzfeldt und Anna Kersten für Diskussion und hilfreiche Kommentare zu früheren Versionen dieses Manuskripts. Christian Vlasak lieferte das Grafikdesign für Abbildung 1 und die Abbildung im Anhang 3. Venela Matz unterstützte uns bei Auswahl und Annotation des Bildmaterials in Abbildung 4. Wir danken Wolfgang Fiedler und einem anonymen Gutachter für hilfreiche Kommentare.

Interessenkonflikte

Tim Schmoll, Guruprasad Hegde, Monika Dorsch und Georg Nehls arbeiten für die BioConsult SH GmbH & Co KG, die gewinnorientiert digitale Offshore-Flugerfassungen nach der HiDef-Methode sowie deren KI-unterstützte Auswertung anbietet.

5 Zusammenfassung

Die zuverlässige Schätzung von Vogelbeständen auf See bildet eine wichtige Grundlage für die Bewertung der Auswirkungen anthropogener Nutzungsansprüche auf die Meeresumwelt. Der Einsatz künstlicher Intelligenz (KI) für eine automatisierte Auswertung von Luftbildern verspricht dabei eine schnellere, kostengünstigere und besser reproduzierbare Analyse im Vergleich zur manuellen Bearbeitung. Unklar ist jedoch, ob ein KI-unterstützter Arbeitsprozess eine Ergebnisqualität erzielen kann, die derjenigen geschulter Beobachtungspersonen vergleichbar ist. Dies wäre eine notwendige Voraussetzung für eine Etablierung automatisierter Objekterkennung als Standard in der maritimen Raum- und Umweltplanung. Wir beschreiben hier Architektur, Training und Testung des Objekterkennungsmodells HiDeFIND (Version 1.0), eines künstlichen neuronalen Netzwerks mit mehr als 400 Schichten und mehr als 86 Millionen Parametern. Das Modell wurde mit mehr als 138 000 auf HiDef-Videomaterial annotierten Objekten (Vögel und Meeressäuger) trainiert und anschließend mit Bildern eines unabhängigen Testbildsatzes konfrontiert, die 111 666 verifizierte biologische Objekte zeigten. Obgleich der Testbildsatz mit mehr als 120 Arten/Artengruppen fast doppelt so viele Taxa aufwies wie der Trainingsbildsatz, fand HiDeFIND insgesamt 96.5 % aller Objekte. In gemischten Modellen, die die hierarchische Abhängigkeitsstruktur des Datensatzes berücksichtigten, erzielte das Modell eine hohe globale mittlere Sensitivität („recall" im maschinellen Lernen) von >99 %. Das Modell entdeckte Seevögel und Meeressäugetiere auf Standbildern digitaler Videoaufnahmen demnach fast genauso gut wie speziell geschulte menschliche Beobachter:innen. Dies schloss die Detektion vieler Schlüsselarten der maritimen Raum- und Umweltplanung wie Sterntaucher *Gavia stellata*, Trottellumme *Uria aalge* oder Dreizehenmöwe *Rissa tridactyla* und unter den Meeressäugern den Schweinswal *Phocoena phocoena* ein (alle mit mittleren Sensitivitäten von >99 %). Die erzielte Sensitivität war unabhängig von der Jahreszeit und weitgehend unabhängig von detektionsrelevanter Umweltvariation mit Ausnahme starker Sonnenreflexion. Diese minderte die mittlere Sensitivität, jedoch mit sehr begrenzter Wirkung für die Gesamtleistung, da stärkere Sonnenreflexion in aller Regel bereits effektiv durch technische Maßnahmen vermieden worden war. Die hohe Sensitivität ging mit einer hohen Rate falsch positiver Detektionen einher, die ebenfalls verstärkt bei stärkerer Sonnenreflexion auftraten. In einem KI-unterstützen Arbeitsprozess wird die daraus resultierende niedrige Präzision ein manuelles Herausfiltern falsch positiver Detektionen vor Übergabe an die sich anschließende Artbestimmung erfordern. Derzeit mindert dies noch die Effizienz des Modells und damit die erzielbare Zeitersparnis, nicht jedoch die dank ausgezeichneter Sensitivität hohe HiDeFIND-Ergebnisqualität. Für die Auswertung digitaler Flugerfassungsdaten in der maritimen Raum- und Umweltplanung, im Umweltmonitoring sowie in der Forschung bietet HiDeFIND damit eine zukunftsweisende Alternative zur manuellen Objektdetektion. HiDeFIND operiert dabei im Rahmen eines integrierten „human-in-the-loop" Arbeitsprozesses, bei dem eine automatisierte Objektdetektion durch KI von einem Qualitätssicherungsprozess durch geschulte Beobachter:innen flankiert wird.

6 Literatur

Borowicz A, Le H, Humphries G, Nehls G, Höschle C, Kosarev V, Lynch HJ 2019: Aerial-trained „Deep Learning" networks for surveying cetaceans from satellite imagery. PLOS ONE 14: e0212532.

Borowiec ML, Dikow RB, Frandsen PB, McKeeken A, Valentini G, White AE 2022: „„Deep Learning"" as a tool for ecology and evolution. Methods in Ecology and Evolution 13: 1640–1660.

Boulent J, Charry B, Kennedy MM, Tissier E, Fan R, Marcoux M, Watt CA, Gagné-Turcotte A 2023: Scaling whale monitoring using „„Deep Learning"": A human-in-the-loop

solution for analyzing aerial datasets. Frontiers in Marine Science 10: 1099479.
BSH 2013: Standard - Untersuchung der Auswirkungen von Offshore-Windenergieanlagen auf die Meeresumwelt (StUK 4).
Buckland ST, Burt ML, Rexstad EA, Mellor M, Williams AE, Woodward R 2012: Aerial surveys of seabirds: the advent of digital methods. Journal of Applied Ecology 49: 960–967.
Burnell R, Schellaert W, Burden J, Ullman TD, Martinez-Plumed F, Tenenbaum JB, Rutar D, Cheke LG, Sohl-Dickstein J, Mitchell M, Kiela D, Shanahan M, Voorhees EM, Cohn AG, Leibo JZ, Hernandez-Orallo J 2023: Rethink reporting of evaluation results in AI. Science 380: 136–138.
Converse RL, Lippitt CD, Koneff MD, White TP, Weinstein BG, Gibbons R, Stewart DR, Fleishman AB, Butler MJ, Sesnie SE, Harris GM 2024: Remote sensing and machine learning to improve aerial wildlife population surveys. Frontiers in Conservation Science 5: 1416706.
Dierschke V, Borkenhagen K, Enners L, Garthe S, Mercker M, Peschko V, Schwemmer H, Markones N 2024: Sensitivität von Seevögeln gegenüber Offshore-Windparks in der deutschen Nordsee im Hinblick auf Lebensraumverluste durch Meidung. Vogelwelt 142: 59–74.
Ditria EM, Buelow CA, Gonzalez-Rivero M, Connolly RM 2022: Artificial intelligence and automated monitoring for assisting conservation of marine ecosystems: A perspective. Frontiers in Marine Science 9: 918104.
Dorsch M, Schmoll T, Nehls G 2024: Zehn Jahre digitale Flugerfassung von Seevögeln und Meeressäugern. Die HiDef-Methode. Seevögel 45: 14–17.
Dujon AM, Ierodiaconou D, Geeson JJ, Arnould JPY, Allan BM, Katselidis KA, Schofield G 2021: Machine learning to detect marine animals in UAV imagery: effect of morphology, spacing, behaviour and habitat. Remote Sensing in Ecology and Conservation 7: 341–354.
Fliessbach KL, Borkenhagen K, Guse N, Markones N, Schwemmer P, Garthe S 2019: A Ship Traffic Disturbance Vulnerability Index for Northwest European Seabirds as a Tool for Marine Spatial Planning. Frontiers in Marine Science 6: 192.
Forstmeier W, Wagenmakers E, Parker TH 2017: Detecting and avoiding likely false-positive findings – a practical guide. Biological Reviews 92: 1941–1968.
Frainer G, Dufourq E, Fearey J, Dines S, Probert R, Elwen S, Gridley T 2023: Automatic detection and taxonomic identification of dolphin vocalisations using convolutional neural networks for passive acoustic monitoring. Ecological Informatics 78: 102291.
Furness RW, Wade HM, Masden EA 2013: Assessing vulnerability of marine bird populations to offshore wind farms. Journal of Environmental Management 119: 56–66.
Greener JG, Kandathil SM, Moffat L, Jones DT 2022: A guide to machine learning for biologists. Nature Reviews Molecular Cell Biology 23: 40–55.
Guirado E, Tabik S, Rivas ML, Alcaraz-Segura D, Herrera F 2019: Whale counting in satellite and aerial images with „Deep Learning". Scientific Reports 9: 14259.
Hurlbert SH 1984: Pseudoreplication and the Design of Ecological Field Experiments. Ecological Monographs 54: 187–211.
Ke T-W, Yu SX, Koneff MD, Fronczak DL, Fara LJ, Harrison TJ, Landolt KL, Hlavacek EJ, Lubinski BR, White TP 2024: „Deep Learning" workflow to support in-flight processing of digital aerial imagery for wildlife population surveys. PLOS ONE 19: e0288121.
Kellenberger B, Veen T, Folmer E, Tuia D 2021: 21 000 birds in 4.5 h: efficient large-scale seabird detection with machine learning. Remote Sensing in Ecology and Conservation 7: 445–460.
Kuhn HW 1955: The Hungarian method for the assignment problem. Naval Research Logistics Quarterly 2: 83–97.
Kuru K, Clough S, Ansell D, McCarthy J, McGovern S 2023: WILDetect: An intelligent platform to perform airborne wildlife census automatically in the marine ecosystem using an ensemble of learning techniques and computer vision. Expert Systems with Applications 231: 120574.
Lenzi J, Barnas AF, ElSaid AA, Desell T, Rockwell RF, Ellis-Felege SN 2023: Artificial intelligence for automated detection of large mammals creates path to upscale drone surveys. Scientific Reports 13: 947.
Li J, Xu W, Deng L, Xiao Y, Han Z, Zheng H 2023: „Deep Learning" for visual recognition and detection of aquatic animals: A review. Reviews in Aquaculture 15: 409–433.
Marchowski D 2021: Drones, automatic counting tools, and artificial neural networks in wildlife population censusing. Ecology and Evolution 11: 16214–16227.
Mcilwaine B, Rivas Casado M 2021: JellyNet: The convolutional neural network jellyfish bloom detector. International Journal of Applied Earth Observation and Geoinformation 97: 102279.
Miao Z, Yu SX, Landolt KL, Koneff MD, White TP, Fara LJ, Hlavacek EJ, Pickens BA, Harrison TJ, Getz WM 2023: Challenges and solutions for automated avian recognition in aerial imagery. Remote Sensing in Ecology and Conservation 9: 439–453.
Nakagawa S, Lagisz M, Francis R, Tam J, Li X, Elphinstone A, Jordan NR, O'Brien JK, Pitcher BJ, Van Sluys M, Sowmya A, Kingsford RT 2023: Rapid literature mapping on the recent use of machine learning for wildlife imagery. Peer Community Journal 3: e35.
Saeed W, Omlin C 2023: Explainable AI (XAI): A systematic meta-survey of current challenges and future opportunities. Knowledge-Based Systems 263: 110273.
Sokolova M, Lapalme G 2009: A systematic analysis of performance measures for classification tasks. Information Processing & Management 45: 427–437.
Tabak MA, Norouzzadeh MS, Wolfson DW, Sweeney SJ, Vercauteren KC, Snow NP, Halseth JM, Di Salvo PA, Lewis JS, White MD, Teton B, Beasley JC, Schlichting PE, Boughton RK, Wight B, Newkirk ES, Ivan JS, Odell EA, Brook RK, Lukacs PM, Moeller AK, Mandeville EG, Clune J, Miller RS 2019: Machine learning to classify animal species in camera trap images: Applications in ecology. Methods in Ecology and Evolution 10: 585–590.
Torney CJ, Lloyd-Jones DJ, Chevallier M, Moyer DC, Maliti HT, Mwita M, Kohi EM, Hopcraft GC 2019: A comparison of „Deep Learning" and citizen science techniques for counting wildlife in aerial survey images. Methods in Ecology and Evolution 10: 779–787.
Tuia D, Kellenberger B, Beery S, Costelloe BR, Zuffi S, Risse B, Mathis A, Mathis MW, van Langevelde F, Burghardt T, Kays R, Klinck H, Wikelski M, Couzin ID, van Horn G, Crofoot MC, Stewart CV, Berger-Wolf T 2022: Perspectives in machine learning for wildlife conservation. Nature Communications 13: 792.

Weiser EL, Flint PL, Marks DK, Shults BS, Wilson HM, Thompson SJ, Fischer JB 2023: Optimizing surveys of fall-staging geese using aerial imagery and automated counting. Wildlife Society Bulletin 47: e1407.

Weiß F, Büttger H, Baer J, Welcker J, Nehls G 2016: Erfassung von Seevögeln und Meeressäugetieren mit dem HiDef Kamerasystem aus der Luft. Seevögel 37: 14–21.

Xu Z, Wang T, Skidmore AK, Lamprey R 2024: A review of „Deep Learning" techniques for detecting animals in aerial and satellite images. International Journal of Applied Earth Observation and Geoinformation 128: 103732.

Žydelis R, Dorsch M, Heinänen S, Nehls G, Weiss F 2019: Comparison of digital video surveys with visual aerial surveys for bird monitoring at sea. Journal of Ornithology 160: 567–580.

Glossar Fachbegriffe

Bounding Box: Im Computersehen grenzen Bounding Boxen Objekte von Interesse in Rechteckform möglichst eng ein. Auf Bildmaterial für Trainingszwecke werden Bilder mit Bounding Boxen annotiert, um dem Modell visuelle Muster der Zielobjekte als Lernbeispiele bereitzustellen. Bei Validierung und Test hingegen repräsentieren Bounding Boxen die vom Modell gemachten Vorhersagen und stellen somit das Ergebnis der automatisierten Objekterkennung dar.

Bodenauflösung *(„ground sampling distance")*: Reale Strecke auf der Erd- oder Meeresoberfläche, die dem Abstand der Zentren benachbarter Pixel im digitalen Bild entspricht.

Convolutional Neural Network (CNN): CNNs sind eine auf „Deep Learning" basierende Variante künstlicher neuronaler Netzwerke, die besonders gut für die Verarbeitung von Bilddaten geeignet ist. CNNs nutzen Faltungsschichten („convolutional layers"), um lokale Muster wie Kanten oder Formen zu erkennen und schrittweise zu komplexeren Strukturen zu kombinieren. CNNs lernen relevante Muster direkt aus ihren Trainingsdaten und sind zentraler Bestandteil moderner automatisierter Bildanalyseverfahren.

„Deep Learning": „Deep Learning" als Teildisziplin des maschinellen Lernens setzt künstliche neuronale Netzwerke mit zahlreichen tief gestaffelten Netzwerkschichten zwischen der Eingabeschicht (z. B. RGB-Werte der Pixel eines Digitalfotos) und der Ausgabeschicht (z. B. Foto enthält Seevogel ja/nein) ein. Dies ermöglicht Computern hocheffizientes autonomes Lernen aus Beispielen. Maschinelles Lernen ist wiederum ein Teilgebiet der künstlichen Intelligenz.

Explainable Artificial Intelligence (XAI): Die Entscheidungen künstlicher neuronaler Netzwerke sind im Einzelfall oft schwer nachvollziehbar („Black Box"). Mit diesem Problem beschäftigt sich eine eigene Unterdisziplin des „Deep Learning", die sogenannte erklärbare künstliche Intelligenz („Explainable Artificial Intelligence"; Übersicht z. B. in Saeed & Omlin 2023). Während in sensiblen Anwendungen wie automatisierter Kreditvergabe hohe Anforderungen an einer Nachvollziehbarkeit des Entscheidungsprozesses bestehen, wird dieser Aspekt im Computersehen generell als weniger entscheidend betrachtet. Die Evaluierung von Modellen im Computersehen ist entsprechend stark ergebnis- und weniger prozessorientiert.

Feste Effekte: Schätzer fester Effekte in statistischen Modellen repräsentieren in der Regel Differenzen von Mittelwerten (für Faktoren) oder Steigungen (für Kovariaten).

Gradientenverfahren *(„gradient descent")*: Im maschinellen Lernen häufig genutzter Optimierungsalgorithmus, um effizient lokale Minima einer Verlustfunktion zu bestimmen. In unserer Anwendung sind dies Minima der Abweichung zwischen tatsächlicher Objektposition und der vom Modell vorhergesagten Objektposition in digitalen Bildern. Die unterliegende Heuristik der Gradientenmethode entspricht dem sogenannten "Bergsteigeralgorithmus mit negativem Vorzeichen": Ein Bergsteiger in dichtem Nebel wird den Gipfel – ein lokales Maximum – auf dem kürzesten Weg erklimmen, wenn er zu jedem Zeitpunkt des Aufstiegs die Route des steilsten Anstiegs wählt. Analog dazu bewegt sich der Optimierungsalgorithmus entlang des steilsten Gradientenabfalls.

Intersection over Union (IoU): Eine wichtige Kenngröße im Computersehen, die den Grad der Überlappung der aus dem Feldvergleich abgeleiteten Bounding Box („ground truth") mit der vom Objekterkennungsmodell vorhergesagten Bounding Box angibt (in Prozent). Je nach geforderter Präzision der Verortung können Schwellenwerte festgelegt werden, ab denen eine Modellvorhersage als korrekt (richtig positiv) bewertet wird.

Präzision: Anteil der richtig positiven Vorhersagen an allen positiven Modellvorhersagen (Summe richtig positiver und falsch positiver Vorhersagen); im maschinellen Lernen als „precision" oder „positive predictive value" bezeichnet.

Sensitivität: Anteil der richtig positiven Vorhersagen an allen tatsächlich positiven Fällen (Summe richtig positiver und falsch negativer Vorhersagen); im maschinellen Lernen als „recall" oder „true positive rate" bezeichnet.

You Only Look Once (YOLO): Frühe Objekterkennungsmodelle basierten auf einem rechenintensiven, zweistufigen Prozess: Zunächst wurden Regionen mit potenziellen Objekten lokalisiert, anschließend in einem zweiten Schritt Objekte klassifiziert (ist Zielobjekt ja/nein). YOLO-Modelle hingegen leisten beide Schritte simultan bei nur einmaliger Präsentation eines Bildes – eine Innovation, die die automatisierte Objekterkennung wesentlich vereinfacht und beschleunigt hat.

Zufallseffekte: Schätzer von Zufallseffekten in statistischen Modellen repräsentieren in der Regel Varianzen, die durch Unterschiede zwischen Gruppen abhängiger Daten bedingt sind. Dabei gilt u. a. die Annahme, dass die im Datensatz vertretenen Gruppen/Stufen eine zufällige Stichprobe aus einer theoretisch unbegrenzten Menge möglicher Gruppen/Stufen dieses Effekts darstellen.

Anhang 1: Beschreibung detektionsrelevanter Umweltbedingungen. – *Appendix 1: Description of detection-relevant environmental conditions.*

Umweltbedingung – ***Environmental condition***	**Beschreibung** – ***Description***	**Skalenniveau** – ***Scale***	**Erfasst pro** – ***Assessed per***	**Fehlerverteilung (abhängige Variable)** – ***Errors (dependent variable)***	**Art des Prädiktors (unabhängige Variable)** – ***Predictor type (independent variable)***
Sonnenreflexion – *Glare*	Grad der Sonnenreflexion auf der Meeresoberfläche. – *Degree of sun reflection on the sea surface.*	Ordinal: Keine, gering, mäßig, stark. – *Ordinal: None, slight, moderate, strong.*	[1]Standbild. – *Frame.*	Normal – *Gaussian.*	Faktor (fester Effekt) mit vier Stufen. – *Fixed effect four-level factor.*
Seegang – *Sea state*	Zustand der freien Meeresoberfläche, erzeugt durch Dünung und Windsee. – *State of the open sea surface produced by both swell and wind sea.*	[2]Ordinal: Spiegelglatt, ruhig, schwach bewegt, leicht bewegt, mäßig bewegt, grob. – *Ordinal: Smooth, calm, weakly moved, slightly moved, moderately moved, rough.*	[3]Reihe horizontal benachbarter Standbilder über die acht Videosequenzen. – *Row of horizontally adjacent frames across the eight reels.*	Normal – *Gaussian.*	Faktor (fester Effekt) mit sechs Stufen. – *Fixed effect six-level factor.*
Lufttrübung – *Air turbidity*	Grad der Trübung des Luftkörpers zwischen Flugzeug und Meeresoberfläche. – *Degree of turbidity of the air body between aircraft and sea surface.*	Ordinal: Keine, gering, mäßig bis stark. – *Ordinal: None, slight, moderate to strong.*	[3]Reihe horizontal benachbarter Standbilder über die acht Videosequenzen. – *Row of horizontally adjacent frames across the eight reels.*	[4]Binomial – *Binomial.*	[4]Faktor (fester Effekt) mit zwei Stufen (Lufttrübung *versus* keine). – *Fixed effect two-level factor (air turbidity versus none).*
[5]Wassertrübung – *Water turbidity*	Grad der durch Schwebstoffe verursachten Trübung des Meerwassers. – *Degree of turbidity of seawater caused by suspended matter.*	Ordinal: Keine, gering, mäßig bis stark. – *Ordinal: None, slight, moderate to strong.*	[3]Reihe horizontal benachbarter Standbilder über die acht Videosequenzen. *Row of horizontally adjacent frames across the eight reels.*	[6]Binomial – *Binomial.*	[6]Faktor (fester Effekt) mit zwei Stufen (Wassertrübung *versus* keine). – *Fixed effect two-level factor (water turbidity versus none).*

[1]Geschätzt auf jeder der acht Videosequenzen eines Transekts. – *Estimated from each of the eight reels of a transect.*

[2]Entspricht unterem Teil der Petersen-Skala zur Klassifikation des Seegangs. Aufnahmen bei mehr als grober See müssen laut behördlicher Untersuchungsstandards verworfen werden. – *Corresponds to lower part of the Petersen sea state scale. Footage taken in more than rough seas must be discarded according to regulatory guidelines.*

[3]Geschätzt nur auf Videosequenzen der virtuellen Kamera #2. – *Estimated only from the reels captured by virtual camera #2.*

[4]Lufttrübung „mäßig bis stark" war sehr selten (0.1% und 0.09% der Beobachtungen im Trainings- und Testbildsatz) und wurde für die Analysen mit „gering" zusammengelegt. – *Air turbidity "moderate to strong" was very rare (0.1% and 0.09% of observations in the training and test data set, respectively) and merged with "slight" for analysis.*

[5]Relevant für (teilweise) untergetauchte Meeressäuger. – *Relevant for (partially) submerged marine mammals.*

[6]Wassertrübung „mäßig bis stark" war sehr selten (0.7% und 0.65% der Beobachtungen im Trainings- und Testbildsatz) und wurde für die Analysen mit „gering" zusammengelegt. – *Water turbidity "moderate to strong" was very rare (0.7% and 0.65% of observations in the training and test data set, respectively) and merged with "slight" for analysis.*

Anhang 2: Eigenschaften von Trainingsbildsatz und Testbildsatz. – *Appendix 2: Attributes of training and test image set.*

Eigenschaft – ***Attribute***	**Trainingsbildsatz** – ***Training image set***	**Testbildsatz** – ***Test image set***
Jahre – *Years*	2017, 2021	2021, 2022
Summe Transekt-Kilometer – *Total transect kilometers*	~11 650	~8040
Analysierte/aufgezeichnete Fläche (km^2) – *Area analysed/observed (km^2)*	~6080/6210	~4180/4280
Analysiertes HiDef Video-Material (h) – *HiDef video footage analysed (h)*	~280	~250
Meeresgebiete – *Marine areas*	Nordsee, Ostsee	Nordsee, Ostsee, Englischer Kanal
Projektgebiete – *Project sites*	4	6
Erfassungsflüge – *Aerial surveys*	21[1]	13[1]
Transekte – *Transects*	291	196
Videosequenzen – *Reels*	1629	1408
Standbilder – *Frames*	79 134[2]	51 218[2]
Von Menschen (nicht vom Modell) markierte Objekte – *Objects marked by humans (not the model)*	138 681[3]	111 666
Von Menschen (nicht vom Modell) markierte individuelle Organismen – *Individual organisms marked by humans (not the model)*	26 635	111 666[4]
Von Menschen (nicht vom Modell) markierte Vogelobjekte – *Bird objects marked by humans (not the model)*	132 718	110 697
Von Menschen (nicht vom Modell) markierte Säuger-Objekte – *Mammal objects marked by humans (not the model)*	5846	703
Andere von Menschen (nicht vom Modell) markierte Objekte – *Other objects marked by humans (not the model)*	117[5]	266[5]
Arten-/Artengruppen – *Species/species group richness*	66	124
Gesamtdiversität – *Species/species group diversity*	4.28[6]	4.05[6]
Gesamtgleichverteilung – *Species/species group evenness*	0.71[7]	0.58[7]
Vogelarten – *Bird species richness*	37	73
Vogel-Artengruppen – *Bird species group richness*	20	36
Summe Vogeltaxa – *Bird taxa richness*	57	109
Diversität Vogelarten – *Bird species diversity*	3.64[6]	3.48[6]
Gleichverteilung Vogelarten – *Bird species evenness*	0.70[7]	0.56[7]

Anhang 2: Fortsetzung

Eigenschaft ***– Attribute***	**Trainingsbildsatz** ***– Training image set***	**Testbildsatz** ***– Test image set***
Diversität Vogel-Artengruppen *– Bird taxa diversity*	4.16[6]	3.99[6]
Gleichverteilung Vogel-Artengruppen *– Bird taxa evenness*	0.71[7]	0.59[7]
Kumulativer Prozentsatz fünf häufigster Vogeltaxa *– Cumulative percentage of five most abundant bird taxa*	53.8[8]	66.4[8]
Kumulativer Prozentsatz zehn häufigster Vogeltaxa *– Cumulative percentage of ten most abundant bird taxa*	78.3[8]	78.6[8]
Kumulativer Prozentsatz zwanzig häufigster Vogeltaxa *– Cumulative percentage of twenty most abundant bird taxa*	94.9[8]	88.8[8]
Säugetierarten *– Mammal species richness*	4	5
Säugetier-Artengruppen *– Mammal species group richness*	3	2
Summe Säugetier-Taxa *– Mammal taxa richness*	7	7
Diversität Säugetierarten *– Mammal species diversity*	0.18[6]	1.21[6]
Gleichverteilung Säugetierarten *– Mammal species evenness*	0.09[7]	0.52[7]
Diversität Säugetier-Artengruppen *– Mammal taxa diversity*	0.83[6]	1.80[6]
Gleichverteilung Säugetier-Artengruppen *– Mammal taxa evenness*	0.30[7]	0.64[7]
Gewichtetes Mittel Sonnenreflexion[9] (95% Konfidenz) *– Weighted mean glare[9] (95% confidence)*	0.61 (0.44-0.78)	0.20 (0.12-0.28)
Gewichtetes Mittel Seegang[9] (95% Konfidenz) *– Weighted mean sea state[9] (95% confidence)*	2.54 (2.26-2.82)	2.11 (1.82-2.40)
Gewichtete Wahrscheinlichkeit von Lufttrübung[9] (95% Konfidenz) *– Weighted mean probability of (some) air turbidity[9] (95% confidence)*	0.17 (0.10-0.27)	0.15 (0.08-0.23)
Gewichtete Wahrscheinlichkeit von Wassertrübung[9] (95% Konfidenz) *– Weighted mean probability of (some) water turbidity[9] (95% confidence)*	0.16 (0.09-0.26)	0.18 (0.11-0.28)

[1]Siehe Abbildung 2 im Text für jahreszeitliche Verteilungen. – *See figure 2 in main text for seasonal distributions.*

[2]Standbilder, auf denen Beobachter:innen mindestens ein biologisches Objekt entdeckt hatten. – *Frames, on which human observers had detected at least one biological object.*

[3]Im Trainingsbildsatz stellt ein auf mehreren Standbildern erfasstes Individuum ein Trainingsobjekt je Standbild dar (siehe Text 2.2.4.3). – *In the training image set, an individual present on multiple frames represents a separate training object on each occasion (see text 2.2.4.3).*

[4]Im Testbildsatz ist jedes Individuum nur einmalig vertreten (siehe Text 2.2.4.3). – *In the test image set, each individual is represented only once (see text 2.2.4.3).*

[5]Zum Beispiel Roter Thun, Mondfisch, Löwenmähnenqualle. – *For example Atlantic Bluefin Tuna, Ocean Sunfish, Lion's Mane Jellyfish.*

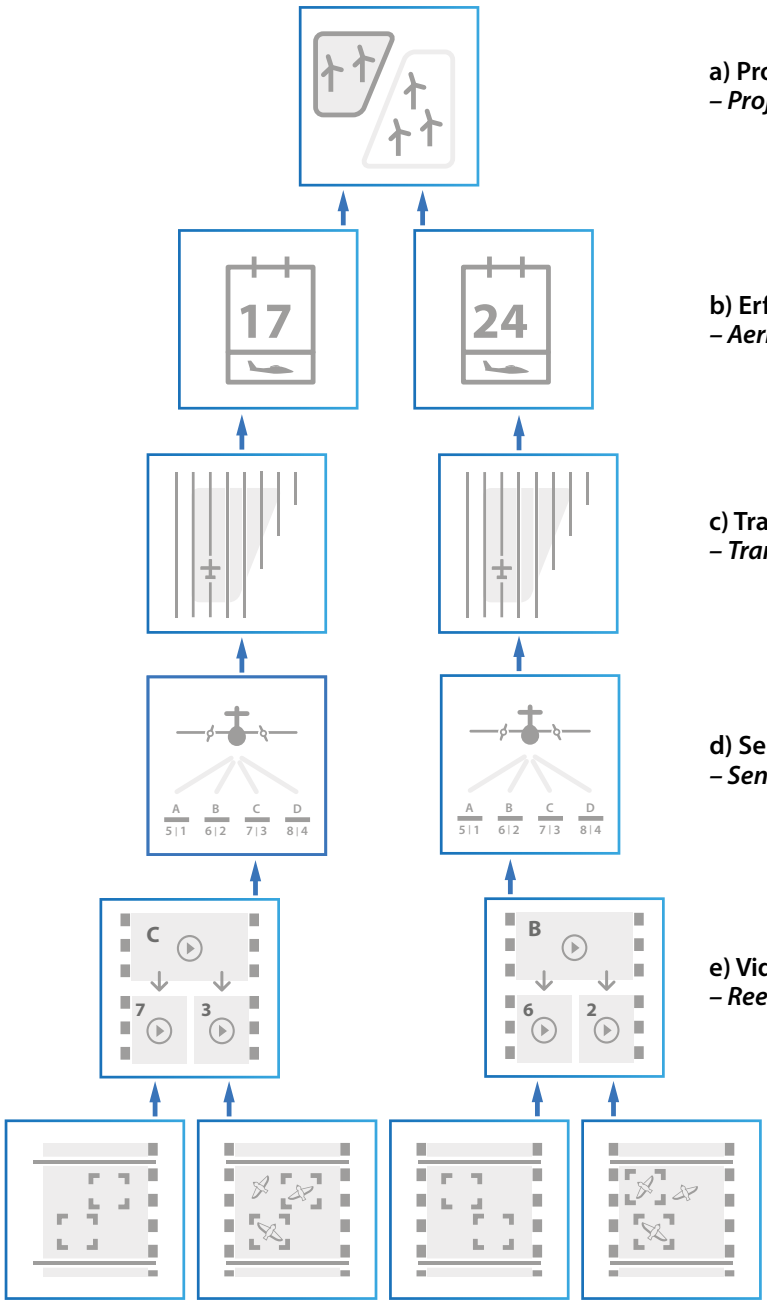
[6]Shannon-Index ($\log_2$). – *Shannon-Index ($\log_2$).*

[7]Shannon-Index/maximal möglicher Shannon-Index. – *Shannon-Index/maximum possible Shannon-Index.*

[8]Siehe Abbildung 3 im Text für Artidentitäten. – *See figure 3 in main text for species identities.*

[9]Siehe auch Anhang 1. – *See also electronic supplementary material 1.*

Anhang 3: Hierarchische Abhängigkeitsstruktur von HiDef-Daten – *Appendix 3: Hierarchical dependency structure of HiDef data.*



a) Projektgebiete (# variabel)
– Project sites (#variable)

b) Erfassungsflüge (# variabel)
– Aerial surveys (# variable)

c) Transekte (# variabel)
– Transects (# variable)

d) Sensoren A–D & virtuelle Kameras 1–8
– Sensors A–D & virtual cameras 1–8

e) Videosequenzen 1–8 (zwei je Sensor)
– Reels 1–8 (two per sensor)

f) Standbilder (# variabel)
– Frames (#variable)

Ursachen für die Nicht-Unabhängigkeit von Detektionswahrscheinlichkeiten können sich auf geteilte Attribute der Objekte selbst, geteilte Attribute ihrer Umwelt im Moment der Erfassung oder auf geteilte Attribute der sensorischen Maschinerie beziehen, die einen Bildsatz erzeugt hat. Die unten beispielhaft benannten möglichen Ursachen stellen einige von vielen plausiblen, sich nicht wechselseitig ausschließenden Ursachen dar. Im Kontext der Leistungsanalysen von HiDe-FIND können Detektionen richtig positive, falsch negative, aber auch falsch positive Modellvorhersagen repräsentieren. In f) sind beispielhaft richtig positive Vorhersagen des Modells durch stilisierte Bounding Boxen mit Zielobjekt dargestellt, falsch positive durch Bounding Boxen ohne Zielobjekt und falsch negative als bloßes Zielobjekt. –*Reasons for the non-independence*

of detection probabilities can refer to shared attributes of the objects themselves, shared attributes of their environment at the moment of acquisition, or shared attributes of the sensory machinery that generated an image set. The potential causes highlighted below in an exemplary fashion are some of many plausible causes that are not mutually exclusive. In the context of the performance analyses of HiDeFIND, detections can represent true positive, false negative, but also false positive model predictions. In f), true positive predictions of the model are represented by stylized bounding boxes with target object, false positive predictions by bounding boxes without target object, and false negative predictions by target objects only.

Abhängigkeiten in HiDef-Datensätzen könnten begründet sein in der: – ***Dependencies in HiDef data sets could be caused by the:***

- **Identität von Projektgebieten (1a)**. Die Detektionswahrscheinlichkeiten des Modells könnten zwischen Projektgebieten variieren, wenn sie artspezifisch sind und sich Projektgebiete im Artenspektrum unterscheiden; oder wenn sie tageszeitabhängig sind und verschiedene Gebiete systematisch zu unterschiedlichen Tageszeiten beflogen wurden (z. B. aufgrund unterschiedlicher Anflugdistanzen). – Identity of project sites (1a). *The model's detection probabilities could vary among project sites if they are species-specific, and project sites differ in species composition; or if they are dependent on the time of day and different areas were surveyed at systematically different times of the day (e.g. due to different approach distances).*

- **Identität von Erfassungsflügen (geschachtelt in Projektgebieten, 1b)**. Innerhalb von Projektgebieten könnten Detektionswahrscheinlichkeiten des Modells zwischen individuellen Erfassungsflügen variieren, wenn Objekte die hydrografischen Bedingungen oder das Wetter auf der Makroskala eines Erfassungsfluges teilen. – Identity of aerial surveys (nested in project sites, 1b). *Within project sites, model detection probabilities could vary among individual aerial surveys, when objects share the macroscale hydrographic or weather conditions of a given aerial survey.*

- **Identität von Transekten (geschachtelt in Erfassungsflügen, 1c)**. Innerhalb von Erfassungsflügen könnten Detektionswahrscheinlichkeiten des Modells zwischen individuellen Transekten variieren, wenn Objekte die hydrografischen Bedingungen von Transekten auf der Mesoskala teilen (z. B. Luv- oder Lee-Seiten um Inseln). – Identity of transects (nested in aerial surveys, 1c). *Within aerial surveys, model detection probabilities could vary among individual transects when objects share the mesoscale hydrographic conditions of transects on this (e.g., windward versus leeward sides around islands).*

- **Identität von Sensoren (geschachtelt in Transekten, 1d)**. Innerhalb von Transekten könnten die Detektionswahrscheinlichkeiten des Modells zwischen individuellen Sensoren variieren, wenn geringfügige Unterschiede bei der Produktion oder Konfiguration ihrer Komponenten existieren. – Identity of sensors (nested in transects, 1d). *Within transects, the model detection probabilities could vary among individual sensors if there are minor differences in the production or configuration of their components.*

- **Identität von Videosequenzen (geschachtelt in Sensoren, 1e)**. Innerhalb von Sensoren könnten die Detektionswahrscheinlichkeiten des Modells zwischen individuellen Videosequenzen variieren, wenn die Qualität des Referenzmaterials (Benchmark) von der Identität der menschlichen Beobachter:innen abhängt, die das Referenzmaterial produziert haben (Videomaterial wird auf Basis von Videosequenzen zufällig zugewiesen). – Identity of reels (nested in sensors, 1e). *Within sensors, model detection probabilities could vary among individual reels if the quality of the reference material (benchmark) depends on the identity of the human observers who produced the reference material (video footage is randomly assigned based on reels).*

- **Identität von Standbildern (geschachtelt in Videosequenzen, 1f)**. Innerhalb von Videosequenzen könnten die Detektionswahrscheinlichkeiten des Modells zwischen individuellen Standbildern variieren, wenn Objekte die hydrografischen Bedingungen oder das Wetter auf der Mikroskala eines bestimmten Standbilds teilen. – Identity of still images (nested in reels, 1f). *Within reels, model detection probabilities could vary among individual still images when objects share microscale hydrographic or weather conditions of a given image.*