High sensitivity of automated detection of seabirds at sea on digital aerial survey footage

Tim Schmoll, Guruprasad Hegde, Monika Dorsch & Georg Nehls

This publication is a translation of "Hohe Sensitivität automatischer Detektion von Seevögeln auf See auf digitalen Luftbildaufnahmen", originally published in Vogelwarte 63, 2025: 191-215 (https://www.do-g.de/veroeffentlichen/vogelwarte).

Abstract

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 from an independent test image set featuring more than 111,000 verified biological objects. Objects from 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, Remote Sensing, Schobüller Str. 36, 25813 Husum.
Email: t.schmoll@bioconsult-sh.de. ORCID: 0000-0003-3234-7335
MD: BioConsult SH GmbH & Co. KG, Remote Sensing, Schobüller Str. 36, 25813 Husum.
Email: m.dorsch@bioconsult-sh.de
GH: BioConsult SH GmbH & Co. KG, Remote Sensing, Schobüller Str. 36, 25813 Husum.
Email: g.hegde@bioconsult-sh.de
GN: BioConsult SH GmbH & Co. KG, Schobüller Str. 36, 25813 Husum. Email: g.nehls@bioconsult-sh.de.
ORCID: 0009-0000-6424-1989

1 Introduction

The reliable recording of seabirds at sea is of great importance for maritime spatial and environmental planning, as well as for assessing seabird population dynamics within the framework of national and international monitoring programs. Since the turn of the millennium, aircraft-based surveys and, for more than ten years, digital aerial transect surveys (Buckland et al. 2012) have been the standard for seabird and marine mammal surveys in many countries. In Germany, this is specified in the *Standard Investigation of the Impacts of Offshore Wind Turbines on the Marine Environment (StUK 4)* (BSH 2013). A recognised and tested method is the HiDef digital aerial survey video recording method (Weiß et al. 2016, Žydelis et al. 2019), which has been applied internationally for over ten years in more than 3,000 survey flights (Dorsch et al. 2024).

Artificial intelligence (AI) in the form of machine learning and, in particular, deep learning has recently become increasingly important in most areas of biology (overview in Greener et al. 2022). This also applies

to ecology in general (overviews in Borowiec et al. 2022, Ditria et al. 2022) and the monitoring of animal populations in particular (overviews in Tuia et al. 2022, Nakagawa et al. 2023, Xu et al. 2024). With regard to the latter, automated image detection has proven to be very useful both in terrestrial (e.g. Tabak et al. 2019) and aquatic environments (e.g. Li et al. 2023). For example, machine learning has been successfully used to detect jellyfish blooms (Mcilwaine & Rivas Casado 2021) and to estimate the abundance of terrestrial mammals (Torney et al. 2019, Lenzi et al. 2023) or the abundance of (breeding) waterbirds and seabirds on drone images (Dujon et al. 2021, Kellenberger et al. 2021, Marchowski 2021). In addition, machine learning also supported, for example, the estimation of whale and dolphin populations on aircraft-based aerial images (Boulent et al. 2023), on satellite images (Borowicz et al. 2019, Guirado et al. 2019) or by means of underwater acoustic detection (Frainer et al. 2023).

While machine learning in combination with drones or satellites has already been successfully used in a wide variety of contexts, there has been little published work on the application of machine learning to aircraft-based surveys of seabirds at sea. A study by Kuru et al. (2023) describes a method for the automated detection of Northern Gannets *Morus bassanus* in aerial photographs. However, this approach was not based on the latest deep learning techniques. In addition, Ke et al. (2024) describe a deep learning detection model as part of a fully automated work process that is supposed to enable population estimates of seabirds at sea during flight and in near real time in the future.

Given the track record of machine learning in estimating animal population abundance, its application to the detection of birds and marine mammals in HiDef video footage is promising. At least four advantages over manual object detection are conceivable. First, automated object detection is expected to be faster than the manual process. Results could then be incorporated into planning decisions more quickly, for example. Second, the resulting higher cost-effectiveness could allow for more frequent and/or extensive surveys, enabling a more detailed recording of the marine environment in terms of both time and space. Third, it is to be expected that the results of automated object detection are more reproducible than the results of manual object detection, which would increase the transparency of planning decisions derived from them. Fourth, automated object detection potentially improves the quality of population estimates of seabirds at sea and marine mammals. However, the interplay of pronounced biological variation (e.g. species, sexual dimorphism, age classes, behaviour, body posture) on the one hand and pronounced environmental variation (e.g. light conditions, sea state, air and water turbidity) on the other makes reliable automated object detection in digital aerial imagery challenging (Miao et al. 2023, Xu et al. 2024). Whether an AI-supported approach is equal to or possibly superior to trained observers therefore requires an evaluation of the performance of each individual object detection model. As long as regulatory authorities do not require proof of high-quality results and no general quality standards have been established for the use of AI in population estimation using remote sensing (Converse et al. 2024), there is a risk that planning decisions or assessments of population dynamics will be made on the basis of unreliable data.

In this article, we first describe the object detection model HiDeFIND developed by us. It is a deep learning-based artificial neural network for detecting seabirds and marine mammals in digital HiDef video material. We document the architecture, training and testing of the model, as well as the extensive HiDef image sets used to train and test HiDeFIND. We then analyse the model's performance in detail and examine its dependence on factors such as species identity, season and detection-relevant environmental conditions. We show that HiDeFIND detects birds at sea and marine mammals almost as well as trained observers throughout the year and under a variety of detection-relevant environmental conditions. Lastly, we discuss the suitability of the system for use in maritime spatial and environmental planning and seabird monitoring.

2 Materials and methods

The development of the HiDeFIND object detection model (version 1.0) was based on digital video recordings obtained using the HiDef method. Object detections from the established manual work process served as the benchmark against which we compared the performance of HiDeFIND. Accordingly, the objects marked by human observers represent the ground truth and not the actual number of potentially detectable objects, which could only be determined in a field comparison. In the following we briefly describe the HiDef standard workflow (for details, see Weiß et al. 2016, Žydelis et al. 2019), followed by the documentation of the architecture, training and testing procedures of HiDeFIND itself.

2.1 The HiDef standard workflow

2.1.1 Data collection

Twin-engine high-wing propeller aircraft (e.g. Vulcanair P 68) were used for digital survey flights at an altitude of approximately 500 m. The aircraft were equipped with sensor systems consisting of four high-resolution digital video cameras (Figure 1a). At the sea surface, these achieved an average ground sampling distance of approximately 2 cm at a frame rate of seven frames per second. Due to a slightly greater distance between the lens and the sea surface, the inner cameras had a slightly higher ground resolution than the

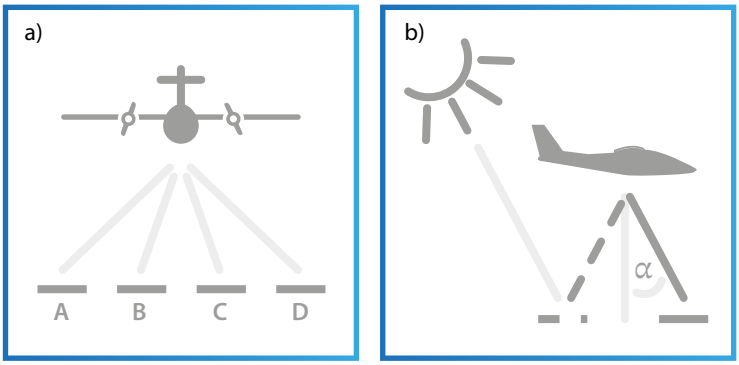


Figure 1: Schematic depiction of the HiDef sensor system in a) frontal view and b) lateral view.

outer cameras. The sensor system was not rigidly aligned perpendicular to the sea surface but inclined by 30° in or against the direction of flight, depending on the course and position of the sun (Figure 1b). This served to avoid interfering sun reflection (glare) on the sea surface, which may impact the detection and identification of target objects. The outer sensors A and D each covered a strip 143 m wide, while the inner sensors B and C each covered a strip 129 m wide. To avoid double counting, gaps of approximately 20 m were maintained between the strips. This resulted in an effective swath width of 544 m, distributed over a total strip width of approximately 604 m.

The aircraft flew at an average speed of approximately 220 km/h (120 knots). A GPS device recorded the position every second, enabling the georeferencing of the objects detected. The collected data was stored on mobile hard drives for later analysis.

2.1.2 Data processing

The video files were processed using StreamPix image capture and management software (NorPix, Montreal, Canada), and the files recorded by each camera were split for more convenient processing. The four cameras thus accommodate a total of eight virtual cameras, which deliver eight video sequences (reels) per transect. For data analysis, trained staff first examined still images (frames), and all recognised objects were digitally marked with points in StreamPix and preclassified for later object identification (e.g. possible bird, mammal, animal or man-made object). Reels or parts thereof that could not be reliably evaluated due to glare or clouds were marked and not included in subsequent analysis steps (in future work processes, an analysis of the recording conditions will first be carried out before reliably analysable parts of the video material are assigned to AI-supported object detection). To ensure consistently high quality, a randomly selected 20% of the video material was processed independently by two observers (without knowledge of each other's results). In addition, the detection-relevant environmental conditions glare, sea state, air turbidity and water turbidity were recorded (see Appendix 1). For sea state, air turbidity and water turbidity, it was assumed that these were identical at the time of recording for the entire swath covered by the four cameras and thus also for the eight reels of each transect resulting from splitting. These environmental conditions were therefore evaluated frame by frame on only a single reel per transect. However, glare can also vary between reels of a transect depending on the position of the sun and the course. It was therefore assessed frame by frame on each of the eight reels of each transect.

In a second step, marked objects were identified at the most precise taxonomic level possible, usually at species level, by experienced staff members skilled in ornithology and marine mammal identification. If species level identification was not possible due to the risk of confusion between very similar species (e.g. Common Tern *Sterna hirundo* and Arctic Tern *S. paradisaea*), these objects were assigned to groups of similar species (e.g. species group Common/Arctic tern). In addition, where possible, sex and age as well as behaviour (e.g. swimming or flying), association (e.g. with individuals of the same or other species) and, where applicable, flight direction were recorded. Furthermore, for quality control purposes, 20% of the marked objects were identified independently by a second person (without knowledge of the other person's results). Any discrepancies between the first and second identification processes were double-checked by a third person and corrected if necessary. Only if there was at least 90% agreement between the two identification processes the data was released for further analysis. If the agreement was less than 90%, systematic errors such as re-occurring misidentifications within certain species groups were discussed and all objects on the affected video material were re-identified.

2.2 Development of a neural network for object detection: HiDeFIND

The development of an artificial neural network model for HiDef object detection comprised the following five main steps:

1. **Base model:** Selection of a base model whose architecture best met our specific requirements.
2. **Data annotation:** Selection and annotation of a training image set that was as comprehensive and diverse as possible.
3. **Training:** During the training process, we exposed the model to various hyperparameter settings and trained and optimised it in a recursive process using the annotated training images.

4. **Validation:** We then tested the model on a small validation image set and repeatedly evaluated its current performance throughout the training process.
5. **Testing:** Finally, we conducted a comprehensive performance test using a large, independent and heterogeneous set of test images.

2.2.1 Model selection and specification

After considering various suitable model architectures, we selected a single-stage object detection algorithm from the You Only Look Once (YOLO) family as our base model, which is widely used in computer vision. Supplemented by user-defined network layers, the resulting artificial neural network HiDeFIND (a *convolutional neural network, CNN*) consists of more than 400 layers and more than 86 million parameters. The input layer processes digital images as a vector of the RGB values of all their pixels, and the output layer delivers bounding boxes generated by HiDeFIND that contain target objects with a specified probability. HiDeFIND performs classification along a single category (semanticised): "Is it a biological object? Yes/No". The main goal of the development was to achieve a detection accuracy that is at least as good as that of the established manual process to ensure the highest standards of result quality. HiDeFIND was therefore configured to mark visual patterns as objects of interest in cases of doubt rather than discarding them, i.e. it generally prioritises sensitivity (*recall* in machine learning) at the expense of precision (see 2.2.5.3).

2.2.2 Object tracking across images

HiDef video material produces time-oriented, spatially overlapping sequences of individual images (frames). Each object on HiDef video material therefore usually appears on more than one frame (depending on its position in the swath and the flight altitude of birds, on up to eight frames). Accordingly, HiDeFIND usually addresses each object more than once (this applies in particular to true positive detections). In order to control and automatically filter out unwanted multiple detections of the same biological object, we have developed an auxiliary algorithm for object tracking based on the Kuhn-Munkres algorithm (*Hungarian matching algorithm*, Kuhn 1955). This ensured that the detection(s) of a tracked biological object were only classified as true positives on a single frame (the frame on which human observers had marked the object). Further detections of the same tracked biological object on other frames were classified as false positives (see 2.2.5.2).

2.2.3 Origin of the image material

To create the training and test image sets, we used the joint archive of digital HiDef video material from BioConsult SH GmbH & Co KG (Husum, Germany, https://www.bioconsult-sh.de), HiDef Aerial Surveying Ltd. (Workington, United Kingdom, https://www.hidefsurveying.co.uk) and Biotope (Mèze, France, https://www.biotope.fr). A detailed characterisation of the image sets used is provided below.

2.2.4 Training

2.2.4.1 Annotation of the training image set

The availability of a large and sufficiently diverse training image set, which includes suitable images with annotations of the training objects in the form of bounding boxes, is crucial for the successful training of an object detection model. In the archive material used, observers had located the objects they detected by placing a digital point marker in the centre of the object. To be able to reuse the existing digital point markers, we developed a customised digital tool that allowed users to import existing point markers onto HiDef material and use them as a basis for manual annotation with bounding boxes.

2.2.4.2 Training process

For training purposes, the model processed the annotated training images during dozens of so-called epochs. In each of these epochs, the model was confronted with the entire training material, and its predictions were represented by model-generated bounding boxes. Their centre coordinates, height and width were used in an iterative gradient descent process to minimise discrepancies between the actual location of the objects (stored as manually set bounding boxes in the training material) and the location predicted by the model (as bounding boxes in the model output) by systematically adjusting the model parameters. For each epoch, we tracked the model's learning progress using a validation image set (just under 35,000 images) that did not overlap with the training image set (over 138,000 images, see 2.2.4.3) or the test image set (over 111,000 images, see 2.2.5.1).

2.2.4.3 Training image set

We trained the model with 138,681 annotated objects from 21 survey flights conducted at different times of the year in four project areas in two marine regions (North Sea, Baltic Sea) (Figure 2a, Appendix 2). Individuals may be represented by more than one image in the training image set, as HiDef video footage typically captured the same objects in several consecutive frames (see 2.2.2), often with different exposures or wing positions, which was beneficial for training (in the test image set, the number of manually detected objects corresponded to the number of manually detected individuals). The training image set contained 66 species/species groups, including 57 bird taxa (see Appendix 2 for details). Common Guillemots (*Uria aalge*) and Razorbills (*Alca torda*), together with the species group Guillemot/Razorbill, accounted for 31% of the total number of objects. Harbour Porpoises *Phocoena phocoena* represented 85% of the marine mammal objects. Figure 3a shows the frequency of the 50 most common taxa in the training image set.

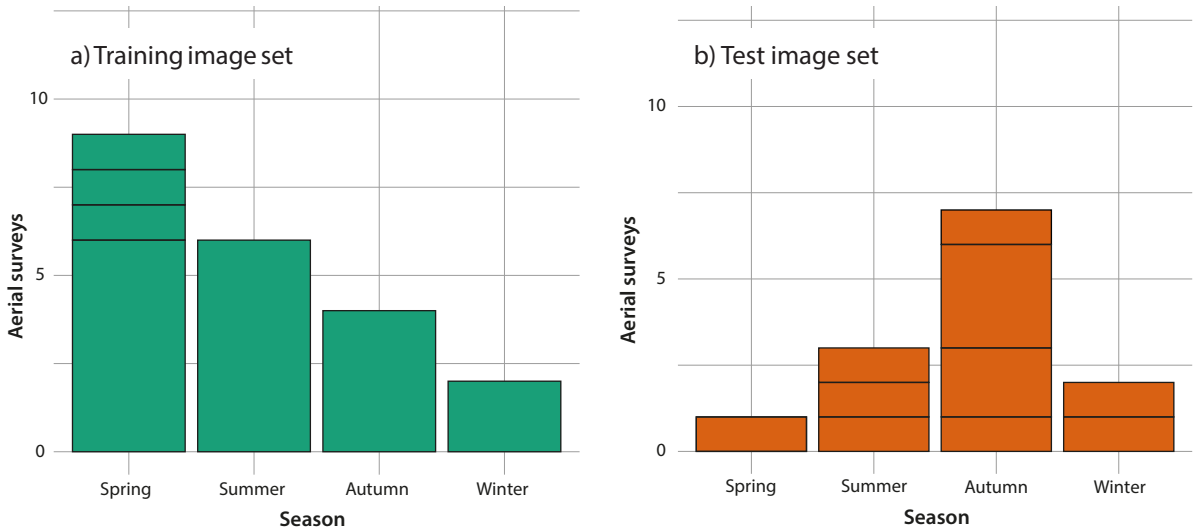


Figure 2: Aerial surveys by season for a) training image set and b) test image set (seasons meteorological). Stacks visualise contributions of different project sites.

2.2.5 Evaluation of model performance

We evaluated the performance of HiDeFIND using a test image set that did not overlap with the training image set or the validation image set.

2.2.5.1 Test image set

For the test image set, we selected survey flights that included different species compositions and a range of detection-relevant environmental variation, allowing us to test the model under a variety of realistic conditions. The test image set comprised 111,666 verified biological objects from 13 survey flights conducted at different times of the year in six project areas in three different marine regions (North Sea, Baltic Sea, English Channel) (Figure 2b, Appendix 2). The test image set contained 124 species/species groups, including 109 bird taxa (see Appendix 2 for details). Common Eiders *Somateria mollissima* and Common Scoters *Melanitta nigra* were the most common, accounting for 23% and 22% of the total number of objects, respectively, while Harbour Porpoises accounted for 58% of marine mammals. Figure 3b shows the frequency of the 50 most common taxa in the test image set. The objects had previously been detected during manual standard analyses of HiDef video material and identified at the species or species group level (see 2.1.2). The corresponding standard analyses were completed before the test design was drafted, so the observers involved could not have known that their performance would be used as the benchmark for evaluating HiDeFIND's performance.

2.2.5.2 Performance evaluation

The evaluation of model performance in classification tasks in general and in object detection in particular is usually based on a confusion matrix, in which actual events are compared with events predicted by the model (Sokolova & Lapalme 2009). The confusion matrix for the HiDeFIND performance evaluation can be specified as shown in Table 1. In addition to birds and marine mammals, we have included a few other representatives of marine megafauna that are regularly recorded in HiDef aerial surveys (e.g. Bluefin Tuna *Thunnus thynnus* or Sunfish *Mola mola*).

The evaluation of an object detection model usually requires annotations derived from field comparisons in the form of bounding boxes, which are then compared

Table 1: Confusion matrix mapping potential outcomes of the HiDeFIND performance evaluation.

		Benchmark (trained staff)	
		Is bird/mammal	Is not bird/mammal
Model prediction	Is bird/mammal	True positive	False positive
	Is not bird/mammal	False negative	Not defined

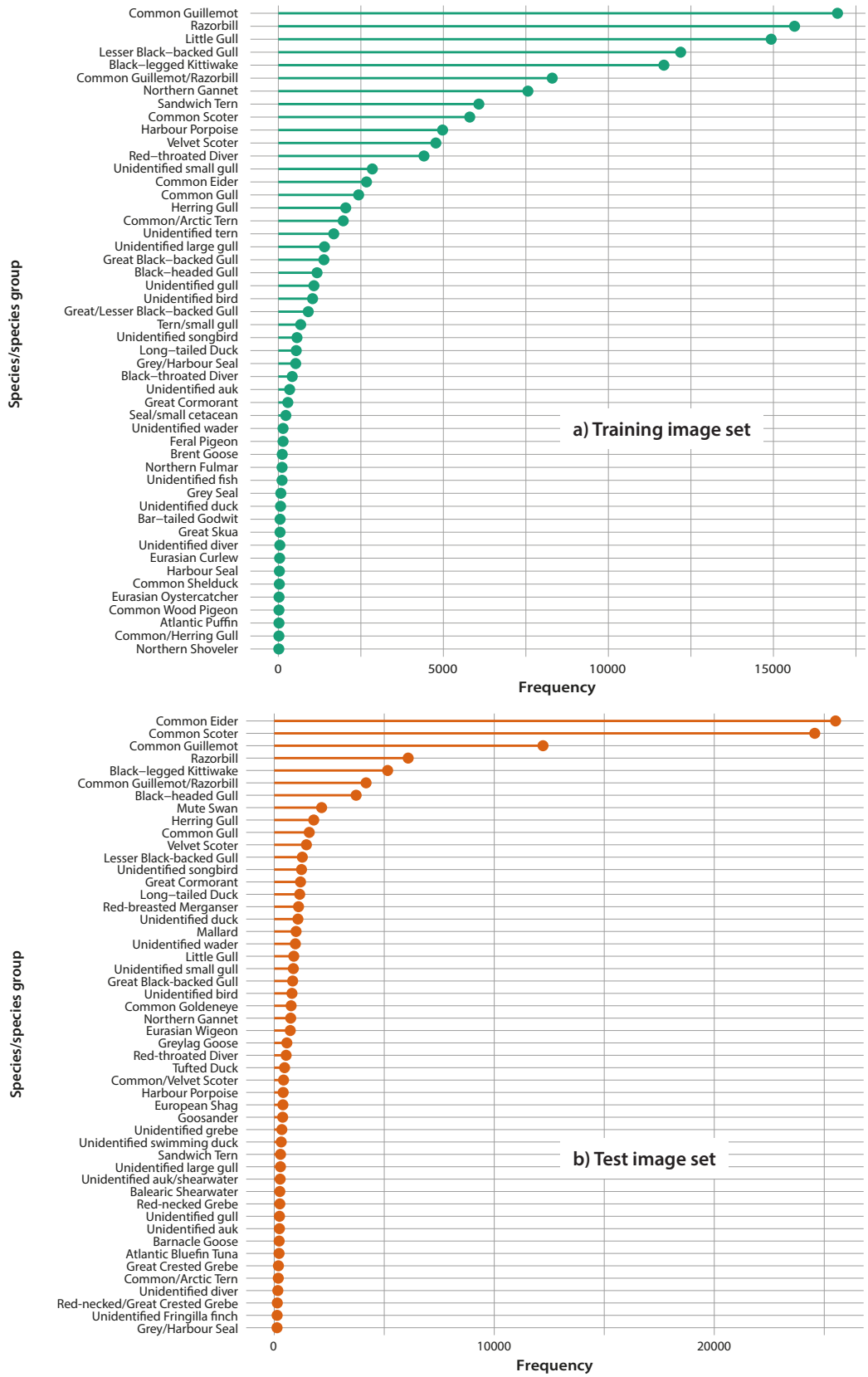


Figure 3: Frequency of the fifty most common taxa in a) the training image set and b) the test image set.

with the bounding boxes predicted by the model to calculate the intersection over union (IoU). In our test design, however, object detections made by observers on the video material were the benchmark against which we compared the performance of HiDeFIND, rather than the actual number of objects potentially detectable in the field comparison (see above). The detections made by observers were located by point markers (see 2.2.4.1). We therefore rated model predictions as true positives if the bounding boxes predicted by the model included the manually set point markers. We rated model predictions as false negatives if manual point markers were not included by model-generated bounding boxes. We classified all other model predictions as false positive detections. For technical reasons, false positive detections therefore also included detections of tracked biological objects that did not occur on the frame on which the observers had point-marked the object (see also 2.2.2). True negative model predictions were not defined in our test design.

2.2.5.3 Performance metrics

We used the following standard metrics for evaluating model performance for classification tasks:

$$\text{Sensitivity/Recall} = \frac{\textit{True Positives}}{\textit{True Positives + False Negatives}}$$

and

$$\text{Precision} = \frac{\textit{True Positives}}{\textit{True Positives + False Positives}}$$

It is recommended for the evaluation of AI models to be reported in a stratified way (Burnell et al. 2023). To analyse how strongly HiDeFIND performance varied depending on species/species group, we therefore report sensitivities not only globally (comprising all species/species groups), but also separately for species/species groups, provided that their frequency in the test image set reasonably supported this. Since false positive detections cannot be assigned to species/species groups in our test design, this was not done for precision. In addition, we analysed to which extent the performance indicators depended on the position of the sensors, the season and potentially detection-relevant environmental conditions (see Appendix 1 and 2.3.3 and 2.3.4).

2.3 Statistical analysis

2.3.1 Background

In contrast to standard test designs in computer vision, we were unable to use an established set of verified images with relevant objects for the HiDeFIND performance test (such image sets are available on the internet for training, validation and testing purposes for many standard applications in object classification). Instead, a sample of completed aerial surveys was taken to compose the test image set (see 2.2.5.1). The resulting data sets were subject to a hierarchical dependency structure (see 2.3.2 and Appendix 3). In addition to the detailed description of model performance, we therefore used statistical approaches to calculate unbiased confidence intervals around the point estimators of performance metrics in order to take sampling error into account. In a further step, we investigated causes of variation in model performance to identify conditions under which HiDeFIND had not yet shown optimal performance. The results of these exploratory analyses are intended to guide, for example, further optimisation of the model architecture or the composition of future training image sets.

2.3.2 Hierarchical dependency structure of HiDef data

The detection probabilities of individual objects on HiDef video material are not independent of each other but are subject to a hierarchical dependency structure (for details, see Appendix 3). This applies in principle to both manual and machine object detection. In order to control pseudoreplication (Hurlbert 1984) and thus avoid type 1 error inflation (Forstmeier et al. 2017), this dependency structure must be taken into account in statistical models of HiDeFIND performance. This is achieved by fitting statistical models to simultaneously estimate both fixed effects and random effects (mixed effects models). A nested structure of relevant random effects accounts for the relevant hierarchical levels as grouping factors in the statistical model (Gelman & Hill 2006).

2.3.3 Sensitivity

To estimate confidence intervals for global and species-specific sensitivities, we used generalised linear mixed models (GLMMs) with a binomial error structure and logit link function, as well as object/individual as the statistical unit. We fitted the overall mean as the only fixed effect (null model) and additionally considered the identity of project areas, transects and reels as nested random effects (see also Appendix 3). We calculated corresponding 95% confidence intervals by multiplying the standard errors of the estimates on the link scale by 1.96 and, for better comprehensibility, present detection probabilities as percentages including re-transformed 95% confidence intervals (the latter are therefore asymmetric). When calculating species-specific sensitivities, convergence problems regularly occurred even for the more common species in the test data set. Due to the smaller sample sizes, the models were unable to distinguish very small random effects from zero (the variances in question were effectively zero). In such cases, we simplified the random effects structure for each species until the models converged, removing random effects in descending order. To test the effects of detection-relevant environmental conditions on sensitivity, we used GLMMs with a binomial error structure and logit link function. Environmental conditions were mod-

elled as fixed effects in separate models (see also Appendix 1, water turbidity for marine mammals only). In addition, we considered the identity of project areas, transects and reels as nested random effects (see also Appendix 3). We tested the statistical significance of fixed effects by comparing a given model with the null model using a likelihood ratio test (R function *anova*).

2.3.4 Precision and number of false positive detections

In our test design, false positive detections can occur on any of the more than two million frames in the test image set, not just on the approximately 51,000 images on which human observers had previously marked a relevant object. The precision in our test design is therefore not comparable with other performance tests in computer vision. Other studies generally show significantly higher precision because they offer substantially fewer designated non-object test images in the test (usually in similar numbers to test images containing a relevant object). We therefore only report the overall precision and instead analyse in detail the number of false positive detections per frame as a function of potentially detection-relevant environmental conditions. This allows us to draw valuable conclusions about conditions under which HiDeFIND was not yet able to deliver optimal precision. We modelled the number of false positive detections per frame using linear mixed effects models (LME) with normal error distribution after logarithmic transformation ($\log_{10}$). The maximum permitted number of detections per frame (sum of true positive and false positive detections) and thus also the maximum possible number of false positive detections was limited to 1000 (only 49 of the more than two million frames had >1000 false positive detections). Environmental conditions were modelled as fixed effects in separate models (see Appendix 1, water turbidity only for marine mammals) and we considered the identity of project areas, transects and reels as nested random effects (see also Appendix 3). We tested the statistical significance of fixed effects by comparing a given model with the null model using a likelihood ratio test (R function *anova*).

2.3.5 Diversity indices

Separately for both data sets, we used i) the number of species/species groups to quantify species or species group richness; ii) the Shannon diversity index

$$H' = -\sum_{i=1}^{S} p_i \, (\log 2 \, p_i)$$

with S = number of species and p_i = relative abundance of species i in the respective data set to quantify species or species group diversity; iii) the Shannon diversity index divided by its maximum possible value for the given species richness of the data set to quantify the evenness of species or species groups:

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

3 Results

3.1 Global sensitivity (across all species/species groups)

Of a total of 111,666 objects in the test image set, HiDeFIND detected 107,778, resulting in an overall global sensitivity of 96.5%. Taking into account the hierarchical dependency structure of the data, the global weighted mean sensitivity was 99.4% (Table 2). Almost two-thirds of the variance in detection probability in the test image set was explained by differences between reels, about a quarter by differences between transects, and roughly 10% by differences between aerial surveys (Table 2).

The detection probability was independent of the position of the sensors (inner *versus* outer), the season and potentially detection-relevant environmental conditions, except for glare (Table 3). With increasing glare, the detection probability decreased significantly.

Figure 4 contrasts false negative with true positive predictions of the model for a selection of species with high relevance for maritime environmental planning (see also Discussion).

3.2 Species- and sex-specific sensitivities

As expected, species-specific sensitivities were generally high: Table 4 shows the species-specific overall sensitivities as well as the weighted mean detection probabilities for the twelve most common taxa in the test image set and for six other important target species in maritime environmental planning (including Harbour Porpoise). Only one of six sexually dimorphic species (all ducks) showed sex-specific sensitivity: male Common Scoters had a slightly lower detection probability than females (Table 4).

3.3 Overall precision and number of false positive model predictions

Of a total of 6,443,717 model predictions, 107,778 were true positive, resulting in an overall precision of 1.7% (the precision calculated here is not comparable to that from other performance tests in computer vision, see 2.3.4). False positive detections occurred in 99.3% of the total 2,096,554 frames and on all reels. The number of false positive detections per frame ranged from 0 to 1,000 (capped at 1,000, see 2.3.4). The median was 1.

The number of false positive model predictions was independent of the position of the sensors (inner *versus* outer) and the season (Table 5, Figures 5a and 5b). However, it increased with increasing glare, particularly in case of strong glare (Table 5, Figure 5c). Significant effects were detected for sea state and air and water turbidity (Table 5), but the effect sizes were small (Figures 5d to 5f).

Table 2: Global weighted mean sensitivity of the artificial neural network HiDeFIND across all 124 species/species groups of the test image set. Results of a generalised linear mixed model with binomial error structure and logit link function.

				Fixed effects	Random effects		
Model	**N (objects)**	**True positive**	**False negative**	**Intercept (95% confidence)[1]**	**Surveys**	**Transects**	**Reels**
All species/species groups	111 666	107 778	3888	99.41 (99.15, 99.59)	0.35	0.61	1.60

[1]Weighted mean sensitivity.

Table 3: 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 generalised linear mixed models with binomial error structure and logit link function.

		Fixed effects				Random effects		
Model	**N (objects)**		**Chisq**	**DF**	**p[1]**	**Surveys**	**Transects**	**Reels**
Inner/outer sensors	111 666	Two-level factor	1.01	1	0.32	0.35	0.61	1.60
Season	111 666	Four-level factor	2.59	3	0.46	0.27	0.62	1.63
Glare	111 666	Four-level factor	26.10	3	<0.001[2]	0.33	0.61	1.55
Sea state	111 564[3]	Six-level factor	3.20	5	0.67	0.29	0.64	1.66
Air turbidity	111 564[4]	Two-level factor	0.81	1	0.37	0.35	0.60	1.60
Water turbidity (marine mammals)	702[5]	Two-level factor	0.03	1	0.85	Effectively zero[6]	9.43	77.70

[1]For factors with more than two levels p value for omnibus test.
[2]Sensitivity drops under moderate and in particular strong glare.
[3]Information on sea state missing for 102 observations.
[4]Information on air turbidity missing for 102 observations.
[5]Information on water turbidity missing for one observation.
[6]Singularity (variance indistinguishable from zero).

Table 4: 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 generalised linear mixed models with binomial error structure and logit link function.

				Fixed effects	Random effects		
Model	**N (objects)**	**True positive**	**False negative**	**Intercept (95 % confidence)[1]**	**Surveys**	**Transects**	**Reels**
Common Eider	25513	25322	191	99.73 (99.57, 99.83)[2]	Effectively zero[3]	0.37	1.63
Common Scoter	24567	23116	1451	98.81 (97.87, 99.34)[4]	Effectively zero[3]	1.77	1.48
Common Guillemot	12215	12146	69	99.99 (99.98, 100)	Effectively zero[3]	Effectively zero[3]	36.40
Razorbill	6087	6027	60	99.99 (99.98, 100)	Effectively zero[3]	Effectively zero[3]	52.76
Black-legged Kittiwake	5154	5124	30	99.99 (99.90, 100)	Effectively zero[3]	Effectively zero[3]	24.09
Guillemot/Razorbill	4174	4127	47	99.99 (99.95, 99.99)	Effectively zero[3]	Effectively zero[3]	46.07
Black-headed Gull	3722	3673	49	99.96 (99.59, 100)	Effectively zero[3]	Effectively zero[3]	19.04
Mute Swan	2153	2151	2	100 (99.92, 100)	Effectively zero[3]	Effectively zero[3]	84.12
Herring Gull	1795	1728	67	100 (99.98, 100)	Effectively zero[3]	Effectively zero[3]	135.70
Common Gull	1590	1575	15	100 (99.96, 100)	Effectively zero[3]	Effectively zero[3]	76.95
Velvet Scoter	1462	1432	30	100 (99.34, 100)[5]	Effectively zero[3]	Effectively zero[3]	24.57

Table 4: Continuation

				Fixed effects	Random effects		
Model	**N (objects)**	**True positive**	**False negative**	**Intercept (95 % confidence)[1]**	**Surveys**	**Transects**	**Reels**
Lesser Black-backed Gull	1275	1275	0	Not estimable[6]	–	–	–
Great Cormorant	1193	1048	145	99.82 (98.37, 99.98)	Effectively zero[3]	0.88	22.63
Long-tailed Duck	1157	1143	14	99.93 (80.07[7], 100)[8]	Effectively zero[3]	Effectively zero[3]	14.72
Northern Gannet	746	744	2	100 (99.48, 100)	24.12	Effectively zero[3]	332.78
Eurasian Wigeon	728	699	29	98.63 (94.90, 99.64)[9]	Effectively zero[3]	Effectively zero[3]	4.33
Red-throated Diver	540	538	2	99.63 (98.86, 99.94)	Effectively zero[3]	Effectively zero[3]	Effectively zero[3]
Harbour Porpoise	406	401	5	99.61 (91.88, 99.98)	2.72	Effectively zero[3]	Effectively zero[3]

[1]Weighted mean sensitivity.
[2]Sex specificity: p=0.51, N=7797 sexed individuals.
[3]Singularity (variance indistinguishable from zero).
[4]Sex specificity: p<0.001, N=16036 sexed individuals; males with 0.7% lower sensitivity.
[5]Sex specificity: p=0.29, N=1029 sexed individuals.
[6]Complete separation: Only true positive predictions.
[7]One of 109 reels with Long-tailed Ducks contributed with 120 individuals >10% of sample (all true positive predictions).
[8]Sex specificity: Not estimable, complete separation (all 13 females detected).
[9]Sex specificity: p=0.73, N=191 sexed individuals.

Table 5: 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.

		Fixed effects				Random effects			
Model	**N (frames)**		**Chisq**	**DF**	**p**[1]	**Surveys**	**Transects**	**Reels**	**Residual**
Inner/outer sensors	2 096 554	Two-level factor	1.86	1	0.17	Not converged	Not converged	0.022	0.090
Season	2 096 554	Four-level factor	5.45	3	0.14	0.003	0.011	0.005	0.090
Glare	2 096 554	Four-level factor	451.35	3	<0.001[2]	0.005	0.011	0.005	0.090
Sea state	2 090 829[3]	Six-level factor	454.20	5	<0.001[4]	0.005	0.011	0.005	0.090
Air turbidity	2 090 829[5]	Two-level factor	5.73	1	0.02[6]	0.090	0.011	0.005	0.090
Water turbidity	2 090 829[7]	Two-level factor	64.40	1	<0.001[8]	0.090	0.017	0.005	0.090

[1]For factors with more than two levels p value for omnibus test.
[2]False positives increase with increasing glare (see Figure 5).
[3]Information on sea state missing for 5725 observations.
[4]Five out of 15 contrasts significant, but no trend (see Figure 5).
[5]Information on air turbidity missing for 5725 observations.
[6]Minimally higher median for level no versus some air turbidity (see Figure 5).
[7] Information on water turbidity missing for 5725 observations.
[8] Higher median for level some versus no water turbidity (see Figure 5).

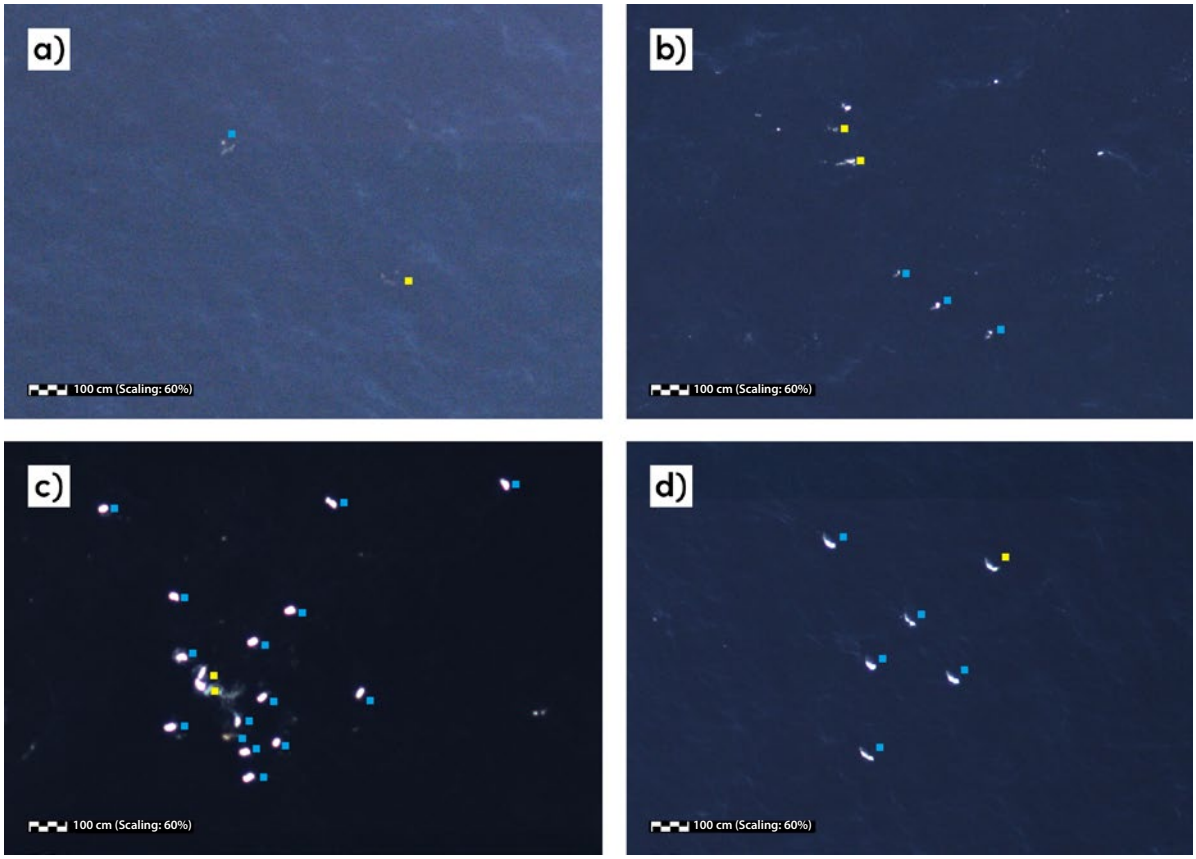


Figure 4: Examples of false negative model predictions. The figure shows in an exemplary manner HiDef frames featuring both true positive detections (blue annotations) and false negative detections (yellow annotations) for the same species. a) Common Scoters *Melanitta nigra*, b) Razorbills *Alca torda*, c) Common Eiders *Somateria mollissima*, d) Red-throated Divers *Gavia stellata*.

4 Discussion

4.1 Sensitivity

Global sensitivity

The global mean sensitivity of HiDeFIND across all 124 species/species groups represented in the test image set (including 109 bird taxa) was very high at over 99%. HiDeFIND was therefore able to detect birds at sea and marine mammals in HiDef video material almost as well as observers specially trained for this purpose. A 95% confidence interval of 99.2-99.6% also indicates that the estimate is very precise and suggests that the model will be able to achieve a comparably high sensitivity for HiDef aerial surveys with similar species compositions and weather conditions. We also found no detectable difference in the mean sensitivity between the inner and outer sensors (Table 3). HiDeFIND thus performed similarly well across the entire swath of effectively 544 m.

The global mean sensitivity estimated using the mixed model was slightly higher than the global overall sensitivity. The former accounted for the hierarchical dependency structure of the HiDef data and weighted the contributions of groups of dependent data to the overall effect. Discrepancies between the two measures can arise, for example, when false negative detections are clustered in relatively few frames and/or reels. In this case, the overall sensitivity tends to underestimate the actual sensitivity. Such clustering was indeed detectable in the test data set: for example, 947 of a total of 3,888 (24.4%) false negative detections were attributable to just ten of the total 51,218 (0.02%) frames that contained at least one biological object.

The objects detected manually as part of the established work process were the benchmark for assessing sensitivity. If objects were overlooked in the manual process, some of the model predictions classified as false positives (see 4.2 below) were in fact true positive model predictions (false negative manual detections). However, our test design was unable to identify these as such. The very high global sensitivity suggests that HiDeFIND could also find some of the objects overlooked in the manual workflow. We will quantify this promising additional potential using independent test image material in future tests with a suitable design.

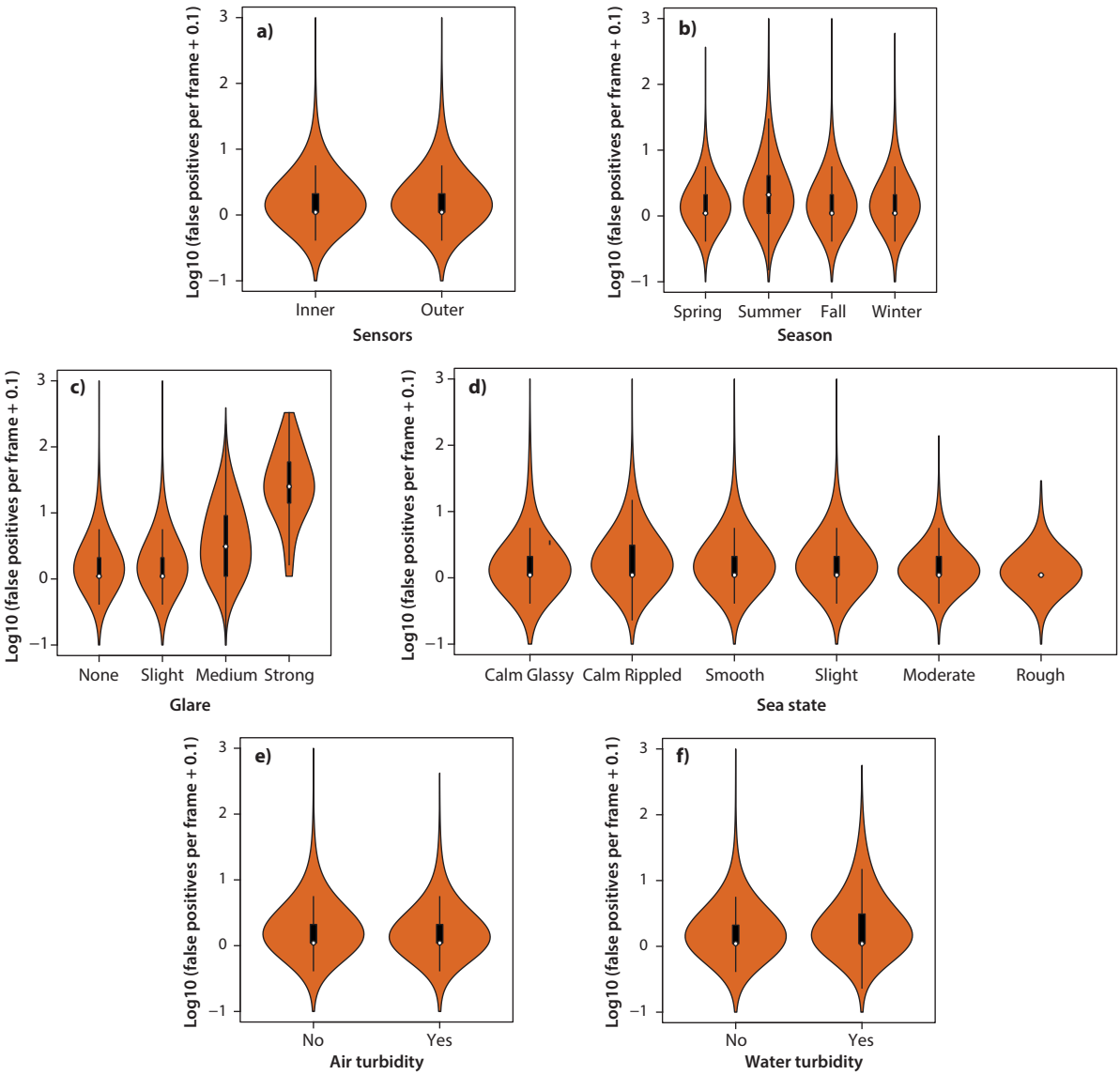


Figure 5: Number of false positive model predictions per frame in relation to a) positions of the sensors, b) seasons, c) glare, d) sea state (Petersen scale), e) air turbidity and f) water turbidity (see also Appendix 1).

Generalisability

Although HiDeFIND was trained with annotated objects from 66 species/species groups and two marine areas, the model achieved its performance on a test image set that included well over a hundred species/species groups from three marine areas (Appendix 2). Therefore, we can conclude that HiDeFIND has learned to generalise. This is a desired feature for artificial neural networks, and our findings suggest that HiDeFIND will be able to achieve high sensitivities even when confronted with expanded or completely new species compositions. We expect a great advantage when applied in new marine areas and in the event of shifts in distribution ranges, e.g. as a result of anthropogenic environmental change.

Species-specific sensitivities

As expected, species-specific sensitivities were very high for most species (Table 4). This included the detection of species of outstanding importance for maritime spatial and environmental planning, such as Red-throated divers, Common Guillemots and Black-legged Kittiwakes (Furness et al. 2013, Fliessbach et al. 2019, Dierschke et al. 2024) and, among marine mammals, Harbour Porpoises. Species-specific sensitivities generally varied only slightly (Table 4), suggesting future development efforts to focus on further optimising a general HiDeFIND model rather than developing several species- or species group-specific HiDeFIND special modules (such as a diver module).

Sex-specific sensitivities
In species with pronounced sexual dimorphism, model sensitivity could be sex-specific. We were able to test this for some duck species, for which we assumed adequate statistical power based on their frequency in the test image set. While the sensitivities for the sexes were not statistically distinguishable for Common Eiders, Velvet Scoters *Melanitta fusca* and Eurasian Wigeons *Anas penelope*, HiDeFIND showed a slightly lower sensitivity for male Common Scoters than for females (Table 4, see also Figure 4a). However, considering the species-specific sensitivity of >98%, this effect was very small, with a difference of 0.7%, and can be neglected for population estimates.

Environmental variation
The global mean sensitivity decreased significantly with increasing glare (Table 3), which is not surprising, as increasing glare can easily outshine and thus mask objects on the sea surface. Compared to the training image set, the test image set also showed less intense glare on average (all other detection-relevant environmental conditions were similar; see Appendix 2), potentially favouring the high global sensitivity in the test data set. However, this would only affect the assessment of model sensitivity if the average glare in the test data set were unrepresentative of future data sets. In any case, the effect on the overall performance of the model is very limited, as stronger glare can be effectively avoided by turning the HiDef sensor system away from the sun (see 2.1.1). Accordingly, only 3504 (0.03%) and 158 (0.001%) objects in the test data set were attributable to video material with moderate or strong glare, respectively. HiDeFIND proved to be robust in relation to other environmental conditions that could potentially affect detection: Neither sea state nor air turbidity or water turbidity (in the case of marine mammals) had a detectable effect on sensitivity (Table 3). The same was true for the seasons (Table 3). We conclude that HiDeFIND can be used all year round and under varying environmental conditions, even suboptimal ones, without restricting result quality.

Causes of false negative detections
Statistically, conditions with which the probability of false negative detections varies can be narrowed down (see previous section). In practice, however, it is impossible to determine why specific objects were not detected by the model. Such limited transparency of the machine decision-making process is characteristic of deep learning-based neural networks, whose successive computations across many network layers largely correspond to so-called black box operations. Consequently, we found quasi-stochastic occurrences of false negative detections, the causes of which can only be speculated upon.

Despite the overall very good generalisability, species or species groups that were not or only very rarely included in the training image set might be less well detected in the test image set than more frequently represented species. For example, the training image set did not contain any Eurasian Coots *Fulica atra*, and the model detected only two out of five individuals in the test image set. However, such cases mainly concern species that rarely occur in the surveyed project areas and therefore are of minor relevance. For more common species, false negative detections could result if characteristics such as body posture or diving status differ in frequency between the training and test image sets. For example, one false negative detection concerned the snapshot of a Harbour Porpoise in the process of diving; its body therefore appeared to be divided into two parts. Figures 4a-4c show three examples of false negative predictions.

We rated a model prediction as true positive if the bounding box predicted by the model included the manually set point marker (see 2.2.5.2). If a point marker was not set precisely in the centre of the object and was therefore located just outside a model-generated bounding box, a false negative detection was recorded, even though the model had detected the object ("false false negative detections"). A random sample check revealed that less than 10% of all false negative detections were attributable to this shortcoming in the test design, including cases where the objects were difficult for human observers to overlook (an example of this is a Red-throated Diver in Figure 4d). The phenomenon is a technical testing issue, which only leads to an underestimation of the model's actual sensitivity, but also to a slight overestimation of the false positive rate, as objects actually detected by the model in immediately adjacent locations were recorded as false positive detections. The HiDeFIND performance evaluation we have conducted in this article should therefore be considered conservative.

Alternative AI models with similar objectives
While machine learning in combination with drones or satellites as platforms has already been used very successfully for monitoring animal populations at sea (see introduction), there are very few studies published on the combination of machine learning with aircraft-based surveys for population assessment at sea. A direct comparison of specific performance indicators of competing models in computer vision is generally only of limited significance, as both training and testing are usually based on vastly different image sets. To our knowledge, fewer than a handful of peer-reviewed studies employ a design similar enough to ours – platform: aircraft; sensor: digital camera system; target object: seabirds at sea – to allow for any comparison at all.

Kuru et al. (2023) describe a method for semi-automatic detection of Northern Gannets on aerial photographs, which are similar in type and resolution to HiDef still images. Kuru et al. (2023) achieved an un-

weighted sensitivity of 97.1% for their model[1], which is comparable to the unweighted sensitivity of 99.7% achieved by HiDeFIND for Northern Gannets (see also Table 4). In contrast to the broad taxonomic scope of our approach (Figure 3, Appendix 2), the method described by Kuru et al. (2023) and/or its quantitative performance analysis was limited to the Northern Gannet, the largest and most conspicuous seabird species in European waters. The suitability of the method described by Kuru et al. (2023) for smaller and less conspicuous species therefore remains uncertain. It is possible that the analysis by Kuru et al. (2023) only included individuals in conspicuous adult plumage (illustrations in the paper only showed adult individuals in flight). In contrast, both our training and our test image set contained not only (sub-) adults but also individuals in the less conspicuous plumage of the 1st and 2nd calendar years (160 in the training image set and 10 in the test image set, the latter all true positive) as well as swimming Northern Gannets (hundreds in each of our two image sets). Furthermore, Ke et al. (2024) report sensitivities of up to 65% in a deep learning approach that is intended to enable population estimates of seabirds at sea during flight and thus almost in real time. Due to very low flight altitudes of only about 25 m to 200 m, this model had access to image data with a very high ground sampling distance of between 0.14 cm and 1.47 cm, which naturally has a positive effect on performance tests of object detection models. In addition, the objects offered by Ke et al. (2024) for training and testing were limited to only two sampling plots and largely to wintering sea ducks. HiDeFIND achieved higher sensitivities even though it was confronted with image data that had been captured from a significantly greater altitude of approximately 500 m for flight safety reasons and therefore had poorer ground sampling distances of approximately 2 cm. Furthermore, the HiDeFIND image material covered a much broader spectrum of species (Figure 3, Appendix 2). Finally, Weiser et al. (2023) presented a method for the automated detection of resting *Branta* geese in protected shallow water at one site in Alaska. Although the model was able to distinguish images with geese from those without, the subsequent automated counting underestimated the populations (lack of sensitivity), requiring time-consuming manual reworking as part of a "human-in-the-loop" work process.

[1] Calculated from Table 3 in Kuru et al. (2023). In contrast to our analysis, their calculations were not based on individuals as a unit, but on aerial photographs that showed either at least one or no Northern Gannet. A detection was considered a true positive if the method recognised an image showing at least one Northern Gannet as such.

4.2 Precision and number of false positive model predictions

Precision

HiDeFIND generated a high number of false positive detections, which was reflected in a low unweighted overall precision. This tendency towards overdetection was desired explicitly per design, as we prioritized sensitivity (as complete detection as possible) over precision (as high efficiency as possible) to avoid compromising result quality compared to the established manual process. In a partially automated "human-in-the-loop" work process, however, low precision entails a high manual effort to separate false positive detections prior to species identification. This reduces achievable efficiency gains through the automation of initial object detection. A primary goal in the further development of an AI-supported work process is therefore to improve precision without compromising too much on sensitivity.

Position of the sensors and environmental variation

The position of the sensors (Table 5, Figure 5a) and the seasons (Table 5, Figure 5b) had no detectable effect on the number of false positive detections. However, as glare increased, the model produced more false positives. This was particularly pronounced for strong glare (Table 5, Figure 5c). Other environmental conditions potentially relevant to detection, such as sea state, air turbidity or water turbidity, showed significant differences (Table 5), but the effects were small (Figures 5d to 5f). Due to the large sample size of >2 million frames, the power of our analyses was high and allowed us to detect even minor differences between individual factor levels. These small effects are unlikely to be relevant in practical application, partly because surveys are usually only scheduled for time slots that promise good detection conditions, such as clear air and calm seas. Nevertheless, the effects of detection-relevant environmental conditions should be investigated in more detail in future performance tests of HiDeFIND itself, as well as of comparable alternative models for the detection of seabirds at sea.

Mitigation of the effects of glare

Improvements in HiDeFIND precision and thus improvements in the efficiency of the entire AI-supported work process could focus on further minimising the effects of glare beyond the established technical measures (see 2.1.1). Glare effects could be mitigated by controlled reduction of the model's sensitivity, e.g. by increasing confidence thresholds in post-production, which define the minimum confidence of the model to trigger a detection. Aiming to achieve the most complete detection possible, no such threshold was defined for the tested HiDeFIND version 1.0. To mitigate false positive detection spikes, the maximum permissible number of detections per frame could be reduced by sorting the model detections in descending order of confidence. If true positive detections have higher confidence levels on aver-

age than false positives, false positive detections would then be primarily capped on frames with a mixture of true positive and false positive detections. For the performance test presented here, the number of permitted detections per frame was limited to 1,000 (49 frames were affected by this cap), but the highest number of relevant objects in a solitary case was only 811. For the remaining 51,217 frames with at least one relevant object, the number of relevant objects was consistently less than 100. Based on a sample of future performance tests, the effects of reducing the maximum number of permitted detections per frame should be examined.

Trade-off precision versus sensitivity
Increasing the confidence thresholds for detections and reducing the maximum number of permitted detections per frame could result in a reduction in global sensitivity. The primary goal of future development will therefore be to increase precision at a given sensitivity to fully exploit the advantages of automation, such as acceleration, increased reproducibility and greater cost-effectiveness. In any case, quality of the results should be at least as high as that of the established manual work process. Achieving this overarching development goal is facilitated by technical innovation in sensors, which now achieve a noticeably better average ground sampling distance of well below 2 cm at comparable flight altitudes and swath widths (BioConsult SH, own data).

4.3 Quality assurance of AI-supported population surveys at sea

AI-supported methods offer great opportunities to evaluate digital aerial survey data in maritime spatial and environmental planning, especially in terms of analysis speed. For the assessment of seabird populations and distributions at sea, and thus the legal certainty of planned projects, the quality, continuity and comparability of collected data are of fundamental importance. When transforming manual to (partially) automated work processes, it should therefore be ensured that the performance and practical suitability of AI models are comprehensively evaluated before application. In addition, the practical application should be accompanied by a quality assurance concept that defines how the quality of results is ensured in the automated evaluation of digital survey flights.

4.4 Conclusion

The HiDeFIND artificial neural network detected seabirds at sea and marine mammals on digital video recordings almost as well as specially trained observers. This also applied to key species in maritime spatial and environmental planning, regardless of the season and largely independent of detection-relevant environmental conditions. The model demonstrated good generalisability, which is likely due to the very extensive and diverse training image set and promises successful use in other marine areas with new species spectra. The high global sensitivity also suggests that HiDeFIND could prospectively find some of the objects overlooked in the manual work process. The high rate of false positive detections currently reduces the efficiency of an AI-supported work process and thus the achievable time savings and the cost-effectiveness of its use. However, thanks to its excellent sensitivity, this does not reduce the effectiveness of the model, i.e. the high quality of the results and the population or density estimates that can be derived from them. HiDeFIND thus offers a forward-looking alternative to purely manual object detection for the analysis of digital aerial survey data in maritime spatial and environmental planning, environmental monitoring and research. HiDeFIND operates within the framework of an integrated "human-in-the-loop" work process, in which automated object detection by AI is accompanied by a quality assurance process conducted by trained observers.

Acknowledgements
We would like to thank Claudia Burger, Kelly Macleod, and especially Hanna Kreutzfeldt and Anna Kersten for discussion and helpful comments on earlier versions of this manuscript. Christian Vlasak provided the graphic design for Figure 1 and the figure in Appendix 3. Venela Matz assisted us in selecting and annotating the images in Figure 4. We would like to thank Wolfgang Fiedler and an anonymous reviewer for their helpful comments.

Conflicts of interest
Tim Schmoll, Guruprasad Hegde, Monika Dorsch and Georg Nehls work for BioConsult SH GmbH & Co KG, a for-profit company that offers digital offshore aerial surveys using the HiDef method and AI-supported analysis.

5 References

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 Investigation of the Impacts of Offshore Wind Turbines on the Marine Environment (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.

Glossary of technical terms

Bounding box: In computer vision, bounding boxes delimit objects of interest in rectangular form as closely as possible. In image material for training purposes, images are annotated with bounding boxes to provide the model with visual patterns of the target objects as learning examples.

Ground sampling distance: Actual distance on the earth's or sea's surface that corresponds to the distance between the centres of neighbouring pixels in the digital image.

Convolutional Neural Network (CNN): CNNs are a deep learning-based variant of artificial neural networks that are particularly well suited for processing image data. CNNs use convolutional layers to recognise local patterns such as edges or shapes and gradually combine them into more complex structures. CNNs learn relevant patterns directly from their training data and are a central component of modern automated image analysis methods.

Deep learning: As a sub-discipline of machine learning, deep learning uses artificial neural networks with numerous deeply layered network layers between the input layer (e.g. RGB values of the pixels of a digital image) and the output layer (e.g. image contains seabird yes/no). This enables computers to learn autonomously from examples in a highly efficient manner. Machine learning is in turn a subfield of artificial intelligence.

Explainable Artificial Intelligence (XAI): The decisions made by artificial neural networks are often difficult to comprehend in individual cases ("black box"). This problem is addressed by a separate subdiscipline of deep learning known as explainable artificial intelligence (overview e.g. in Saeed & Omlin 2023). While sensitive applications such as automated lending place high demands on the traceability of the decision-making process, this aspect is generally considered less crucial in computer vision. The evaluation of models in computer vision is therefore strongly results-oriented and less process-oriented.

Fixed effects: Estimators of fixed effects in statistical models typically represent differences in means (for factors) or slopes (for covariates).

Gradient descent: An optimisation algorithm frequently used in machine learning to efficiently determine local minima of a loss function. In our application, these are minima of the deviation between the actual object position and the object position predicted by the model in digital images. The underlying heuristic of the gradient method corresponds to the so-called "mountaineer algorithm with negative sign": A mountaineer in thick fog will climb to the summit – a local maximum – by the shortest route if he chooses the steepest ascent at every step during the climb. Similarly, the optimisation algorithm moves along the steepest gradient descent.

Intersection over Union (IoU): An important parameter in computer vision that indicates the degree of overlap between the bounding box (ground truth) derived from the field comparison and the bounding box predicted by the object detection model (in percent). Depending on the required precision of the localisation, threshold values can be set above which a model prediction is evaluated as correct (true positive).

Precision: Proportion of true positive predictions out of all positive model predictions (sum of true positive and false positive predictions); also referred to as *positive predictive value* in machine learning.

Sensitivity: Proportion of true positive predictions out of all actual positive cases (sum of true positive and false negative predictions); in machine learning, also referred to as *recall* or *true positive rate*.

You Only Look Once (YOLO): Early object detection models were based on a computationally intensive, two-step process: first, regions with potential objects were located, then, in a second step, objects were classified (is target object yes/no). YOLO models, on the other hand, perform both steps simultaneously with only a single presentation of an image – an innovation that has greatly simplified and accelerated automated object detection.

Random effects: Estimators of random effects in statistical models typically represent variances caused by differences among groups of dependent data. Among other things, it is assumed that the groups/levels represented in the data set are a random sample from a theoretically unlimited set of possible groups/levels of this effect.

Appendix 1: Description of detection-relevant environmental conditions.

Environmental condition	Description	Scale	Assessed per	Errors (dependent variable)	Predictor type (independent variable)
Glare	Degree of sun reflection on the sea surface.	Ordinal: None, slight, moderate, strong.	[1]Frame.	Gaussian.	Fixed effect four-level factor.
Sea state	State of the open sea surface produced by both swell and wind sea.	[2]Ordinal: Calm (glassy), calm (rippled), smooth, slight, moderate, rough.	[3]Row of horizontally adjacent frames across the eight reels.	Gaussian.	Fixed effect six-level factor.
Air turbidity	Degree of turbidity of the air body between aircraft and sea surface.	Ordinal: None, slight, moderate to strong.	[3]Row of horizontally adjacent frames across the eight reels.	[4]Binomial.	[4]Fixed effect two-level factor (air turbidity versus none).
[5]**Water turbidity**	Degree of turbidity of seawater caused by suspended matter.	Ordinal: None, slight, moderate to strong.	[3]Row of horizontally adjacent frames across the eight reels.	[6]Binomial.	[6]Fixed effect two-level factor (water turbidity versus none).

[1]Estimated from each of the eight reels of a transect.

[2]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]Estimated only from the reels captured by virtual camera #2.

[4]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 for (partially) submerged marine mammals.

[6]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.

Appendix 2: Attributes of training and test image set.

Attribute	Training image set	Test image set
Years	2017, 2021	2021, 2022
Total transect kilometers	~11 650	~8040
Area analysed/observed (km²)	~6080/6210	~4180/4280
HiDef video footage analysed (h)	~280	~250
Marine areas	North Sea, Baltic Sea	North Sea, Baltic Sea, English Channel
Project sites	4	6
Aerial surveys	21[1]	13[1]
Transects	291	196
Reels	1629	1408
Frames	79 134[2]	51 218[2]
Objects marked by humans (not the model)	138 681[3]	111 666
Individual organisms marked by humans (not the model)	26 635	111 666[4]
Bird objects marked by humans (not the model)	132 718	110 697
Mammal objects marked by humans (not the model)	5846	703
Other objects marked by humans (not the model)	117[5]	266[5]
Species/species group richness	66	124
Species/species group diversity	4.28[6]	4.05[6]
Species/species group evenness	0.71[7]	0.58[7]
Bird species richness	37	73
Bird species group richness	20	36
Bird taxa richness	57	109
Bird species diversity	3.64[6]	3.48[6]
Bird species evenness	0.70[7]	0.56[7]

Appendix 2: Continuation

Attribute	Training image set	Test image set
Bird taxa diversity	4.16[6]	3.99[6]
Bird taxa evenness	0.71[7]	0.59[7]
Cumulative percentage of five most abundant bird taxa	53.8[8]	66.4[8]
Cumulative percentage of ten most abundant bird taxa	78.3[8]	78.6[8]
Cumulative percentage of twenty most abundant bird taxa	94.9[8]	88.8[8]
Mammal species richness	4	5
Mammal species group richness	3	2
Mammal taxa richness	7	7
Mammal species diversity	0.18[6]	1.21[6]
Mammal species evenness	0.09[7]	0.52[7]
Mammal taxa diversity	0.83[6]	1.80[6]
Mammal taxa evenness	0.30[7]	0.64[7]
Weighted mean glare[9] (95% confidence)	0.61 (0.44-0.78)	0.20 (0.12-0.28)
Weighted mean sea state[9] (95% confidence)	2.54 (2.26-2.82)	2.11 (1.82-2.40)
Weighted mean probability of (some) air turbidity[9] (95% confidence)	0.17 (0.10-0.27)	0.15 (0.08-0.23)
Weighted mean probability of (some) water turbidity[9] (95% confidence)	0.16 (0.09-0.26)	0.18 (0.11-0.28)

[1]See Figure 2 in main text for seasonal distributions.

[2]Frames, on which human observers had detected at least one biological object.

[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]In the test image set, each individual is represented only once (see text 2.2.4.3).

[5]For example Atlantic Bluefin Tuna, Ocean Sunfish, Lion's Mane Jellyfish.

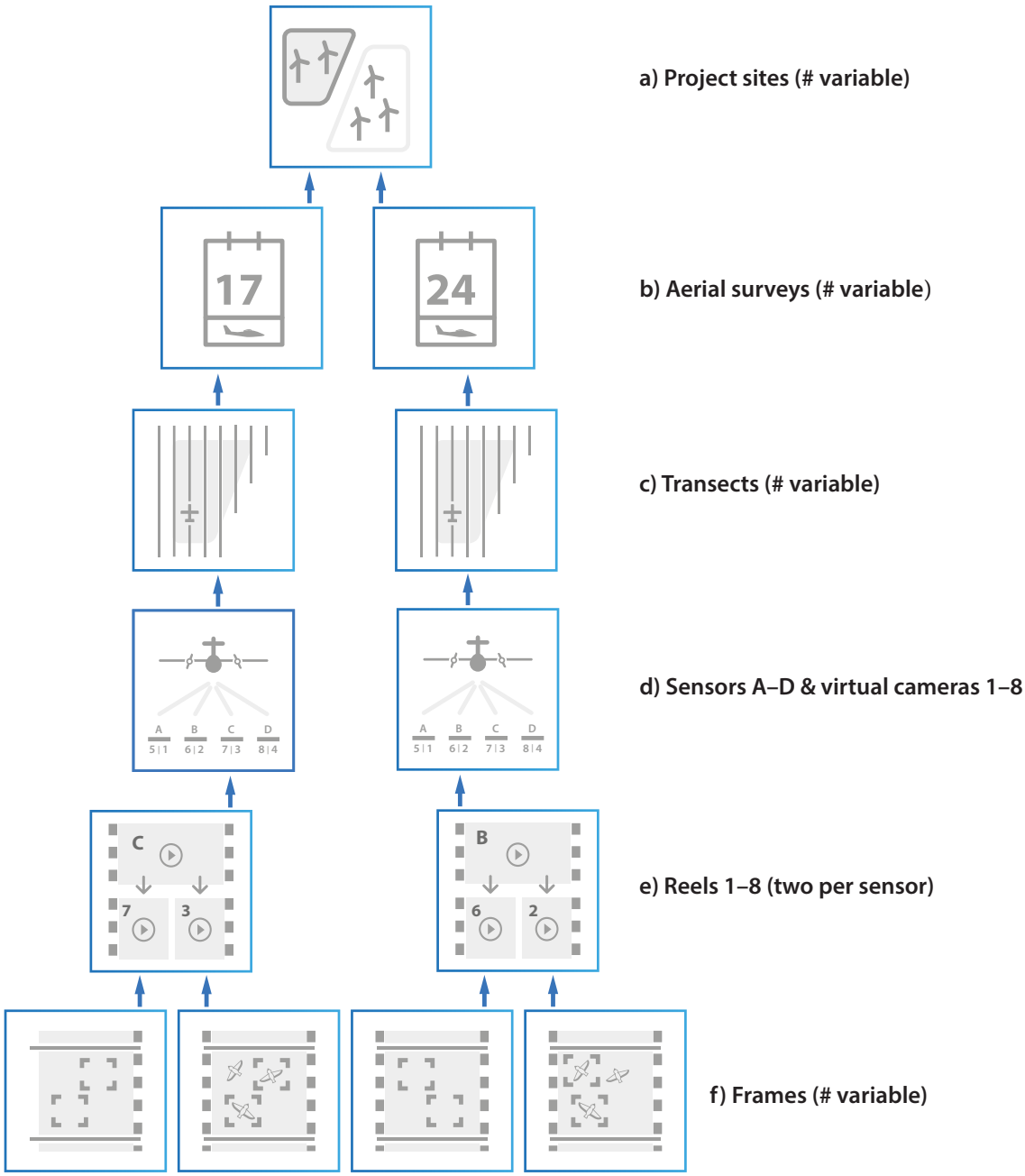
[6]Shannon-Index ($\log_2$).

[7]Shannon-Index/maximum possible Shannon-Index.

[8]See Figure 3 in main text for species identities.

[9]See also Appendix 1.

Appendix 3: Hierarchical dependency structure of HiDef data.



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.

Dependencies in HiDef data sets could be caused by the:

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).

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.

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 (e.g., windward versus leeward sides around islands).

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.

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).

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.