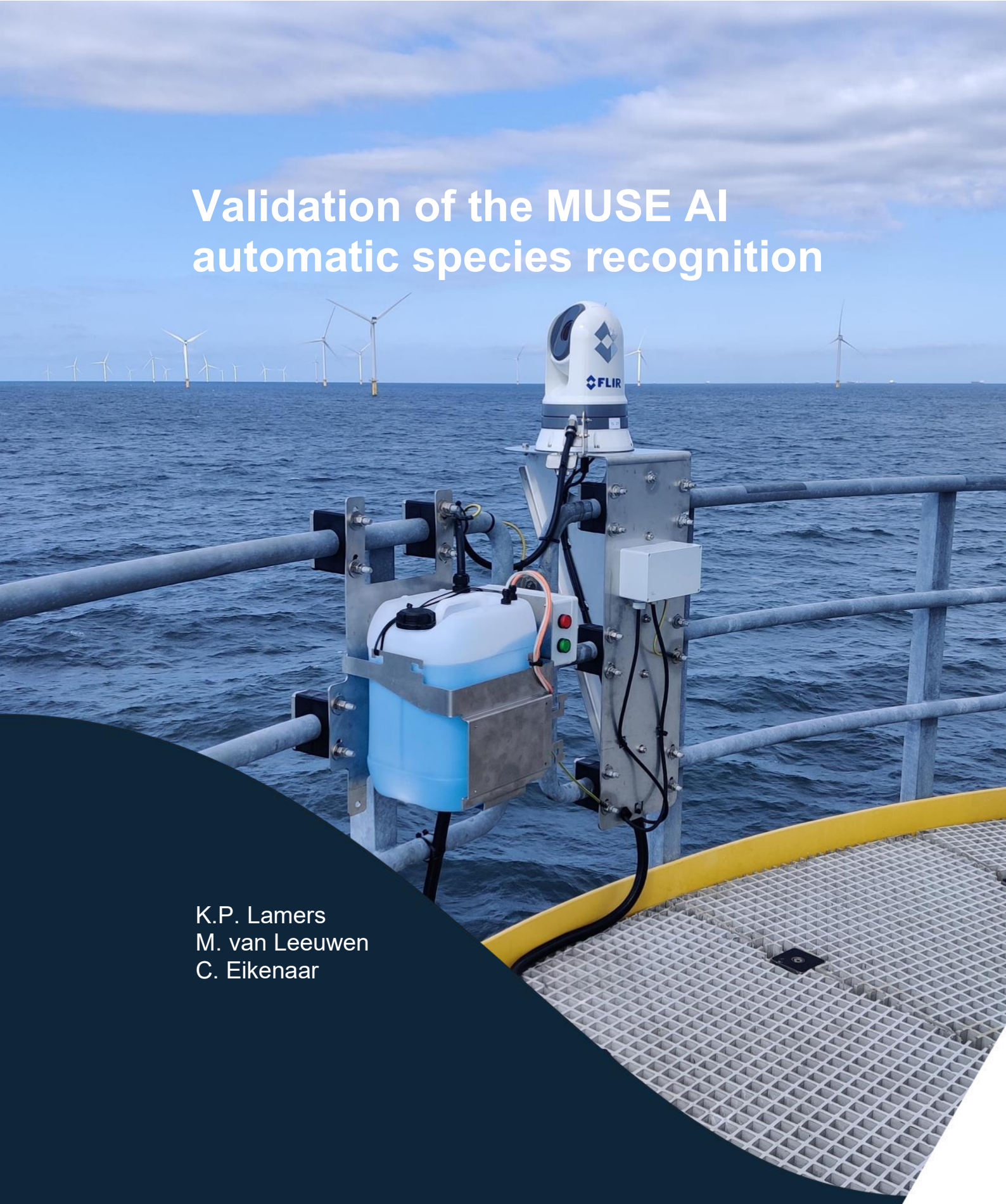
Validation of the MUSE AI automatic species recognition



K.P. Lamers
M. van Leeuwen
C. Eikenaar

WAARDEN BURG Ecology

DUTCH GOVERNMENTAL OFFSHORE WIND ECOLOGICAL PROGRAMME

we consult nature.

Validation of the MUSE AI automatic species recognition

Results of the performance in wind farm Luchterduinen

K.P. Lamers, M. van Leeuwen, C. Eikenaar

Status: Final report

Report nr:	25-251
Project nr:	23-0466
Date of publication:	12 August 2025
Photo credits cover page:	DHI
Project manager:	Koosje Lamers, MSc.
Second reader:	dr. Abel Gyimesi
Name & address client:	Rijkswaterstaat Zuiderwagenplein 2 8224 AD Lelystad Postbus 2232 3500 GE Utrecht
Reference client:	31207062/Oldert
Signed for publication:	Team Manager Waardenburg Ecology dr. A. Gyimesi

Please cite as: Lamers, K.P., van Leeuwen, M., Eikenaar, C. 2025. Validation of the MUSE AI automatic species recognition. Report 25-251. Waardenburg Ecology, Culemborg.

Keywords: DHI, MUSE, automatic species recognition, validation, sea birds, collision risk, offshore wind farm, Luchterduinen

Waardenburg Ecology Varkensmarkt 9, 4101 CK Culemborg, 0345 512710
info@waardenburg.eco, www.waardenburg.eco

WAARDEN BURG Ecology

Preface

This report describes an ornithological validation study of the MUSE automatic species recognition algorithm, using videos collected at Windpark Luchterduinen in the months July 2024 – March 2025. This document details the methodology used by the observers validating the videos, presents the results of this work, lists possible improvements that could be made to the camera system and software, and provides discussion of the validation results.

The work was commissioned by Rijkswaterstaat (RWS) and contracted to Waardenburg Ecology (WE). The work was completed in collaboration with the Danish Hydraulic Institute (DHI), who provided the videos, the annotation tool, and the matched dataset with the MUSE AI species recognition results and validated observations.

The following persons participated in the work described in this report:

Cas Eikenaar	Video annotation and reporting
Mark van Leeuwen	Video annotation
Koosje Lamers	Project management, analyses, reporting
Abel Gyimesi	Quality control

We are grateful to the team at DHI, and in particular Jannie Fries Linnebjerg (DHI), for their support. We thank Jos de Visser and Henri Zomer (Rijkswaterstaat) for the smooth collaboration in this project.

Contents

1 Introduction

1.1 Background

The Dutch government has set climate goals for 2050, which include extensive plans to scale up offshore wind energy. This process requires knowledge about the ecological effects of the construction and operation of large-scale wind farms in the North Sea, particularly the risk for seabirds of collisions with these wind farms. Hence, identifying which bird species are present in and around offshore wind farms and in what numbers, is crucial information to better determine and assess collision risk. Until recently, monitoring of bird collisions in offshore wind farms hardly took place, in part due to the logistic challenges. In the past few years however, advancements in camera systems have made automatic monitoring possible and have been adopted in multiple wind farms.

At the offshore wind farm Luchterduinen in the Netherlands the MUSE monitoring system linking a radar to four cameras has been installed. The radar detects birds and the MUSE system communicates the positions to the cameras which then film the birds. This system was part of the now completed Monitoring and Evaluation Programme (LUD-MEP) research project by Eneco and carried out by the Danish Hydraulic Institute (DHI). During the evaluation of the LUD-MEP study (Skov & Tjørnløv, 2022), it was concluded that the sample size of camera observations should be increased and the quality of the videos improved, in order to provide more realiable measurements of collisions and avoidance rates. Rijkswaterstaat (RWS) has recently taken ownership of the MUSE system and in 2023/2024 the cameras were replaced by DHI by optically higher performing models to obtain images of better quality. The measurements from these four cameras, in combination with the connected bird radar, can be used to further fill the knowledge gap about collisions and avoidance behavior.

Waardenburg Ecology (WE) was previously contracted by RWS to collaborate with DHI in a Wozep bird research project in wind farm Luchterduinen. One goal of the project was to have ornithologists analyze the camera images from DHI to classify the recorded birds by species. Due to the delays in the overarching project, this part was never carried out. After the LUD-MEP study, DHI trained an artificial intelligence algorithm for automated species recognition of the birds recorded by the cameras. Due to this recent AI development, manual species identification may possibly no longer be necessary. RWS has therefore requested WE to perform this research project to validate DHI's AI species recognition software. These validation measurements by WE can then be used by DHI to make species recognition more accurate, and thereby aid in making the species-specific avoidance and collision data more reliable.

1.2 Automatic species recognition

Bird radars can provide detailed data on the flight patterns of birds in and around the wind farm, but they cannot recognize species. Wind farm Luchterduinen has multiple radars, but in order to gather species-specific information of bird presence (and hence collision risk) in the wind farm and behavioural responses to the turbines (i.e. avoidance behaviour), RWS contracted DHI to install cameras. Hence, DHI's MUltiSEnsor bird detection system (MUSE) was deployed in 2023 and commissioned on July 1st 2024. The MUSE system is a software program that connects a number of pan-tilt-zoom cameras and radars, where cameras are instructed to turn towards the direction of a flying bird, focus, zoom in, and track the birds independently using an object detection algorithm. Recent further developments of the MUSE system have added an automatic species recognition algorithm using AI to perform species classifications. This will allow species-specific radar tracks to be collected.

With the MUSE system and its new automatic species recognition algorithm, radar tracks are thus linked to videos, allowing large amounts of species-specific information to be collected. Its AI algorithm was developed and trained with a dataset of video images in which the species had been identified by seabird experts. However, to ascertain that the AI species recognition reaches a sufficient accuracy level when applied in practice, the automatic classifications also need to subsequently be validated by an independent party. The project described in this report aims to do precisely this: it is a validation study of the automatic species classifications by the MUSE AI algorithm.

1.3 Research approach

For this purpose, DHI supplied WE with images collected by the four cameras in Luchterduinen during the period July 2024 – March 2025. WE ornithologists with extensive seabird determination experience selected a cross-section of the video footage distributed across different months and cameras, and reviewed these in the DHI validation tool to identify the birds present in these videos to species level. Subsequently, DHI performed the automatic species recognition with the MUSE software on the manually reviewed images. This matched species classification dataset (AI classification conclusion versus the classification of WE bird ecologists) were analyzed to assess the accuracy of the automatic recognition and to evaluate effects of seasonal patterns and weather conditions.

Chapter 2 documents the video selection, annotation procedure and describes the performance analysis. The results of the validation study are presented in Chapter 3. These are further evaluated and discussed in Chapter 4 and recommendations are made to help further improve the algorithm and the quality of the camera measurements. Finally, Chapter 5 provides a brief summary of our conclusions.

2 Materials and methods

2.1 Camera systems and recognition software

2.1.1 Wind farm Luchterduinen

This validation study was carried out using video images collected by the MUSE monitoring system at wind farm Luchterduinen. The wind farm constitutes 43 wind turbines on the North Sea, off the Dutch coast, west of Zandvoort. This wind farm includes two bird radars, of which the one on the centrally located Offshore High Voltage Station (OHVS; Figure 2.1) is part of MUSE system and used to activate the four cameras. The most commonly observed birds at this wind farm (as per an earlier WE study) are: lesser black-backed gulls, great cormorants, black-legged kittiwakes, northern gannets, species of the *alcid* family (among them common guillemots), greater black-backed gulls, mew gulls (a.k.a. common gulls), songbirds and herring gulls (Leemans *et al*. 2022).

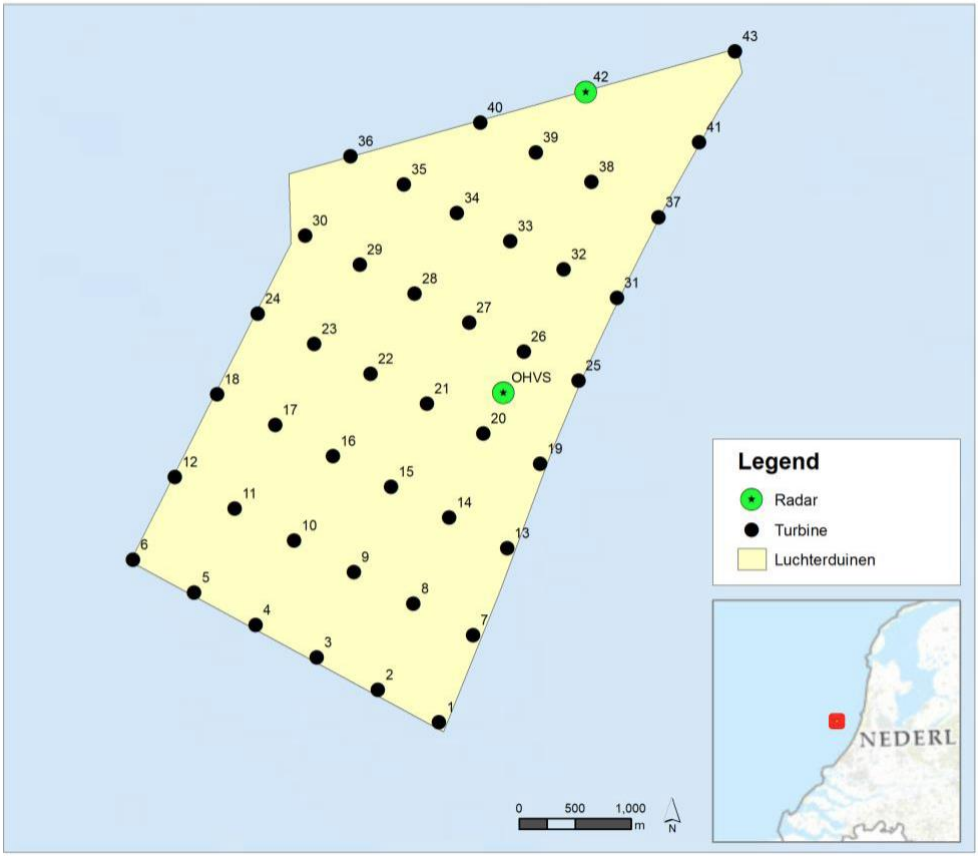


Figure 2.1 *Location and layout of wind farm Luchterduinen. The bird radar on the Offshore High Voltage Station (OHVS) station is used to activate the cameras.*

2.1.2 Camera systems of the MUSE monitoring system

The MUSE system constitutes a fully integrated monitoring system of one horizontal radar (Furuno FAR-3000 S-band) on the OHVS platform and four daylight pan-tilt cameras (FLIR M300C; Figure 2.2). The cameras are positioned on four wind turbine platforms spread throughout the wind farm (Figure 2.3).

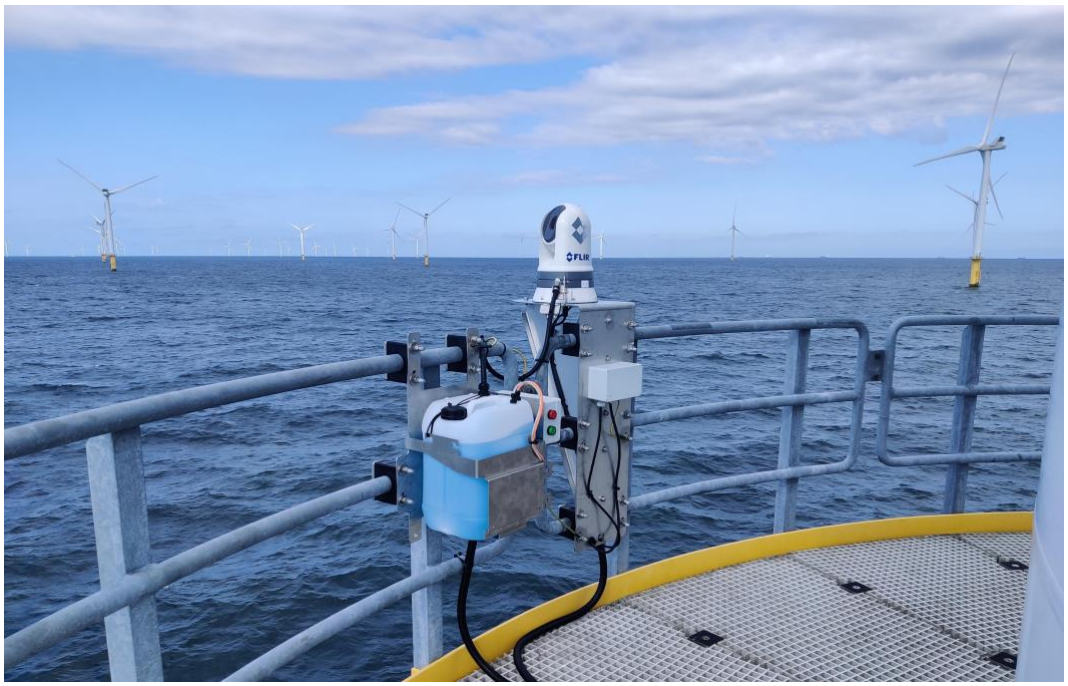


Figure 2.2 *One of the four daylight vision cameras, positioned on a wind turbine platform. Photo taken by DHI © DHI MUSE A/S.*

The horizontal radar has an effective range of approximately 3 kilometers, allowing automated scanning for birds within the 1 kilometer ranges of the cameras (Figure 2.3). Dynamic sea and rain clutter filters are used by the radar to prevent false positive detections of non-birds, such as high waves. This potentially reduces the detection probability of birds, but will not affect this validation study, as the goal is to study the species classifications of detected birds (detection probability is not assessed). Once the radar detects a bird, the DHI MUSE ('multi-sensor') system instructs the most relevant camera to target the bird. The camera, using object detection software in a separate system component, then independently tracks and films the bird.

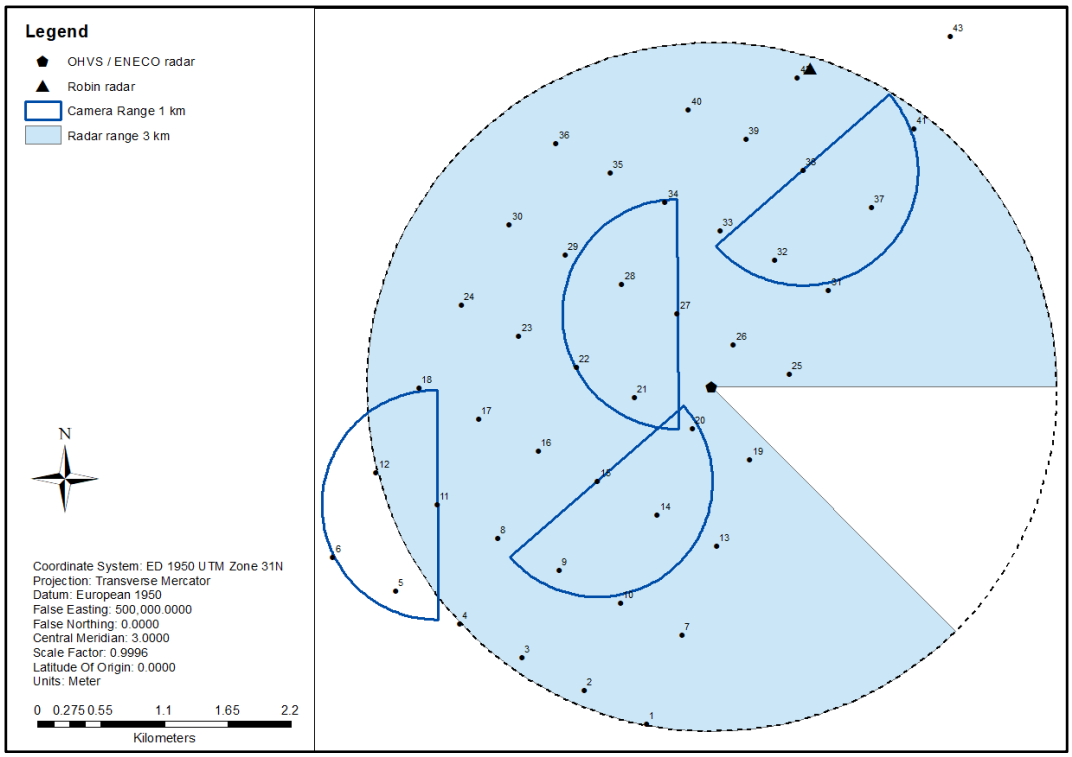


Figure 2.3 *Location and range of the four camera systems within Wind farm Luchterduinen as presented in the LUD MEP project report (Skov & Tjørnløv, 2022).*

2.2 Validation procedure

2.2.1 Video selection

In order to create a representative sample of the birds present and the AI species recognition over the seasons and under different light and weather circumstances, we selected videos by taking a cross section by choosing days spread evenly over the months and scoring all videos taken that same day at all cameras. The annotated videos spanned the months July 2024 to March 2025, though there were few videos in July, as data was only available for the first week.

2.2.2 Annotation procedure

Videos were supplied by DHI and uploaded into the MUSE Annotation tool (Figure 2.4). The annotation tool generates bird tracks from the video and allows the observer to annotate the species of all birds present in the video.

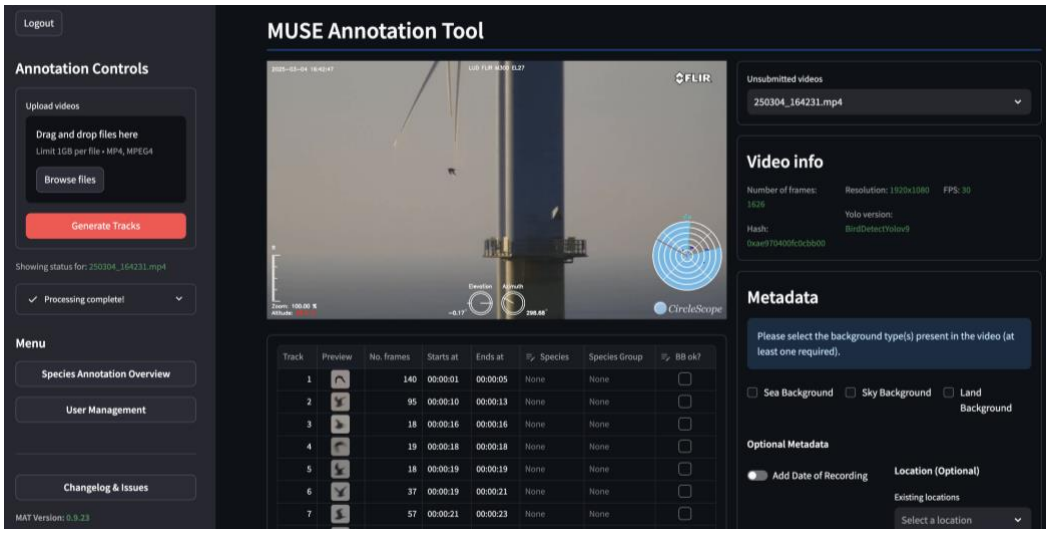


Figure 2.4 View of a video being annotated in the MUSE annotation tool.

Videos of birds consisted of one or more consecutively numbered bird tracks. In the videos in the annotation tool, numbered 'bounding boxes' were placed around the bird being tracked by the recognition software to indicate which object/bird to annotate. These bounding boxes actually often made species recognition more difficult, because these regularly block the bird's wing tips and bill (Figure 2.5) and these are features of importance for seabird identification (especially in gulls). As a consequence, the observers in practice usually first watched the videos in a different program from the MUSE Annotation tool.

Figure 2.5 Example of how the bounding box may obscure part of the body of the bird.

Each of the bird tracks (separately named bounding boxes) had to be categorized for the species name if possible. Most videos contained a single individual bird, but some contained multiple individuals, sometimes of different species. When all bounding boxes (a.k.a. bird tracks) had been categorized, metadata of video background were always submitted for the sky condition (i.e. cloud cover), and optionally for the sea state and/or land conditions (if applicable). If the bird track could not be classified to species level by the observer, it was assigned a broader (species) category (see Figure 3.1 for all species

and species-category names assigned by the observers). Figure 2.6 shows examples of images of birds as they typically occur in the videos. In some cases, the software placed bounding boxes around an object, such as the railing of a wind turbine platform. These were scored as the category 'non-bird' and omitted from analyses. All annotated species were also categorized to species groups by the MUSE annotation tool and software, grouping them into for instance large gull, small gull, gannets, cormorants, passerines, and so forth (see Table *3.2* for group names).

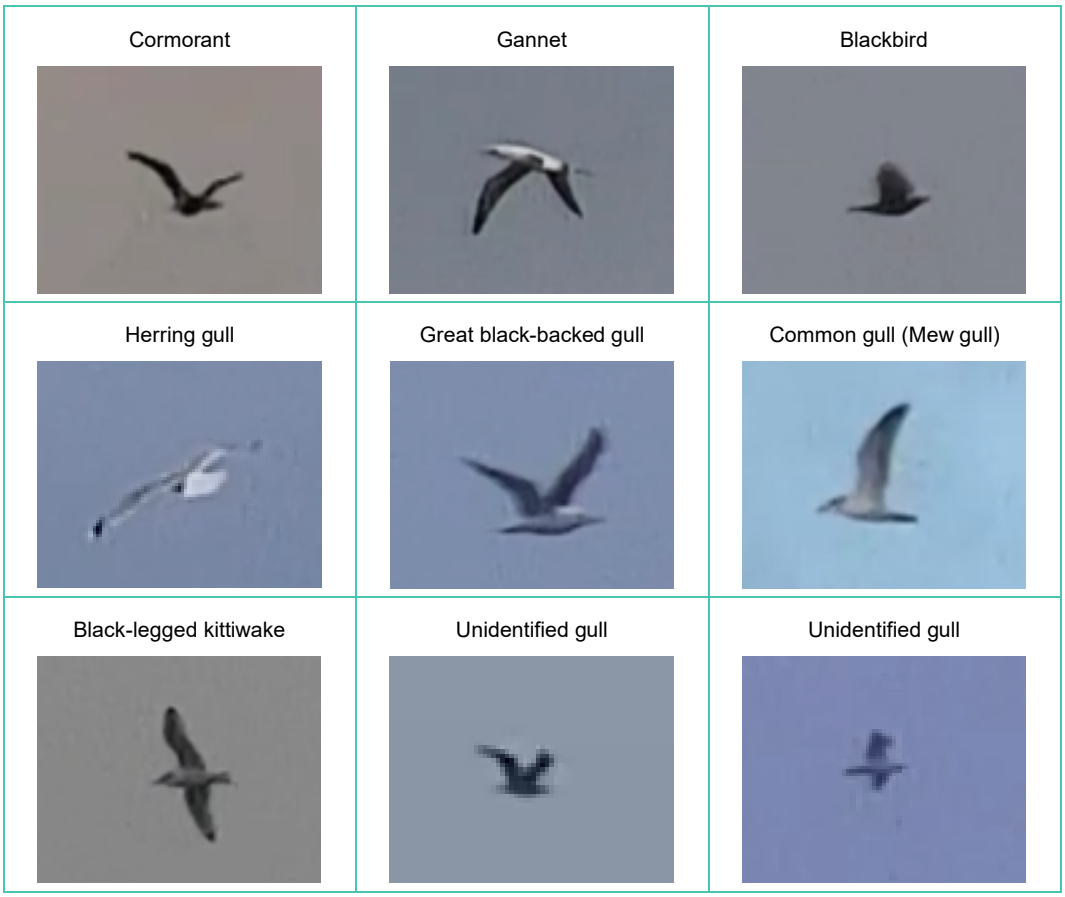


Figure 2.6 *Examples of images (bounding boxes removed for clarity) taken by the cameras and annotated in this study, showing birds that could be classified to species level and two typical images of gulls that could not be further identified by the observers.*

2.3 Data analysis

2.3.1 Data selection

Once annotation was completed, DHI used the MUSE AI to predict the species of the annotated birds. The matched dataset with the annotation results of the WE observers and the prediction conclusions were shared with WE. The MUSE AI predictions consisted of: a predicted species, predicted group and category weights for the predicted species and for the predicted group. The final dataset was comprised of 9,088 bird tracks, based on 633 videos, with both an annotated and predicted result for species and species group. As

unique birds could have been detected multiple times, this can lead to multiple tracks per bird. Analyses were performed using the bird tracks as individual data points. This was chosen because the species recognition software was run at the track level, we had no way of determining which bird was detected multiple times and which was not, and the observers also checked all the individual tracks. Note however that since videos contain varying numbers of tracks, birds with more detections are weighted more heavily in the performance metrics. The analyses evaluating the performance of the species recognition software were performed at both the species level and at the group level.

As some bird tracks could not be identified to species level by the human observers and hence were annotated as a broader unspecified category such as 'lesser or greater black-backed gull' or 'unidentified small gull'. In these cases when a true single-species or single-group annotation was lacking, it was impossible to ascertain whether the AI species recognition software's conclusion was correct. A match between the annotation result and the predicted result would not be possible for such bird tracks lacking a single species- or group-level annotation, because the MUSE algorithm is only trained to recognize individual species and species groups (not vague categories such as: 'unidentified gull' or 'lesser or greater black-backed gull', or group categories such as 'gulls'). In short, if the human observers could not determine the species or the species group of a bird track, it was impossible to test whether the AI algorithm could correctly classify these tracks. Hence, bird tracks lacking a single true species annotation or true group annotation were labelled as 'not-predictable' by DHI and removed for analyses of the AI algorithm's performance parameters. Analyses relating to the species levels were thus limited to the bird tracks for which the annotators had assigned a single annotated species name (n = 5,691) and analyses at the group-level were limited to bird tracks for which the annotated species fit into a distinct group classification (n = 8,080).

2.3.2 Performance parameters

Overall accuracy (at the species level and the group level) was calculated by dividing the number of correct predictions by the total number of predictions. Precision was calculated separately for each species and group category, as the number of tracks correctly predicted divided by the total number of tracks predicted to belong to that particular species (or group) category. Moreover, confusion matrices were generated for the species level and the group level, to show in detail how the model classifies and misclassifies the different categories.

Note that the aforementioned removal of 'not-predictable' bird tracks could have affected the performance results of the species recognition software. When the ornithologists annotating the videos were unable to identify a bird track to species level with certainty, this was likely due to the bird being a great distance from the camera or due to poor image quality (e.g. backlighting causing the bird to appear as a silhouette). As a consequence, the bird tracks used in the analyses potentially originated from video footage that was of above-average quality. Therefore, the accuracy and precision rates reported here are only representative for the situation that human observers can identify the birds. The prediction accuracy and precision beyond what humans can identify is unknown. Presumably it is

lower than for birds that can be identified by humans, due to less information being available upon which the classification can be based.

2.3.3 Seasonal patterns and weather conditions

To evaluate the effects of weather conditions and seasonal patterns, accuracy rates were compared between the different seasons and weather conditions. Since sky condition was collected as additional metadata for all videos and the observers annotating the videos noted this as of particular importance, this was analyzed in further detail. A generalized linear model was fitted to statistically test the effect of the sky condition (clear, clouds in frame, or overcast) on species recognition accuracy. This was performed at the level of the video (n = 392), to prevent potential pseudoreplication in statistical testing. For this, we modelled the proportion of 'true' (= predicted and observed classification matched) bird tracks per video as the binomial response variable (with a logit link function) and fitted the sky condition as explanatory variable. This model was compared with a null model (without explanatory variable), and when found significant, a post-hoc test was performed to compare the three different sky condition categories.

3 Results

3.1 General overview of the data

3.1.1 General observations during annotation

During the annotation process, the WE ornithologists that scored the videos observed that successful species identification clearly depends on multiple factors. Distance is important, with the probability of being able to identify the bird to species level decreasing with distance from the camera. This is especially the case for smaller birds; a gannet can be more easily recognized at a large distance than a kittiwake. Body size in itself appears not to be an issue; one video contained footage of common blackbirds, a passerine bird much smaller than most seabirds. Videos 'looking' directly into the sun result in silhouettes only, making species identification harder. Similarly, overcast weather diminishes colour in the videos. Strong winds make the camera shake, again obstructing identification. The influence of these factors is not the same for all species. Some species, such as gannets, are easily identified, even only as a silhouette at a large distance. The various gull species pose most difficulties, not only because of their similarities, but also because body size is often hard to judge in a video without reference to an object of known size.

Importantly, the radar and therefore camera appear to fail to pick up birds flying low over the surface; at least, virtually all annotated videos start with a bird against a sky background. This means that species such as razorbill and common guillemot, which typically fly just above sea level, are very unlikely to be detected and filmed. Therefore, such species, given they are present in the study area (Leemans *et al*. 2022), almost certainly are underrepresented in the videos (also see §3.1.3). This can be explained due to the fact that the radar in this project is a 2D radar, which has a default verticle angle and is more likely to pick up objects within a certain altitude range.

3.1.2 Data overview

A total of 633 videos spanning the months July 2024 – March 2025, each containing one or multiple tracks, were annotated by the two observers. This selection of annotated videos contained a total number of 9,088 tracks. The number or birds present is likely lower, as birds can be detected multiple times and thereby split into multiple tracks. The analysis was performed using the individual bird tracks and tracks for which the species or group was uncertain were not included in the respective analysis (see §2.3.1 for further explanation). Analyses relating to the species level were thus limited to bird tracks for which the observers had annotated a single species (n = 5,691) and analyses at the group level to bird tracks for which the annotated species fit into a single group classification (n = 8,080).

3.1.3 Species composition

The tracks from the annotated videos revealed that the vast majority of birds were small gulls (such as the mew gull, a.k.a. common gull, and black-legged kittiwake), cormorants, large gulls (great black-backed gull, herring gull and lesser black-backed gull), and gannets (Figure 3.1). In 2.5% of cases (225 tracks), tracks were scored by the observers to belong to non-bird objects such as a wind turbine railing or platform. This 'not a bird' category was henceforth removed from the data. Out of all tracks that were actual birds, a total of 69% could be classified to species level by the observers. The rest of the bird tracks in the videos (31%) could not be classified to species level with enough certainty. This was due to a combination of the image quality, a large distance of birds relative to the video camera, the difficulty of assessing bird size from video footage, and the subtle differences in gull plumage characteristics (Figure 2.6). These tracks were therefore classified to the broader species categories: 'unidentified small gull', 'unidentified gull', 'unidentified large gull', 'lesser or greater black-backed gull', 'unidentified goose', 'herring gull or lesser black-backed gull', 'mew or herring gull', 'unidentified seabird' and 'not identifiable'. The species composition was strongly affected by the month March, when large numbers of videos and bird tracks were recorded and the observers annotated many mew gulls and unidentified small gulls (potentially also mew gulls). In all other months combined, the most common species (and species groups) were in order of frequency: black-legged kittiwake, great cormorant, great black-backed gull, herring gull, northern gannet, and unidentified gull.

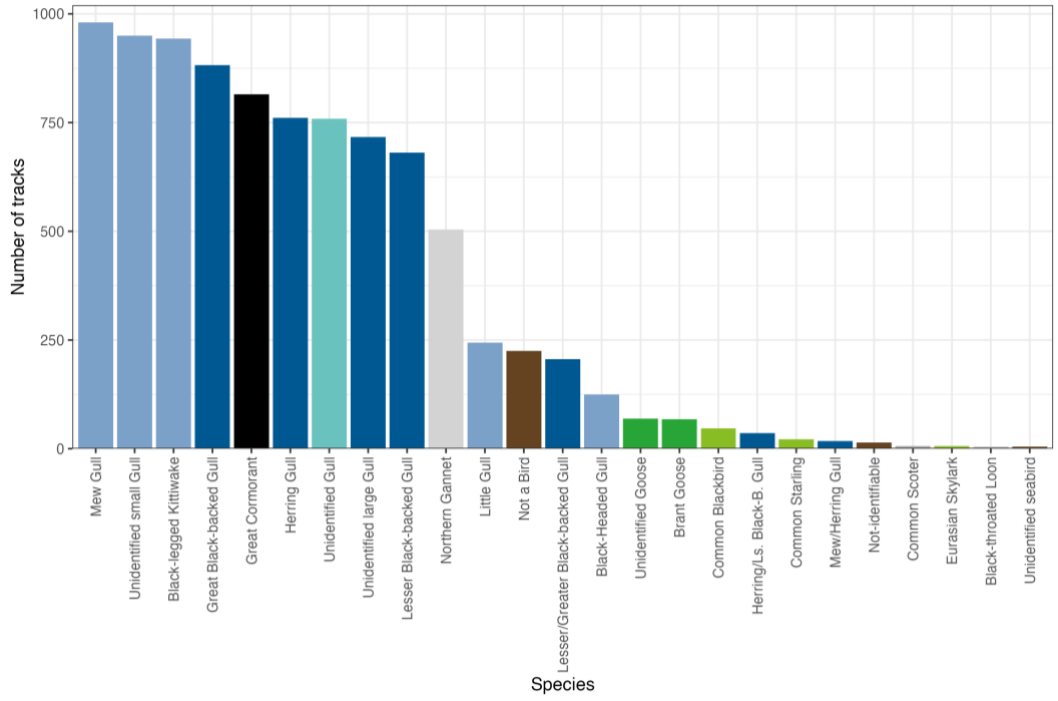


Figure 3.1 *Species composition of the tracks in the annotated videos recorded at wind farm Luchterduinen in July 2024 – March 2025, as scored by two bird ecologists. Bar colours represent species groups: small gulls (light blue), large gulls (dark blue), cormorants (black), other gulls of uncertain size class (turquoise), gannets (light grey), not a bird or not identifiable (brown), geese (green), passerines (light green), ducks (darkgrey) and divers (grey). Mew gulls are also known as common gulls.*

The species composition presented above largely matches the results of earlier visual counts (§2.1.1), though lesser black-backed gulls were less common than in Leemans *et al.* (2022). One important difference is that razorbills and common guillemots were never observed in the video dataset by the ornithologists, whereas these species are known to be present at wind farm Luchterduinen. This supports the observation that detection by the MUSE system mostly occured against a sky background and was very limited for low-flying species that typically only occur just above the sea level (§3.1.1).

3.2 Evaluation of species classification performance of the MUSE software

3.2.1 Accuracy

The overall accuracy, indicating what percentage of bird tracks was classified correctly by the MUSE algorithm, was 34% at the species level and 58% at the group level. The species group (Small Gulls, Large Gulls, Cormorants etc.) is thus better recognized, which may be due to how similar some of these species are. However, overall accuracy can be strongly affected by common species that are predicted very poorly or very well, hence it is most useful to evaluate for each species (and group) separately how accurately it is recognized.

3.2.2 Confusion matrix

How the MUSE automatic species recognition software performs separately for each species was analyzed with a confusion matrix (Table 3.1). This shows how often species are correctly recognized and how species are being misclassified. Since a confusion matrix can only be made for species which were observed by the annotators and which the model was trained to recognize, only species for which the model was trained are shown. This restricts our analysis to tracks annotated as: mew gull, black-legged kittiwake, great black-backed gull, great cormorant, herring gull, lesser black-backed gull, northern gannet, and black-headed gull. The model was not trained to identify the songbirds, geese, ducks and divers which were scored by the observers. Also, it is important to note here that the little gull was present at the site with reasonable frequency (3%; Figure 3.1), but the model was not trained to recognize this species. This is a limitation of relevance for wind farm Luchterduinen, as it means that little gulls are never recognized.

Herring gulls and great cormorants were overall well-recognized by the MUSE software with an accuracy of respectively 77% and 55% (Table 3.1). The three most common species at the site (mew gull, black-legged kittiwake, and great black-backed gull), were recognized with low accuracy (< 30%). This will have negatively affected the overall accuracy at the species level reported in the prior paragraph. Interestingly, the model appears to not be well-trained to recognize lesser black-backed gulls (0% accuracy). On the other hand, it exceptionally often classified birds as herring gulls. In fact, just over 50% of all bird tracks were classified as herring gulls (2,880 out of 5,691). This led to many false classifications as herring gull for all other gull species (> 40%). Finally, it is interesting to note that multiple non-marine species (i.e. kites, corvids, pigeons) which were never seen by the ornithologists were in fact predicted by the algorithm, albeit rarely.

Table 3.1 Confusion matrix showing the species names of bird tracks annotated by the observers (columns) and the MUSE AI predictions (rows). Green-highlighted cells on the diagonal indicate correct predictions (and % accurate). The percentages in brackets are rounded and add up to ±100% over the column length, but are omitted if not greater than 0%. Red text colour indicates species that the software frequently (> 25%) misclassifies the species of interest to.

Predicted (MUSE AI species recognition) \ Observed (annotated by observers)	Mew Gull	Black-legged Kittiwake	Gr. Black-backed gull	Great Cormorant	Herring Gull	Ls. Black-backed Gull	Northern Gannet	Black-Headed Gull
Mew Gull	**197 (20%)**	128 (14%)	58 (7%)	9 (1%)	61 (8%)	88 (13%)	2	6 (5%)
Black-legged Kittiwake	29 (3%)	**272 (29%)**	1	7 (1%)	26 (3%)	0	5 (1%)	13 (10%)
Gr. Black-backed Gull	30 (3%)	41 (4%)	**205 (23%)**	2	61 (8%)	105 (15%)	27 (5%)	0
Great Cormorant	4	7 (1%)	4	**449 (55%)**	0	1	9 (2%)	0
Herring Gull	**649 (66%)**	**383 (41%)**	**495 (56%)**	58 (7%)	**585 (77%)**	**444 (65%)**	**193 (38%)**	**73 (58%)**
Ls. Black-backed Gull	0	0	1	0	0	**1 (0%)**	0	0
Northern Gannet	1	3	1	2	0	0	**200 (40%)**	0
Black-Headed Gull	29 (3%)	23 (2%)	9 (1%)	0	2	10 (1%)	2	**26 (21%)**
Barnacle Goose	0	0	1	1	0	0	0	0
Black Kite	0	2	1	4	1	2	0	2 (2%)
Common Buzzard	1	17 (2%)	1	7 (1%)	0	1	2	0
Common Crane	0	0	3	24 (3%)	0	0	14 (3%)	0
Common Eider	0	0	1	16 (2%)	0	0	0	0
Common Goldeneye	0	0	0	0	1	0	1	0
Common Kestrel	0	0	0	5 (1%)	0	0	1	0
Wood Pigeon	0	0	1	2	0	0	0	0
Eurasian Hobby	0	1	0	0	0	1	0	0
Sparrowhawk	3	5 (1%)	6 (1%)	28 (3%)	0	1	3 (1%)	0
Honey Buzzard	0	0	0	9 (1%)	0	0	0	1 (1%)
Golden Eagle	0	0	1	1	0	0	0	0
Hooded Crow	2	0	3	2	1	0	1	0
Mallard	0	0	0	2	0	0	1	0
Northern Harrier	1	1	0	0	0	0	0	0
Northern Raven	0	6 (1%)	1	2	0	1	1	0
Osprey	0	0	0	1	0	2	0	0
Red Kite	3	5 (1%)	3	27 (3%)	0	3	3 (1%)	0
Red-breast. Merganser	0	0	0	0	0	0	3 (1%)	0
Rook	2	2	0	4	0	1	0	1 (1%)
Rough-legged Buzzard	0	0	0	1	0	0	0	0
Sandwich Tern	0	7 (1%)	1	0	0	0	8 (2%)	2 (2%)
Western Jackdaw	0	0	0	3	0	0	0	0
Marsh Harrier	0	0	0	22 (3%)	0	0	0	0
White-tailed Eagle	3	1	24 (3%)	87 (11%)	0	0	7 (1%)	0
Yellow-legged Gull	26 (3%)	39 (4%)	61 (7%)	40 (5%)	23 (3%)	20 (3%)	21 (4%)	1 (1%)
Total	**980**	**943**	**882**	**815**	**761**	**681**	**504**	**125**

The confusion matrix at the group level shows better overall results (Table 3.2). Distinguishing between the large and small gull groups however presents difficulties, as small gulls were often classified as large gulls (58%). This may indicate that the AI model currently may not be well able to judge size from the video footage. Overall, the group-level classifications appeared biased towards classifying groups as large gulls, as many groups were confused with this category. Though the model was not trained to recognize individual songbird species, the overarching passerines group was well recognized (accuracy 81%).

Table 3.2 *Confusion matrix showing the group classifications of bird tracks annotated by the observers (columns) and the MUSE AI predictions (rows). Green-highlighted cells on the diagonal indicate correct predictions (and % accurate). The percentages in brackets are rounded and add up to ±100% over the column length, but are omitted if not greater than 0%. Red text colour indicates groups that the software frequently (> 25%) misclassifies the group of interest to.*

Predicted (MUSE AI species recogniton) \ Observed (annotated by observers)	Large Gulls	Small Gulls	Cormorants	Gannets	Geese	Passerines	Ducks
Large Gulls	**2825 (86%)**	**1888 (58%)**	104 (13%)	**259 (51%)**	**35 (26%)**	0	0
Small Gulls	377 (11%)	**1194 (37%)**	17 (2%)	13 (3%)	24 (18%)	1 (1%)	0
Cormorants	8	16	**432 (53%)**	8 (2%)	**55 (40%)**	0	**2 (33%)**
Gannets	0	3	0	**186 (37%)**	0	0	0
Geese	0	1	3	0	**11 (8%)**	0	0
Passerines	16	50 (2%)	17 (2%)	1	1 (1%)	**61 (81%)**	0
Ducks	3	2	18 (2%)	3 (1%)	8 (6%)	3 (4%)	**4 (67%)**
Raptors	66 (2%)	70 (2%)	**208 (26%)**	16 (3%)	3 (2%)	8 (11%)	0
Cranes	3	0	16 (2%)	11 (2%)	0	0	0
Terns	1	18 (1%)	0	7 (1%)	0	0	0
Doves	2	0	0	0	0	2 (3%)	0
Total	**3301**	**3242**	**815**	**504**	**137**	**75**	**6**

3.2.3 Precision per species and per group

Though accuracy is important to ascertain whether species and groups are being correctly recognized, the ratio of true predictions to total predictions (the precision) per category is equally important. For example, when species-specific radar tracks generated by automatic species recognition are used to monitor great cormorant collision risk, it is of paramount importance that other species are not falsely being classified as great cormorants at a high rate. Such false positive classifications would otherwise obscure the true collision risk or flight behaviour of the species of interest. Therefore, we calculated the precision rate at the species level and the group level (Table 3.3).

The precision values show that the predictions for some species and group categories have a high precision rate: 95% of birds classified as great cormorants and 97% of birds classified as northern gannets were in fact those species (Table 3.3). The precision of the

classifications of the various gull species was lower. As seen earlier in the confusion matrix (Table 3.1), particularly many birds are falsely being classified as herring gulls: only 20% of birds classified as herring gulls were also annotated as herring gulls by the ornithologists.

Table 3.3 *Precision of the classifications of the MUSE AI species recognition algorithm for the various species and groups. Precision is calculated as the number of bird tracks correctly being predicted to belong to a certain category (species or group), divided by the total number of bird tracks predicted to belong to that same category (hence including false classifications). High percentages indicate that nearly all birds being classified as this species are correctly classified. Low values signal that many bird tracks are falsely being classified as this category.*

Category	Number of correct classifications	Total number of classifications (true and false)	Precision (%)
Species level			
Mew Gull	197	549	**36%**
Black-legged Kittiwake	272	353	**77%**
Gr. Black-backed gull	205	471	**44%**
Great Cormorant	449	474	**95%**
Herring Gull	585	2880	**20%**
Ls. Black-backed Gull	1	2	**50%**
Northern Gannet	200	207	**97%**
Black-Headed Gull	26	101	**26%**
Group level			
Large Gulls	2825	5111	**55%**
Small Gulls	1194	1626	**73%**
Cormorants	432	521	**83%**
Gannets	186	189	**98%**
Geese	11	15	**73%**
Passerines	61	146	**42%**
Ducks	4	41	**10%**

The species and groups which were not annotated by the observers, but for which the model was trained and generated predictions for (i.e. the groups raptors, cranes, terns and doves; and all species in the lower half of Table 3.1: barnacle goose up to yellow-legged gull), are not shown in this table of precision values. However, because they were never predicted correctly (these species and groups were not found to be present in the videos), the precision of classifications for these species and groups is 0%.

3.3 Seasonal and weather effects

3.3.1 Seasonal effects on accuracy

Due to the strong variations in species composition between the months and seasons and the large differences between species in how well they were recognized, performing fair comparisons between the seasons in how well the software performed was difficult. Overall accuracy of the species classifications varies between the months and seasons, but this would likely be mostly attributable to the fact that the MUSE AI algorithm performed very differently for different species. For instance, mew gulls, which were relatively poorly recognized by the software (Table 3.1), were extremely common in March but hardly present in other months. Likewise, lesser black-backed gulls were rarely recognized correctly and are not present in the winter due to their migratory nature, hence increasing overall winter species recognition accuracy. Therefore, we limited the analysis of the seasonal effects to a brief inspection of the accuracy of the species classifications for two species that were present at the site during the majority of the year (great cormorant and great black-backed gull). Overall, no uniform seasonal difference was apparent for how well great cormorants and great black-backed gulls were recognized by the software.

Importantly, since our dataset does not span the full year (several months in the breeding season are lacking: April, May and June), some species are likely underrepresented. This is almost certainly the case for the lesser black-backed gull. This summer visitor to the North Sea was the most commonly observed species in visual counts (Leemans *et al.* 2022), but trailed behind many other gull species in our video dataset. The lack of data from the breeding season (video collection started 1 July 2024) would likely affect the overall accuracy score, as lesser black-backed gulls were poorly recognized (0% accuracy; §3.2.2). Further study using video data spanning a full year, is therefore recommended to be able to better assess species recognition performance over the seasons, and in the critical breeding season.

3.3.2 Effect of weather conditions

As already noted by the observers during the annotation process (§3.1.1), the sky conditions affect the ease of classifying bird tracks on videos to species level: sunny conditions for example can render a video image of a bird to just a colourless silhouette. Not only will this affect the ease of classifying videos for human observers, but it may also influence the accuracy of classifications made by the AI species recognition software. This appeared to potentially be the case: accuracy of the species recognition was 31% for bird tracks recorded during clear skies, 32% when clouds were present in the frame, and 43% for overcast conditions. A statistical test (generalized linear model, fitted at the video level, followed by a post-hoc test) indeed confirmed that bird tracks from videos in overcast conditions were classified with a significantly higher accuracy than bird tracks from videos taken when skies were clear or cloudy.

Other weather variables, such as wave height (a.k.a. sea condition), are more likely to affect detection probability rather than the accuracy of species classifications. Hence, as

we only have data on species classifications and not on detection probability and not all videos had a known sea state, the effect of sea condition was not investigated.

3.4 Effect of hardware update

Prior to the old (Rvision) daylight cameras being replaced with the current FLIR M300C cameras, the number of gulls that could not be identified to the level of the species by observers screening the videos was high. Summarizing the numbers reported in the LUD-MEP study (Table 3 in Skov & Tjørnløv, 2022) shows that 68,5% of all video data was not annotated to a single species but instead to various species categories (unidentified large gull, great/lesser black-backed gull, unidentified seabird, etc.). This contrasts with this study, in which only 31% of bird tracks could not be identified to a single species conclusion (§3.1.3). These strongly improved numbers could indicate improved quality of the video footage and/or differences between observers scoring the videos. Given the better technical specifications of the new cameras, it is expected that the hardware update resulted in better image quality and thereby enabled our observers to more often distinguish species-specific morphological traits.

4 Discussion

4.1 Video footage quality

The video footage collected by the MUSE system at wind farm Luchterduinen has previously only been used to manually score the species of the recorded birds. In a previous study by DHI, where ornithologist scored the videos, only 31% of video data was classified to a single species (the majority were labelled as categories such as 'unidentified gull'). Since then, the cameras have been updated to better quality models and this has coincided with an improved ability of ecologists scoring the videos to identify the species (69%). Though this is no conclusive proof, it appears significant strides have been made in video quality by updating the hardware. This is an important first step in facilitating succesful automatic species recognition software using AI.

4.2 Performance of the recognition software

In the sample of video data scored and analyzed in this study, the overall accuracy (= the percentage of bird tracks of that species or group recognized correctly) of the MUSE species recognition software was 34% for species level classifications and 58% for group level classifications. Accuracy varied greatly per species and group: herring gull and great cormorant were recognized successfully more than 50% of the time. On the other hand, mew gull, greater black-backed gull, lesser back-backed gull and black-headed gull were recognized correctly for less than 25% of bird tracks. At the group level, accuracy was high for large gulls and passerines (>80%) and particularly low (< 50%) for small gulls, gannets and geese. Precision (the percentage of bird tracks classified as a species that indeed belonged to that species of interest) also varied widely: many bird tracks were wrongly classified as herring gulls (precision 20%), whereas classifications of northern gannets by the model were almost always applied to true northern gannet tracks (97%). At the group level, precision was particularly high for gannets and cormorants (>80%), but low for passerines and ducks (<50%).

However, accuracy and precision need to be combined in order to establish whether the model performs well at recognizing a certain species or group. As a result, the great cormorant is the only species recognized with an accuracy of 50% or higher and a precision of 50% or higher. At the group level, this was only the case for large gulls and cormorants. Though no specific target was set for how well the species recognition software should perform, these results likely do not yet allow data collected from fully automated species recognition to be used to reliably monitor bird behaviour in this wind farm. That said, multiple improvements have been identified that can be used to further optimize the model and its effectiveness for this site (Chapter 5).

4.3 Considerations for using automatic species recognition

A number of observations were made during this study relating to how the MUSE recognition software operates and what data is and is not collected. Below, we outline some of these observations and considerations that should be taken into account.

Uncertainty in species identification and video quality

It is important to note that our performance analysis only included bird tracks that had been identified by observers as belonging to a specific species or group. Tracks with uncertain species classifications —often due to poor image quality, lighting, visibility, or distance from the camera—were excluded. Hence, the omission of these bird tracks of uncertain species may overestimate the performance of the species recognition model. Despite the fact that the hardware update improved the image quality, the footage stills varies in quality due to different weather conditions and distances relative to the camera. Not only did the observers note that backlighting (causing silhouettes) affected their ability to identify a bird, our analysis also confirmed that sky conditions had an effect. Though further research on this topic, using additional videos (spanning additional dates and weather conditions) and data extracted from nearby weather stations for e.g. precipitation and sight conditions (e.g. mist) could be conducted to better identify conditions that may hamper species recognition. To improve future use cases, it may be helpful to develop a strategy for handling images of poor quality. For example, analyses using this data could be restricted to conditions that support better species recognition (e.g. favorable lighting showing contrast as well as colour). As an alternative or complimentary strategy, the score for the prediction may be used to exclude possibly erroneous predictions. Determining how to optimally use the prediction certainty score to exclude erroeous predictions, whilst balancing this with preserving the sample size and without creating biases in the data, could be a topic of further study.

Individual birds with multiple tracks

Birds that are detected may be split into multiple bird tracks. If no further processing is performed to reduce this track-splitting, abundance estimates could be affected. Moreover, since it is unknown whether there are differences between species in the propensity for multiple tracks to be generated for one individual, species composition estimates could be affected as well. As this validation study used the level of the bird track for many analyses, this is a limitation of our analyses and results as well.

Tailoring the model to offshore-specific species

Though the list of species and groups which the MUSE software was trained to predict is substantial, it is currently not fully tailored to this offshore wind farm. The software has the option of classifying birds as many species which are not likely to be present at this site (i.e. raptors such as black kites, golden eagles) and lacks others that are known to be present (songbirds, little gulls, several ducks, divers and geese). Tailoring the species recognition to offshore North Sea sites may be beneficial. The model could be trained to recognize additional species which are present at the site, such as the little gull. Likewise, in order to improve recognition of species of specific interest, it could be considered to exclude species which are non-existent or extremely rare at this site if confusion occurs (as happens for sparrowhawk, common crane, and white-tailed eagle).

Detection near the surface

Low-flying species, particularly those of the *alcid* family, are known to be present at wind farm Luchterduinen but were never observed in the videos scored by the ornithologists. As the MUSE system at wind farm Luchterduinen is dependent on a 2D horizontal radar, which does not provide data on the altitude of the birds, a default verticle angle is used that may bias towards birds flying in the default altitude range. Low-flying species may thus be underrepresented and we cannot evaluate how well the MUSE AI performs for classifying radar tracks of these types of seabird species. This should be considered depending on the goal in future use cases: whilst this underestimation of low-flying birds may be an issue when for example estimating overall species composition, it is unlikely to be a problem when monitoring which species occur in the rotor swept zone and are at risk of collision. Further research, using offshore counts and altitude measurements, may compare the species composition at rotor height measured in the field to the species composition in the annotated videos to validate the representativeness of the detections of the MUSE system.

5 Summary and recommendations

This validation study assessed multiple aspects of the MUSE AI species recognition algorithm, among them the effect of the hardware update, the overall performance of the species classifications, and how this depended on weather conditions. For this, a total of 633 videos were scored by two WE ornithologists in the MUSE DHI annotation tool. Subsequently, and a matched dataset of the observed annotated species and the species predictions by the MUSE algorithm was analysed. The results show that the hardware update appeared to have improved the image quality, allowing the majority of bird tracks to be identified by observers. However, the current level of accuracy (34% at the species level) and precision of the model classifications are underperforming, especially for gull species. These numbers likely do not yet allow species-specific data collected with fully automated species recognition to be used to reliably monitor bird behaviour in this wind farm. However, multiple improvements could be identified that may help optimize the model and improve its effectiveness for this site.

In order to improve the classifications by the MUSE AI species recognition software at wind farm Luchterduinen, we recommend to consider:

- adapting the species recognition to focus on offshore-occurring species
- training the model for additional species present at the site (particularly little gull)
- automatically marking videos of poor quality or birds at great distances to potentially enable exclusion of this data (depending on the research goal)
- evaluating how to use the uncertainty score to optimally exclude poor predictions from further analyses without creating biases
- using additional annotated videos to train the algorithm.

The data collected in this study (the videos annotated by our ornithologists) – though intended as a means to evaluate the recognition software – in itself provides a great resource to further train the model and improve its recognition performance in the future.

References

Skov, H., Tjørnløv, R.S., 2022. Monitoring bird collisions - meso and micro avoidance at offshore wind farm Eneco Luchterduinen. DHI Report 11821366.

Leemans, J.J., R.S.A. van Bemmelen, R.P. Middelveld, J. Kraal, E.L. Bravo Rebolledo, D. Beuker, K. Kuiper & A. Gyimesi, 2022. Bird fluxes, flight- and avoidance behaviour of birds in offshore wind farm Luchterduinen. Bureau Waardenburg Report 22-078. Bureau Waardenburg, Culemborg.