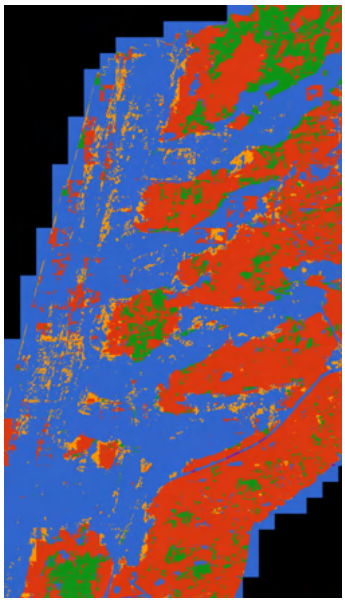
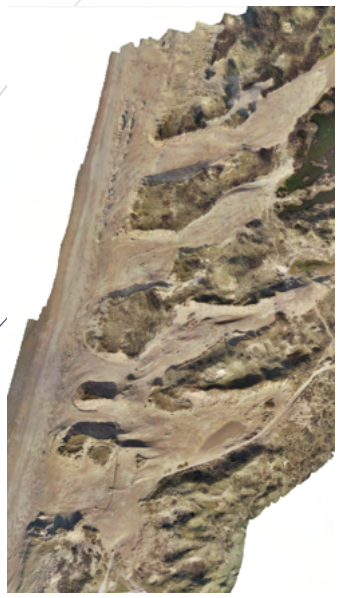
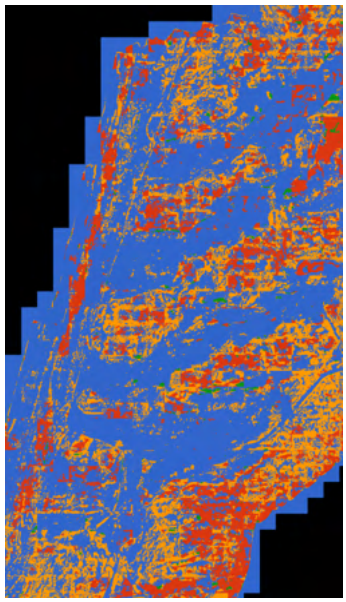
03/07/2023

Seasonal Effects on Habitat Mapping Using Convolutional Neural Networks and Human-In-The-Loop Machine Learning

Exploring Automated Habitat Monitoring Potential







Michael Urson (8890382)

SUPERVISED BY: GERBEN RUESSINK

EXAMINED BY: TIMOTHY PRICE

PROGRAMME: MASTERS IN APPLIED DATA SCIENCE

Table of Contents

List of Tables

List of Figures

Acknowledgements:

I'd like to express my thanks to Gerben for his valuable guidance in writing this thesis, Tino Ouwerkerk for his assistance in resolving software-related issues, and to my wife Maria for supporting and inspiring me to keep my thesis (and mental health) in check.

Abstract:
In the Netherlands, management of coastal dunes has changed focus from short-term flood prevention and drinking water protection to long-term biodiversity promotion and back-dune development. To understand the impact of dune management activities, efficient monitoring methods are required. The use of unmanned aerial vehicles (UAVs) to provide images of areas of interest, and convolutional neural networks (CNNs) to segment these images, are promising developments in remote sensing. Additionally, human-in-the-loop machine learning (HITLML) provides an efficient way to annotate remotely sensed images for land cover classification. Vegetation phenology has shown to impact identifiability from aerial images in some applications. However, the viability of the combined application of these four elements (UAVs, CNNs, HITLML and phenology) to coastal dune monitoring is yet to be assessed. Here we show that these elements can potentially be effective in monitoring coastal dune development. We found that phenology impacts CNN performance, and CNNs trained on multi-season data perform more consistently than single-season CNNs. Additionally, increasing training data volume does not always improve CNN performance. Overall multi-season models' performance (measured by f1 score) was around 5% better than season-specific CNNs, which is in line with findings of other research. Predictions on specific seasons revealed roughly equivalent performance by season-specific and multi-season CNNs. The impact of increasing the dataset size had minimal effect on model performance. White dunes were most error-prone of all habitat types, being most frequently incorrectly predicted as grey dunes. Our results demonstrate that in the context of dune habitat analysis, training data should comprise seasonal diversity, and that beyond a certain point, increasing the sample size (without additional temporal diversity) is likely to have a limited effect. We anticipate our research to lead to improved practices in automated dune monitoring. Diversity of data appears to have a greater impact on model performance than volume alone, and CNNs provide a useful way to create habitat maps for automated dune monitoring. Additional research should incorporate other factors, such as the impact of weather conditions on results, as well as using ensemble models.

1. Introduction

1.1 Motivation and Context
Historically, dune management in the Netherlands has been focussed on safety, such as flood prevention and drinking water protection, at the expense of biodiversity and back dune development (Jackson & Nordstrom, 2011). More recently, the focus of dune management has become to restore aeolian transport, thus simultaneously promoting sustainable dune development and biodiversity (Arens et al., 2013). One strategy to do this has been to excavate notches in the foredune (e.g., in Zuid-Kennemerland) and monitor dune and vegetation development over time (Ruessink et al., 2018). However, manual monitoring is resource intensive, and thus an automated solution is needed. Convolutional neural networks (CNNs) have been successfully used to distinguish between vegetation types from aerial images with high accuracy (Kattenborn et al., 2019). CNNs, when combined with human-in-the-loop machine learning (HITLML), could present a viable solution to monitoring dune development (Buscombe et al., 2022), because they have the potential to automatically generate habitat maps with limited human input. Prior research suggested the relevance of vegetation phenology to CNN performance (Yang et al., 2019), however more research in this direction is required (Katal et al., 2022). In this project, CNNs will be used to segment vegetation types from aerial images, based on images from different seasons, to assess the viability of this automated dune monitoring approach.

1.2 Literature Review

1.2.1 Coastal Dune Habitats
The lifecycle of dunes consists of several components, progressing from embryonic dunes (which are youngest and closest to the sea) to shrub/woodland (most mature and furthest from the sea) (Hesp, 1991). To allow for more practical and informative monitoring, four specific habit types will be focussed on in this project: embryonic dunes, white dunes, grey dunes and shrubs. Embryonic dunes are characterized by mostly sand, and some pioneering marram grass (Natura 2000, 2008a). White dunes

consist of mainly marram grass, with sand visible between grass growths due to mosses not yet having had time to develop (Natura 2000, 2008b). Grey dunes are similar to white dunes in containing marram grass, but will typically have a mossy base (resulting in them appearing grey in spring and green in summer), and may contain low herbaceous vegetation in addition to or instead of marram grass (Natura 2000, 2008c). Finally, shrubs, the most mature of these habitat types, consist of larger herbaceous vegetation (such as sea buckthorn, or hippophae rhamnoides, as well as other larger plants) (Natura 2000, 2008d). Monitoring dune development by focusing on these four habitat types allows for insight into dune lifecycle without unnecessary complications.

1.2.2 UAVs for Habitat Analysis

UAVs are remote-controlled aircraft that have been used in generating landscape imagery over large areas. They can efficiently generate high-quality multi-spectral imagery over a large area in a short time, making them a promising alternative (or compliment) to land-based surveying. Aerial images can identify areas that field-based methods miss, however, they can also produce false-positives (e.g., due to shadows and water), for an agreement of around 70% between field-based and aerial habitat assessments (Oldeland et al., 2021). In terms of efficiency, UAVs can cover an area of 18ha over two 30-minute flights, thus evidencing the rapidity of data collection and the potential for drone imagery to be used instead of field data for habitat classification. Recent research successfully used drone imagery to map habitats of dune systems. For example, Cruz et al. (2023) applied random forest models to UAV dune images, comprising various layers of spectral and topographical information, and combined using principal component analysis resulting in accuracy of around 92%. This study indicates the potential for using UAV images and machine learning techniques for creating accurate habitat maps of coastal dune areas.

1.2.3 Human-in-the-Loop Machine Learning:

One of the challenges of using UAV images for habitat classification is the resource-intensity of labelling data to be used to train models (Vitousek et al., 2023). To address this, a new development in artificial intelligence combines the efforts of humans and machines: human-in-the-loop machine learning (Mosqueira-Rey et al., 2023). This takes several forms, however most relevant to the problem at hand (classification of land cover) is termed assisted annotation and predictive annotation (Buscombe et al., 2022), which lead to the development of the application DashDoodler. With DashDoodler, humans create sparse annotations (‘doodles’), which act, together with the image itself, as inputs to a multilayer perceptron (Bishop, 2006) and conditional random field models (Kumar & Hebert, 2006). These machine learning models serve to complete the annotation on a per-pixel basis that the human annotator has indicated. These models are re-initialized for each image, and the approach is iterative, ensuring that the result is acceptable to the human annotator. This approach is 3 – 10 times faster than manual annotation, and it avoids some its pitfalls (including incomplete labels, as well as detecting finer patterns that are too time consuming for humans to annotate) (Buscombe et al., 2022). Refer to Figure 1 for an illustration of a DashDoodler workflow.

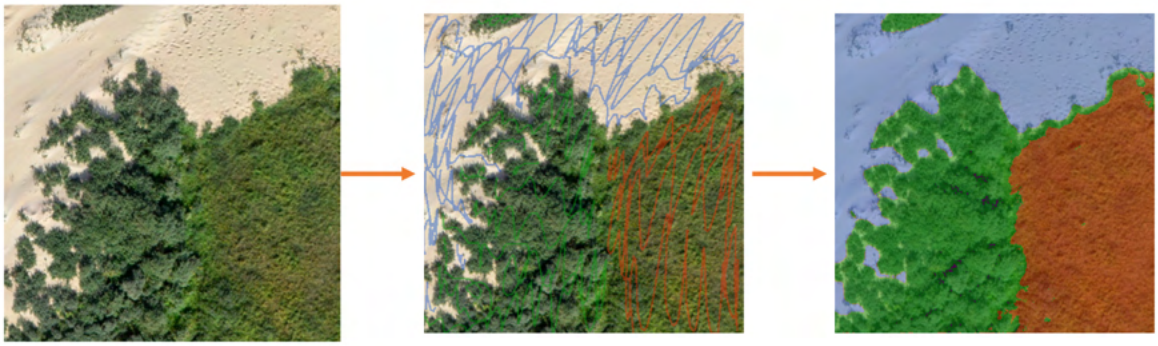


Figure 1: Illustration of the DashDoodler workflow. The original image tile (left) has sparse annotations (doodles) applied to it (middle) by the annotator, from which the annotated image is generated (right) using machine learning models.

1.2.4 CNNs for image segmentation

CNNs have been used successfully in the segmentation of images captured by UAVs for the purposes of habitat mapping (Kattenborn et al., 2019). Image segmentation involves separating an image, on a per-pixel basis, into components to enable analysis of the image (Kattenborn et al., 2021). What distinguishes CNNs from other neural network architectures is that they contain at least one convolutional layer, which applies the convolution operation to the input feature map (LeCun et al., 1998). Convolution involves passing a filter (also known as a kernel) of a specified size across the entire input feature map – refer to Figure 2). This is done via a 'sliding window,' which ensures that patterns between neighbouring pixels are considered by the model. From this process, a new feature map is created, which is populated based on the results of this filtering operation. When a CNN is trained, kernel values are estimated such that the patterns most relevant to the task are learned. Typically, throughout a CNN, multiple convolutional layers are used, with differing numbers of layers and feature map sizes, which allow for the detection of patterns. As a result, CNNs have a key advantage over other machine learning methods relevant to image segmentation, as they consider not just pixel values in isolation, but also relationships between pixels. In addition to being theoretically well-suited to image segmentation tasks, CNNs have also proven to be useful in segmenting habitats from UAV images. One study found that their feed-forward neural network outperformed their random forest model in the classification of vegetation types from UAV images (Oldeland et al., 2021).

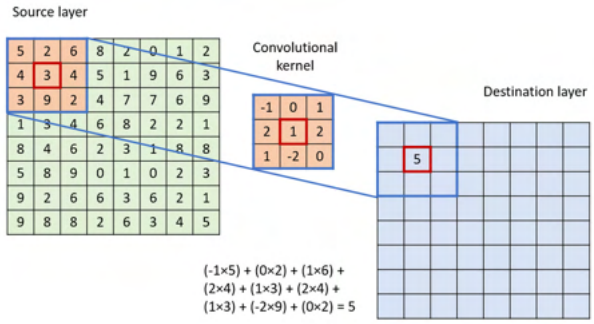


Figure 2: illustration taken from Podareanu et al. (2019) of convolution operation, using a 3x3 convolutional kernel and 8x8 source layer.

A CNN's architecture refers to the specifications and arrangement of the components of its various layers. One common architecture of CNNs used for image segmentation is known as UNet (Ronneberger et al., 2015). This method was originally developed for biomedical image segmentation, but has been used successfully in numerous UAV image segmentation tasks, with studies achieving accuracies from around 84% (Kattenborn et al., 2019). The network gets its name from the shape of the arrangement of its layers, which are arranged in a U-shape and allows for localization of predictions (i.e. a class label per pixel), as well as efficient use of training data (Ronneberger et al., 2015).

Recently, ResUNet (Zhang et al., 2018) was developed as an improvement on UNet, which makes use of residual learning (He et al., 2015) to further optimize use of training data. Residual learning involves creating skip connections between layers, which allow for the passage of information from earlier layers to later layers, without being first passed through each intermediate layer. These skip connections allow for a reduced number of parameters to train, without negatively impacting performance. Zhang et al. (2018) achieved a 1% increase in accuracy using a ResUNet compared to a standard UNet, with having only ¼ of the parameters. One study using a ResUNet in remote sensing image segmentation (Diakogiannis et al., 2020) achieved an overall F1 score of 92.9% on the ISPRS Potsdam dataset. The results indicate that higher performance is possible with ResUNets compared to UNets, while requiring less training data.

Due to their increasing popularity in the segmentation of geographical images (Kattenborn et al., 2021), a pipeline has been developed for the implementation of various CNN architectures for these tasks (Buscombe & Goldstein, 2022). This software, called Segmentation Gym, provides a platform for training and fitting CNN models, and has been developed for use in conjunction with DashDoodler

(Buscombe et al., 2022). Segmentation Gym also allows for augmentation of training datasets, whereby new training data is created through the manipulation (e.g. rotations, cropping, and flipping the images and associated labels) of existing training data, thus allowing for more robust models without additional annotation. Segmentation Gym will be used in training and fitting CNNs used in this thesis.

1.2.5 Phenology and CNN habitat classification

Several studies have reported increased species identification accuracy when plant phenology is taken into consideration when using UAV images. For example, a study that used RGB UAV imagery at 10cm resolution to identify invasive tree species found that incorporating phenological information (i.e., including summer images, when the tree species of interest displayed red flowers), improved the accuracy of neural network models by around 5% (Pearse et al., 2021). Another study used multi-spectral UAV imagery and random forest models to identify dune habitats in Ireland (Cruz et al., 2023). They found that incorporating multi-seasonal data in their model increased accuracy by around 8%. Finally, Pöttker et al. (2023) used convolutional neural networks to map plant communities in grasslands and found that classification accuracy increased by around 5-10% when multi-phenological (i.e. spanning multiple phases of plant life cycles) data was incorporated into their convolutional neural network models. All of the above suggests that vegetation phenology impacts its ability to be identified by CNNs from UAV images, prompting a systematic review (Katal et al., 2022) to highlight the need for further research in considering vegetation phenology in habitat identification.

1.3 Research Questions:

The overarching objective of this thesis is to contribute to the automated mapping of dune habitats from high-resolution UAV imagery. Based on the literature review, the following two research questions were formulated:

1. How do drone images captured in different phenological phases or seasons affect the differentiability of habitat types in the Dutch dune environment?
2. What are some of the benefits and drawbacks of using HITLML for habitat mapping?

1.3.1 Translation of research question into data science question

The intention behind the first research question is to investigate whether a model trained on data from an earlier temporal period could be useful in predicting a habitat map in a later period. To get the best performance, should training data of two seasons be combined, or should a model trained only on the intended prediction season be used? Additionally, what happens to model performance when a non-matching season is predicted, as well as when a later period is predicted? Finally, what differences in performance are evident when comparing a model trained on a 100% combined season dataset, compared to a model trained on a random selection of half of each season's training datasets. To answer these sub-questions, a total of four CNNs are required – refer to section 3.1.2 for details.

2. Data

The dataset (Ruessink, 2023) used for this project comprises several orthomosaics that were stitched together from 3-banded (red, green and blue: RGB) photographs taken with a UAV. The EPSG:28992 coordinate reference system was used, with the following spatial extent (in meters): easting from 98300 to 99200, northing from 492900 to 494100. Images have a resolution of 18000×24000 pixels, with spatial resolution of 5cm/pixel. For additional specifications of the dataset, refer to Ruessink et al. (2018). To answer the first research question, the following spring and autumn datasets were selected, see Table 1. Spring 2021 data was unfortunately unavailable, and thus spring 2022 data was used in its place. Two time periods (2018 and 2021/22) were used to investigate how the performance of models trained on an earlier period perform in creating habitat maps for a later period. For all datasets, spring was prior to the start of the growth season, while autumn is at the end of it. Visualizations are presented prioritizing first temporal period, and then season, for ease of comparison. The second research question was answered in the course of answering the first research question.

Table 1: Datasets used

File reference	Date captured	Meteorological Season & year
20211013_005.tif	13/10/2021	Autumn_2021
20220302_005.tif	02/03/2022	Spring_2022
20180917_005.tif	17/09/2018	Autumn_2018
20180419_005.tif	19/04/2018	Spring_2018

2.1 Selected data exploration results

Refer to Figure 3 for an overview of the datasets used in this project. The vegetation in the spring images have limited contrast with the surrounding sand, while the vegetation in autumn appears to be lusher and greener and thus has more contrast with the surrounding sand. Additionally, there appear to be differences in weather conditions in the images. For example, spring 2018 and autumn 2021 were captured in cloudy conditions, while autumn 2018 and spring 2022 in sunny conditions (as evidenced by the strong shadows and harsh lighting in the latter two). These conditions will likely impact CNN segmentation performance.

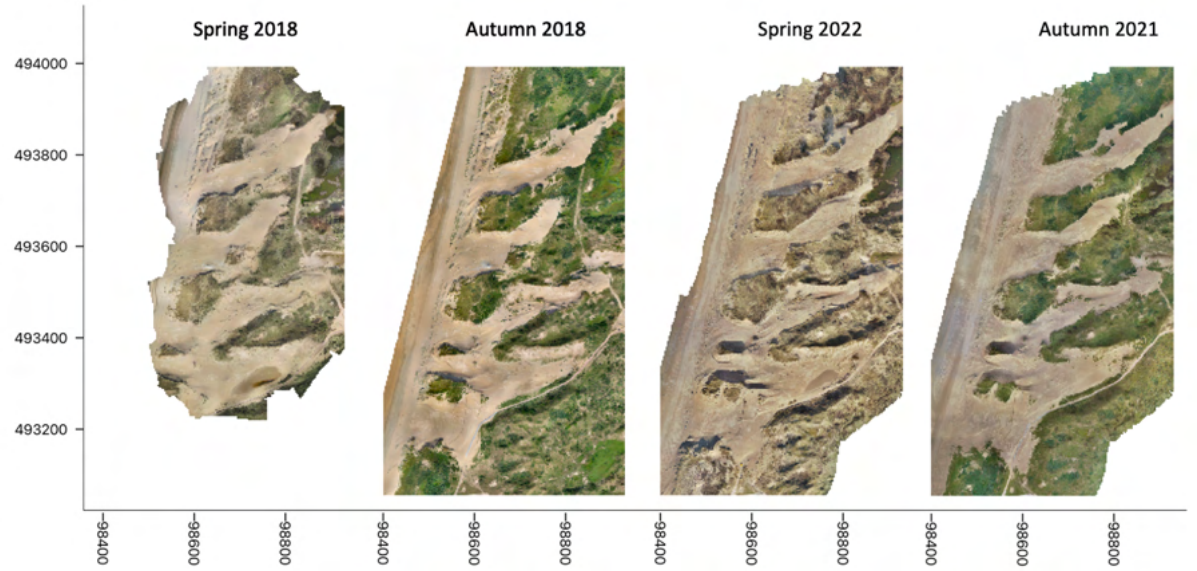


Figure 3: Overview of datasets used. Axes are denominated in meters, representing eastings and northings, using the EPSG:28992 coordinate reference system.

To better understand the differences in differentiability (based on colour) between the different habitats, a specific location was chosen in each of spring 2018 and autumn 2018 and annotated using DashDoodler. The original image was then masked with each label and a histogram was plotted of the resulting RGB values (Figure 4). From the visualization it seems that white and grey dunes are more similar in colour signatures for spring 2018 than in autumn 2018, as shown by the locations of the mode pixel counts of each colour (i.e., the histogram peaks). In spring 2018, they appear to be closer together than in autumn 2018, and in autumn 2018 they are more spread out, implying greater differentiability. The same can be said about sand and shrubs – in spring 2018 they have similar modal pixel values for each colour band (both are clustered at around 200), while in autumn 2018 the modal values are further apart (shrubs: roughly 100, sand: roughly 200).

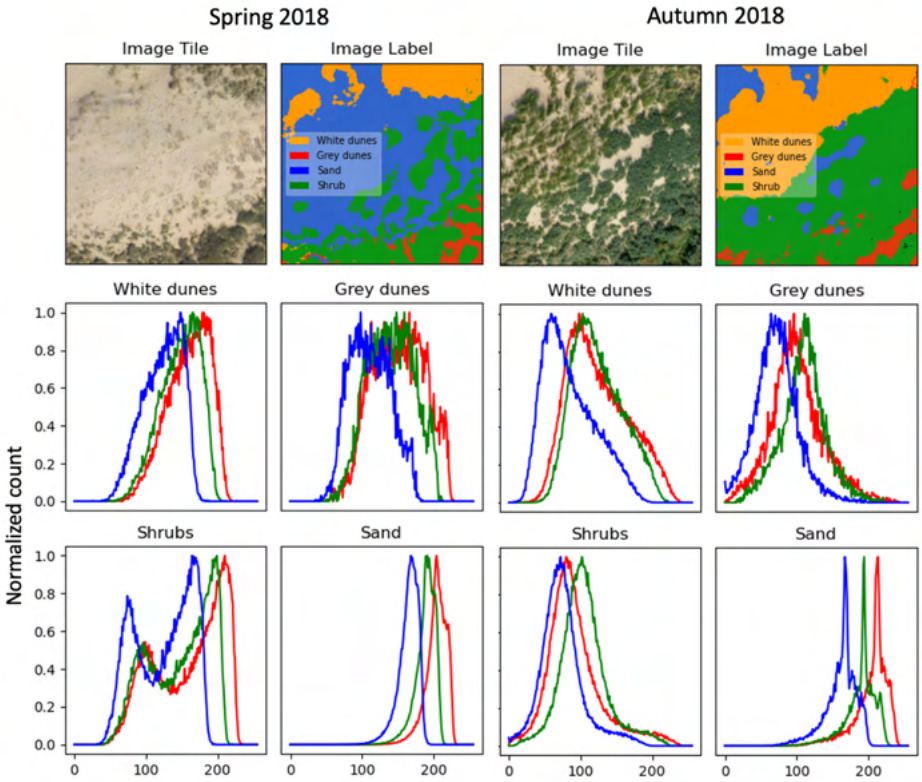


*Figure 4: Comparison between Spring and Autumn habitat type differentiability. The histograms were generated by masking the image tile (500*500 pixels) for each image label, and then counting the number of occurrences of each red, green and blue pixel value.*

2.2 Ethical and legal considerations of data:

The dataset used (Ruessink, 2023) is available under the Creative Commons Attribution 4.0 license, which allows for free use for the dataset provided that the source is credited and that no additional restrictions have been placed (Creative Commons, 2023). As due credit has been given and no further restrictions placed on readers of this paper, the legal considerations of this dataset have been met. This dataset contains images of people, however, the resolution is far too low (5cm/pixel), and the angle inadequate (top-down) to uniquely identify them. Additionally, the people that may appear in the dataset are not the subject of research. Therefore, the relevant ethical and legal requirements have been met.

3. Methods

The methods section of this thesis has been split into two components, first of which is the pre-processing and annotation of data, focussing on the preparation of data for training and testing models. This is followed by model training and evaluation, which focusses on using the data from the first component to train and test the models.

3.1 Pre-processing & Annotation

Figure 5 outlines the data preparation and analysis steps.

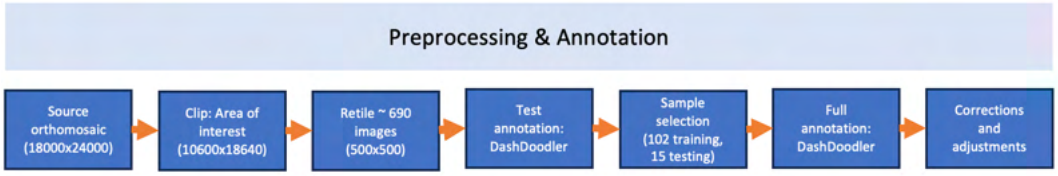


Figure 5: Process flow for data preprocessing and annotation.

3.1.1. Clipping and Retiling

Beginning with stitched, georeferenced orthomosaics (as described in section 2), the images were first clipped to the area of interest (easting range 98400 to 98930, northing range 493058 to 493990). For the purposes of annotation and model fitting, the images were required to be retiled, as otherwise they would be too large. This was performed using GDAL's retile Python API, and both .jpg and .tif files were generated in this process (.jpg for model training and testing, .tif for geolocating model predictions for re-stitching). A resolution of 500x500 pixels was selected because it represents an area of 25mx25m, which is a practical size for containing sufficient habitat diversity, while still being small enough to train a model on without consuming too much memory. Habitat diversity in training data is important for the CNN to be able to differentiate between multiple habitat types in the same image tile.

3.1.2. Test annotation: DashDoodler

DashDoodler (Buscombe et al., 2022) was used in the annotation of testing and training data. Prior to sampling training and testing data, two people annotated 15 tiles from Autumn 2021 to get a sense of which habitat types could be difficult to discern from one-another. The results of this are presented in the confusion matrix in Figure 6, which indicates that sand and grey dunes were generally agreed on by the annotators (93% and 88% of the time respectively). White dunes appear to be mislabelled as grey dunes (35% of the time), and shrubs as grey dunes (26% of the time). The latter findings were used to inform the sampling of training data, resulting in sampling more images appearing to contain white dunes and shrubs than others. The goal of sample selection in this context was to provide a training set that captures the complexity of the AOI (Area of Interest) – not to have habitat types in the same proportion as the AOI.

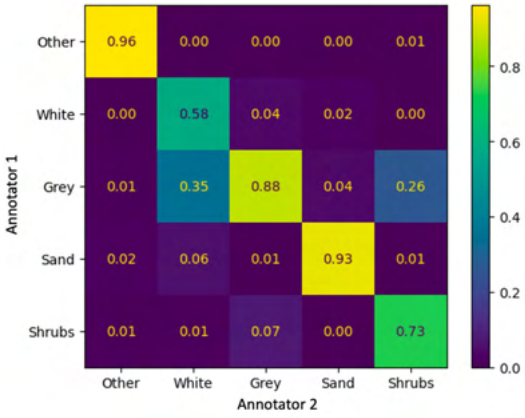


Figure 6: Confusion matrix of test sample results, showing pixel-level habitat type confusion.

Following the discussion of differences, a further sample of 5 images was selected to determine the consistency of annotations. For this sample, Cohen's kappa (Cohen, 1960), a measure of inter-annotator reliability, was calculated to be 0.84, which indicates almost perfect agreement between annotators. Although ultimately all sampled images were annotated by a single annotator, this test provides some evidence for the between-sample consistency of annotation.

3.1.3. Sample Selection

The spring 2018 dataset (see Figure 3), is relatively incomplete (i.e. contains substantially more white space than the other images), likely due to the weather conditions on the day. Figure 7 compares the same region in spring 2018 with autumn 2018, illustrating the relative blurriness of spring 2018.

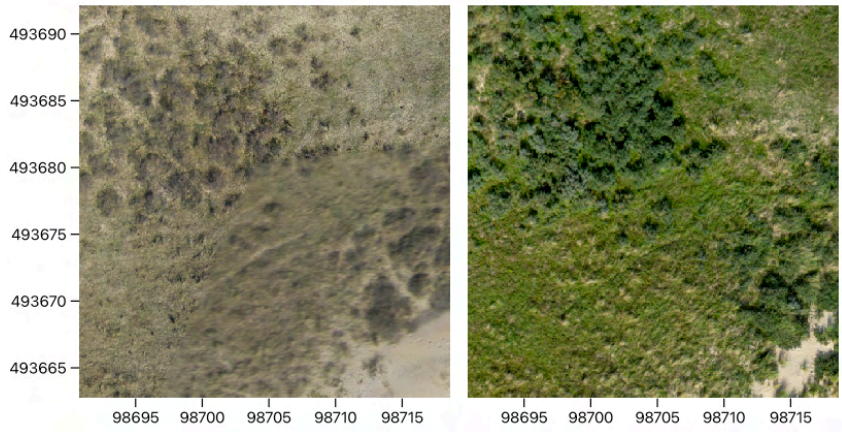


Figure 7: Blurry images in spring 2018 (left) likely due to windy conditions, compared to clear image from autumn 2018 (right). This also resulted in a less complete orthomosaic as shown in Figure 3.

As a result of the somewhat compromised image availability in spring 2018, that orthomosaic was used as a basis for sample selection, as there should be an existing tile in other seasons for every tile in spring 2018, but not necessarily the other way around. In total, 102 images were selected from spring 2018 and autumn 2018 to form the training datasets, representing a total of 2550m^2 of land area in each aforementioned season. The testing datasets comprised 15 images (representing 375m^2) from each orthomosaic presented in Figure 3. Given the data volume requirements of the models, as well as taking the spread of habitat types in the AOI and the number of habitat classes intended to be predicted into consideration, the testing and training dataset sizes were considered to be sufficient.

The same geographic locations were sampled across all testing and training datasets. This ensured consistency between datasets, allowing for the assessment of the impact of phenological/seasonal factors without confounding results with location variability. See Figure 8 for a visualization of the locations sampled.

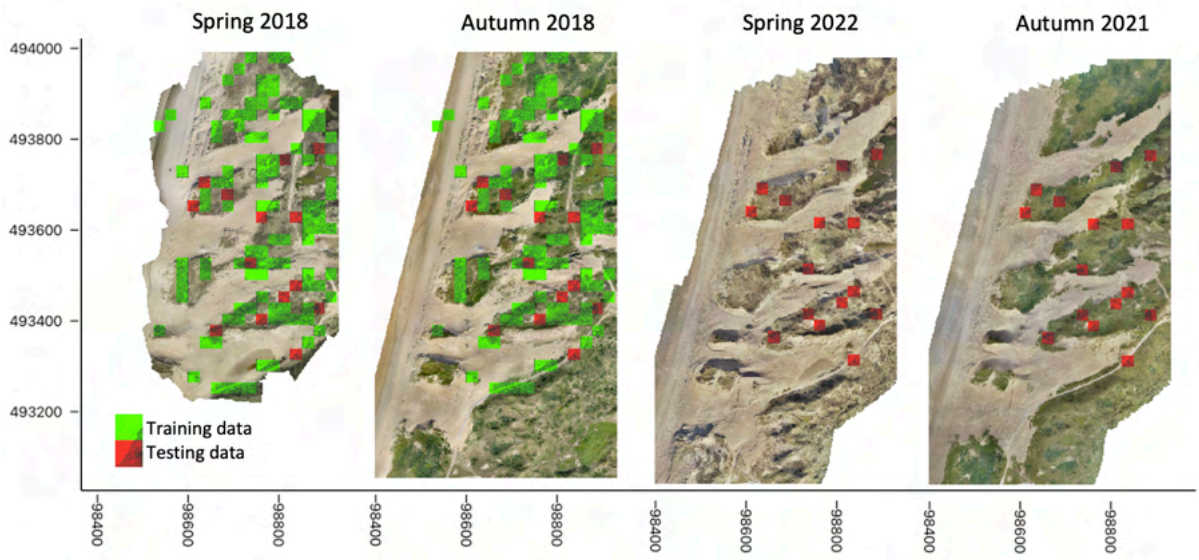


Figure 8: Visualization of location consistency of training (green) and testing (red) split between datasets.

Despite being sampled from the same geographic locations, in the training datasets there are more shrubs and white dunes visible in autumn 2018 compared to spring 2018, see habitat land coverage splits in Figure 9. This follows from the points discussed around Figure 4, where, based on colour, shrubs appeared more sand-like in spring 2018, while in autumn 2018 they were more differentiable. Additionally, the 'Other,' class is larger in spring 2018 than in autumn 2018. This arose from water in the northern part of the AOI, which received this classification, and was not present in autumn 2018 (wherein, 'Other', comprised mainly roads and people). Additionally, the testing and training datasets are quite different from each other, and not representative of the orthomosaic as a whole (which would contain far more sand cover, which is easier to predict based on 3.1.2). One reason for this is the seasonal

visibility of certain habitat types, however another is that the testing dataset was designed to provide a challenging benchmark to assess model performance, to see how they would perform in difficult conditions (large degree of habitat diversity).

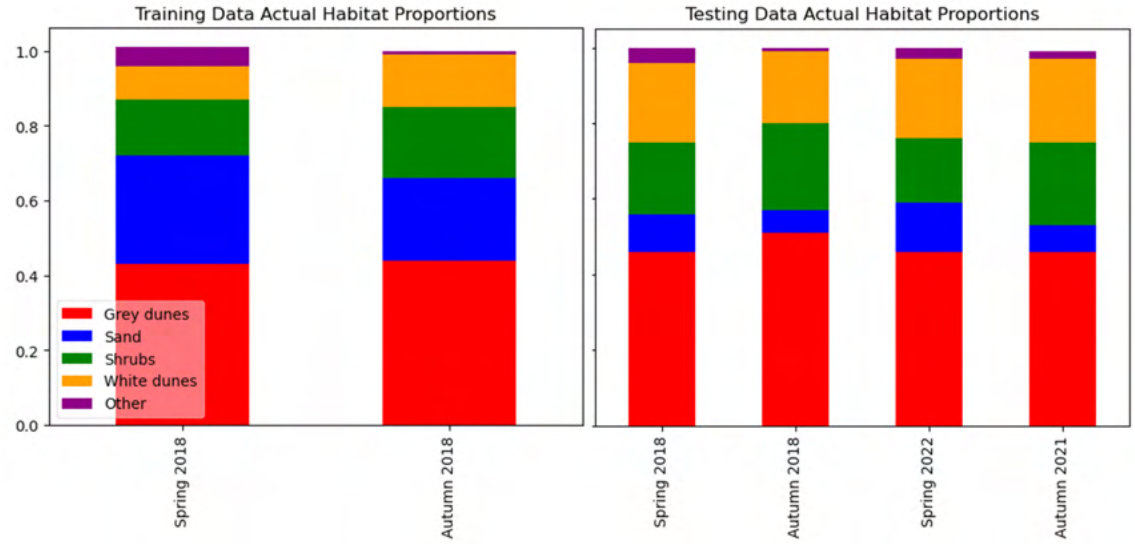


Figure 9: Habitat proportions in training (left) and testing (right) datasets.

3.1.4. Full annotation: DashDoodler, Corrections and Adjustments

The testing and training datasets were annotated using DashDoodler, however some errors were noted in this process, specifically in the annotation of low-contrast images, as well as images containing shadows. The annotation of low-contrast images generally led to DashDoodler performing in unexpected ways (see Figure 10). Without sufficient contrast, the models used could not segment the

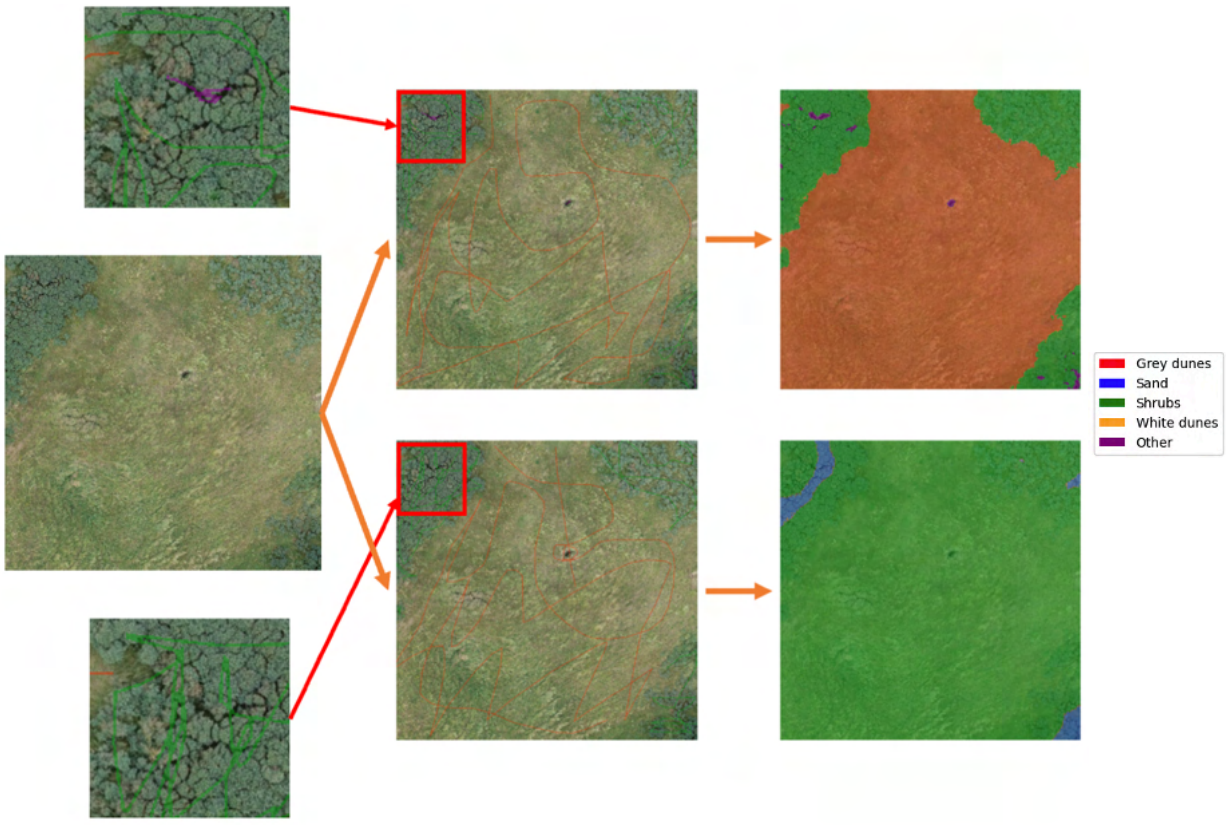


Figure 10: A low contrast image (left middle) is doodled in two different ways. One way separately annotates shadows (top middle – see the purple annotation), while the other does not (bottom middle). The results show that to get close to the desired results (top right), shadows in low contrast images need to be separately annotated as 'Other', otherwise unexpected results (e.g. the annotation of sand, despite it not having been doodled) may occur (bottom right).

images as intended by the user, and so classes were annotated seemingly at random. To correct this, an additional class was annotated ('Other') for shadow regions, which sufficiently reduced within-class variance for an annotation to be generated in line with what was expected.

Another consequence of shadows in images is that DashDoodler assigned null labels to certain pixels (represented as black in the annotated images) – see Figure 11. There were 66 labels exhibiting black pixels, 65 of which occurred in autumn 2018, while only one in spring 2018. The labels were adjusted by means of a Python script, which replaced black pixels with purple pixels (representing the 'Other' habitat classification). This allowed the images to be used in training and testing the models, at the expense of some noise (i.e. shadows being labelled as 'Other').

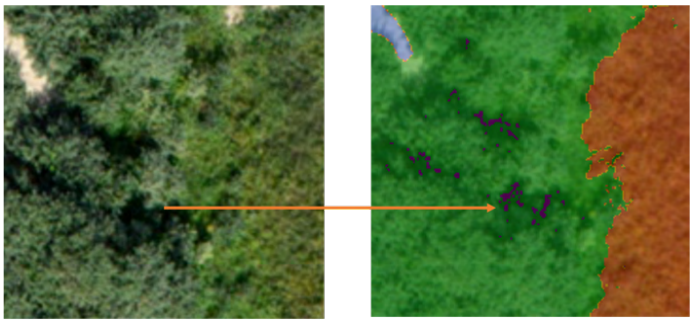


Figure 11: Portion of source image (left) and black pixels predicted by DashDoodler (right) for the shadow region, indicating errors in classification.

3.2 Model Training and Evaluation

The process followed in training and evaluating the models is presented in Figure 12.

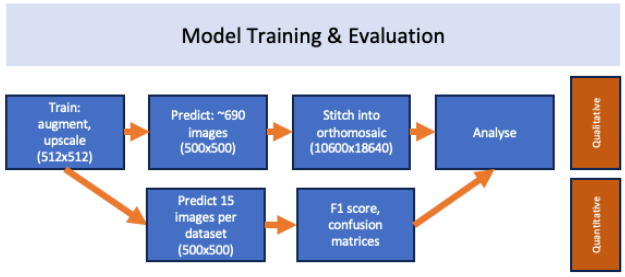


Figure 12: Process flow for model training and evaluation.

This section does not go into detail about each step of the training and evaluation process. Instead, the focus is centred around translating the first research question into data science questions that can more easily be tested, as well as providing information and justification of key modelling assumptions.

3.2.1. Model Training – Assumptions and Specifications

Refer to Table 2 for an overview of the datasets used to train each model.

Table 2: Model and training dataset overview

Model	**Spring 2018 training data**	**Autumn 2018 training data**
spring_2018	102 tiles	Not applicable
autumn_2018	Not applicable	102 tiles
combined_large	102 tiles	102 tiles
combined_small	51 tiles (randomly sampled)	51 tiles (randomly sampled)

3.2.1.1. Model Training: Specification and Testing Data Augmentation

Segmentation Gym (Buscombe & Goldstein, 2022) was used to train and specify the models. All models had the same hyperparameters and thus differed only by their respective training datasets. The

hyperparameters that have the most impact on model performance are the batch size and loss function (Buscombe & Goldstein, 2022), which, together with model architecture and input layer dimensions, is discussed below.

The models were trained on a desktop having an Nvidia GTX 1080 with 8GB RAM. Model architecture was specified as 'resunet' (Diakogiannis et al., 2020). Batch size should generally be as large as graphics card RAM allows (Buscombe & Goldstein, 2022), as this determines how many samples are considered before updating model parameters. However, a larger batch size ensures a closer fit to the data, thus too high a batch size can be detrimental due to overfitting. The batch size used in all models for this thesis was 5 images.

A model's loss function determines how the model is penalized in relation to differences between actual and predicted values, which, in turn, determines how model weights are updated through training. Given that the dataset is relatively unbalanced between classes (see Figure 10), the loss function selected should be relatively insensitive to class imbalances. Therefore, the Dice loss function (Sudre et al., 2017) was used as it meets this criterion.

Input image size was specified as 512x512 pixel resolution, as the model architecture is only compatible with image resolutions in increments of 128 pixels, so as a result, the 500x500 input images were upscaled to 512x512 pixel resolution. For details on the configuration files containing the hyperparameters used in each model, see Appendix 1.

Data augmentation (Eaton-Rosen et al., 2018) refers to creating new samples from the training data by adjusting training images and their labels, thus creating new training samples. These adjustments can include mirroring, rotating and cropping/zooming. The purpose of this is to prevent the model from memorizing the data (i.e., overfitting), as well as to make the model more generalizable to changes in the form of images. Such augmentation was performed in the training of these models, thus increasing the number of tiles in each training dataset five-fold.

3.2.2 Model Evaluation – Quantitative and Qualitative

Two model evaluation methods (quantitative and qualitative) was used to assess model performance (see Figure 13). For a quantitative performance evaluation, models were fitted to testing data and F1 scores were calculated for each dataset-model pair. The qualitative evaluation involved fitting each model to entire orthomosaics of the spring 2022 and autumn 2021 datasets. The resulting tiles were then re-stitched using metadata from the original .tif file (refer to retiling of images in section 2: Data), and predicted orthomosaics were formed. Once stitched, a comparison was made between each orthomosaic to determine how each model performs on a broad level on each dataset.

3.2.2.1 Evaluation – Quantitative Metrics

The quantitative performance of the neural networks will be evaluated using the F1 score, as well as confusion matrices, both of which have been commonly used in evaluation performance of multi-class classifiers (Sokolova & Lapalme, 2009). A confusion matrix is a table presenting proportions of actual and predicted classifications for each class, providing insight into which specific classes are sufficiently similar to become confused with one-another. This is useful to gain insight into the performance of the model relating to specific classes. The F1 score is calculated as the harmonic mean of precision and recall. Precision is calculated as true positives/total predicted positives, while recall is calculated as true positives/total positives. The F1 score combines these two measures into a single number for a concise view of model performance, having a maximum value of 1 (indicating only correct predictions) and a minimum of zero (indicating only incorrect predictions).

$$Precision = \frac{true\ positives}{total\ predicted\ positives}$$

$$Recall = \frac{true\ positives}{total\ actual\ positives}$$

$$F1\ score = 2 * \frac{Precision * Recall}{Precision + Recall}$$

3.2.2.2 Evaluation – Qualitative

Although quantitative metrics do help to understand some aspects of model performance, they do not provide complete insight into it. Some factors, such as the general appearance of predictions (e.g. visibility of tiles), as well as ecological plausibility (e.g. the locations of predicted habitat types). To better understand model performance in respect of these factors, complete orthomosaics were predicted by each model on the spring 2022 and autumn 2021 datasets and analysed further.

4. Results and Analysis

This section has been split into two components, focussing on a quantitative (4.1) and a qualitative evaluation (4.2) of model performance. Figure 13 presents a training history of each model, illustrating how each model's iou (intersection over union, an accuracy measure) and loss developed as the number of epochs (the number of times the entire training set has been fed into the model) increased. All models except for spring 2018 stopped after around 50 epochs, which is likely due to the activation of early stopping conditions to prevent overfitting. Spring continued for longer (to around 90 epochs) probably because of the limited differentiability between habitat types previously discussed, requiring more training.

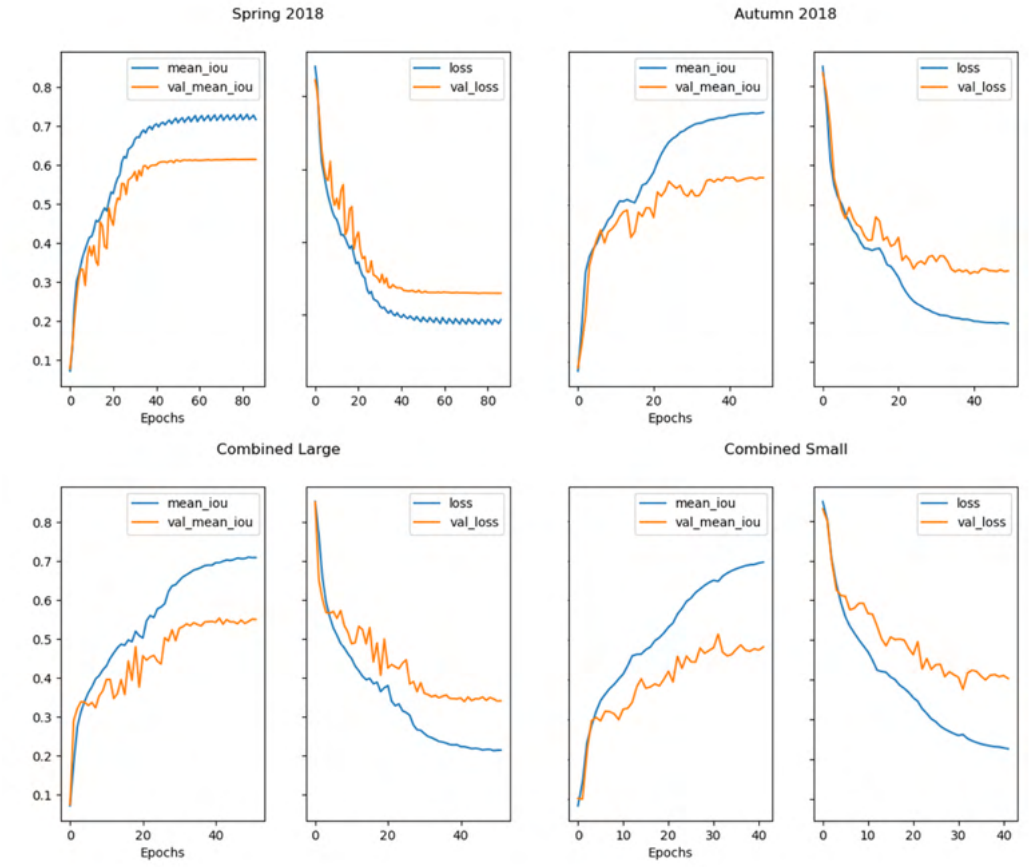


Figure 13: Training history of each model. mean_iou refers to mean intersection over union, which is an accuracy measure.

4.1 Quantitative Performance Analysis

As illustrated in Figure 14, the combined models had the best overall performance, with an overall F1 score of around 5% - 10% higher than that of season-specific models and remaining closer to 5% when only considering datasets from an unseen time period (overall_2022). Additionally, for specific seasons, the combined models perform as well as or better than the season-specific models. This implies that the combined models learned additional patterns by which vegetation can be recognized through exposure to training data from different seasons, and possibly lighting conditions. The difference between the two combined models is minor, with combined_small sometimes outperforming combined_large. Performance of all models were relatively low on the spring 2022 dataset which seems to be related to

the weather conditions on the day. For season-specific models, performance deteriorated by around 10% (spring) - 15% (autumn) when applied to a matched season at a later (unseen) period. The same degradation can be seen when applying a season-specific model to its unmatched season. However, in both cases, the degree of degradation is greater for the autumn model than the spring model, which could result from the spring model having been relatively underfit compared to the autumn model.

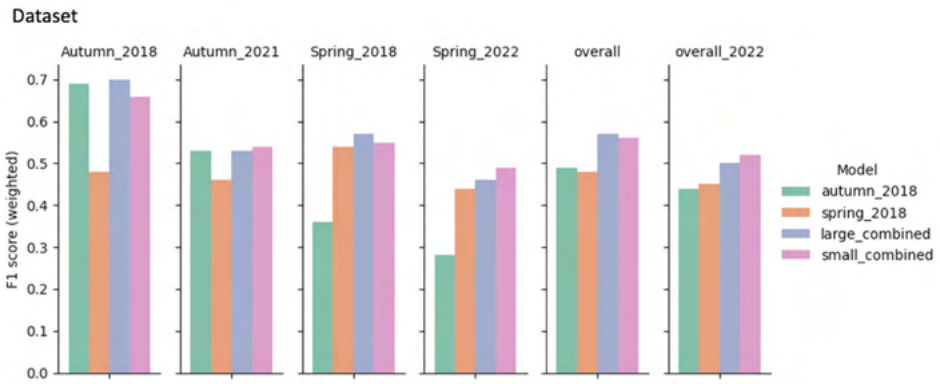


Figure 14: F1 Score per model and dataset. Overall_2022 contains Spring 2022 and Autumn 2021 test data.

Figure 15 shows the differences in lighting conditions between each testing and training dataset. Spring 2022 appears to have the most prominent shadows and harshest lighting of all datasets. These lighting conditions appear to be most similar to autumn 2018 than the other datasets. This could at least partially explain why combined models perform better than the season-specific model on the spring 2022 dataset. For combined models, training data included tiles that are season-consistent (i.e., spring 2018) and lighting-consistent (i.e., autumn 2018) with spring 2022.

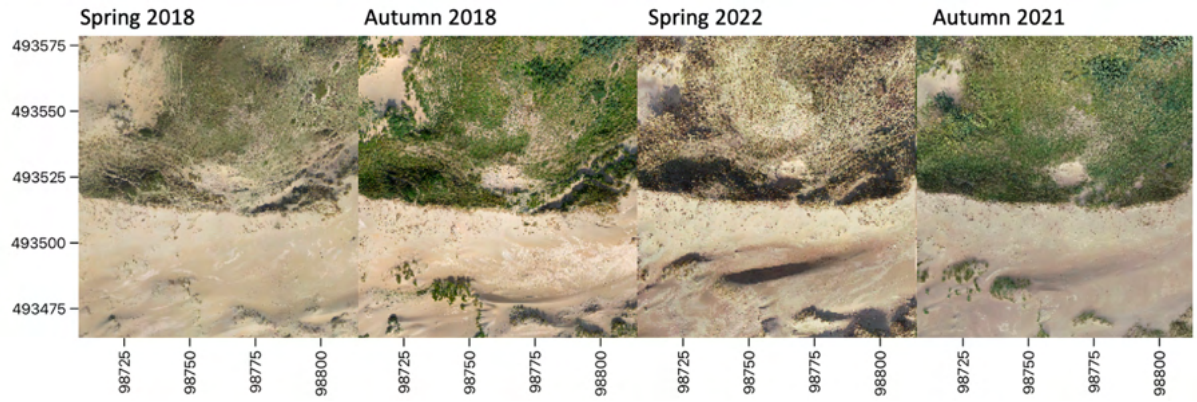


Figure 15: Different lighting conditions between the datasets. Spring 2022 has particularly harsh shadows, with lighting conditions most comparable with autumn 2018.

The confusion matrices in Figures 16 and 17 provide insight into the proportions of classes predicted for each true class, which give an indication of which classes might have similar appearances to other classes. The performance of all models degraded less for the autumn 2018 to 2021 datasets, compared to the spring 2018 to 2022 datasets. This is demonstrated by values close to 1 in the diagonal lines extending from the top left corner to the bottom right corner for autumn, but not for spring, suggesting that the spring 2022 dataset was more challenging for all models.

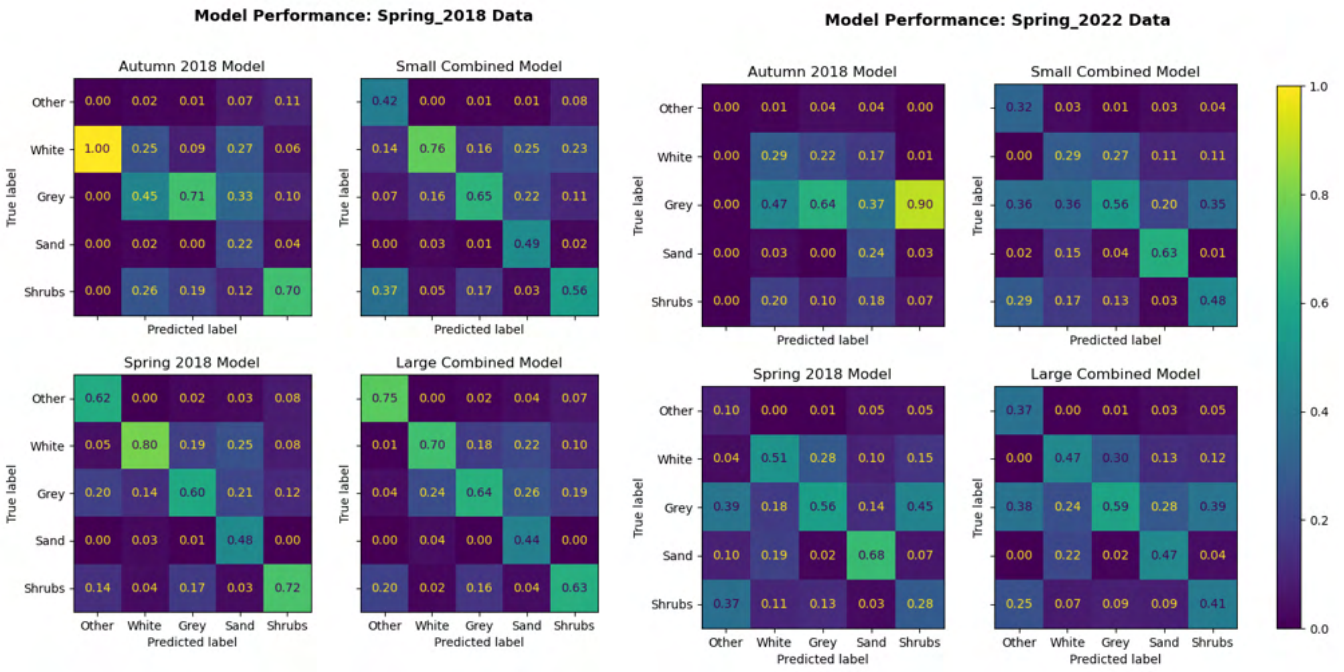


Figure 16: Comparison of spring datasets between periods and models.

Figure 17: Comparison of autumn datasets between periods and models.

The extract from the confusion matrices in Table 3 suggests that using a season-matched model is better than using a season unmatched model. For example, the spring 2018 model predicted white dunes as grey dunes 88% of the time for the autumn 2018 dataset (2021: 58%), while for the spring 2018 dataset it only made this mistake 14% of the time (2022: 18%). The same pattern emerges for the autumn 2018 model, which made this mistake 45% of the time for spring 2018 (2022: 47%), while for the autumn 2018 dataset this was only 11% (2022: 10%).

Table 3: Erroneous predictions of white dunes as grey dunes - higher error rates occur with season-mismatched models.

	Datasets			
Models	Spring 2018	Autumn 2018	Spring 2022	Autumn 2021
Spring 2018 model	14%	88%	18%	51%
Autumn 2018 model	45%	11%	47%	10%

Combined models generally appear to perform as well as, if not better than, the season specific models, and there does not appear to be much of a difference between the combined model trained on a larger dataset compared to a smaller dataset. For example, concerning the autumn 2021 dataset, the small combined model outperforms the autumn 2018 model in certain areas (identification of 'Other' habitats, as well as slightly in the identification of sand and shrubs). On this dataset, the large combined model outperforms the small combined model only in the identification of white dunes (96% vs 78%), in which it also outperformed all other models for this dataset. However it did not outperform the small combined model outright in the other habitats. This performance was repeated in the spring 2022 dataset, whereby the combined models outperformed the season-specific model in classifying shrubs and grey dunes, which comprised most of the testing dataset. The season-specific model was marginally better in classifying white dunes (correct 51% of the time, compared to 47% and 29% for the large combined and small combined models respectively). There was little performance difference between the large combined and small combined models.

Generally, we see a tendency of models to confuse grey dunes with white dunes, especially for the spring 2018 model, wherein these dune habitats appeared very different to how they appeared in autumn. This model performed especially poorly on autumn 2018 data (confusing these habitat types 88% of the time), while this improved on autumn 2021 data (58% of the time). These results are in line with Figure 6, which showed that between human annotators, there was confusion between grey and white dunes, as well as between grey dunes and shrubs.

The spring 2022 dataset was challenging for all of the models, demonstrated by the low values of top-left to bottom-right diagonals in Figure 16. Even the best performing model by F1 score (the small combined model) confused white dunes and shrubs with grey dunes 36% and 35% of the time respectively. This could be due to the shrubs casting shadows, making for difficult predictions by the models in the spring 2022 dataset.

4.2 Qualitative Performance Analysis

The purpose of this section is to further analyse and compare the performance of each of the models on each dataset, reaching beyond the scope of the testing dataset and quantitative analysis.

4.2.1 Analysis of overall habitat maps

Overall, the predictions of the entire orthomosaics appear to present promising results for use in automated habitat modelling. The overall habitat maps (Figure 18), show that there are differences in model performance that appear to be in line with the findings identified in the preceding section (for example, that that autumn 2018 model struggles to correctly identify shrubs and white dunes, classifying them as grey dunes 90% and 47% of the time respectively). However, there are some new insights that are gleaned from the orthomosaics themselves that are absent from the confusion matrices.

Firstly, the orthomosaics display signs of tiling, whereby it becomes visible where exactly the stitching of the tiles making up the orthomosaic took place. This is particularly visible in Figure 19, at around 98475:493390, however, it can be seen in all the orthomosaics. One possible solution could be to retile the source image with different tile locations, predict habitat maps on each of those tiles, stitch them together and then create a new habitat map based on the modal habitat type for each pixel. Another solution could be to use ensemble modelling, which averages out the predictions of multiple models.

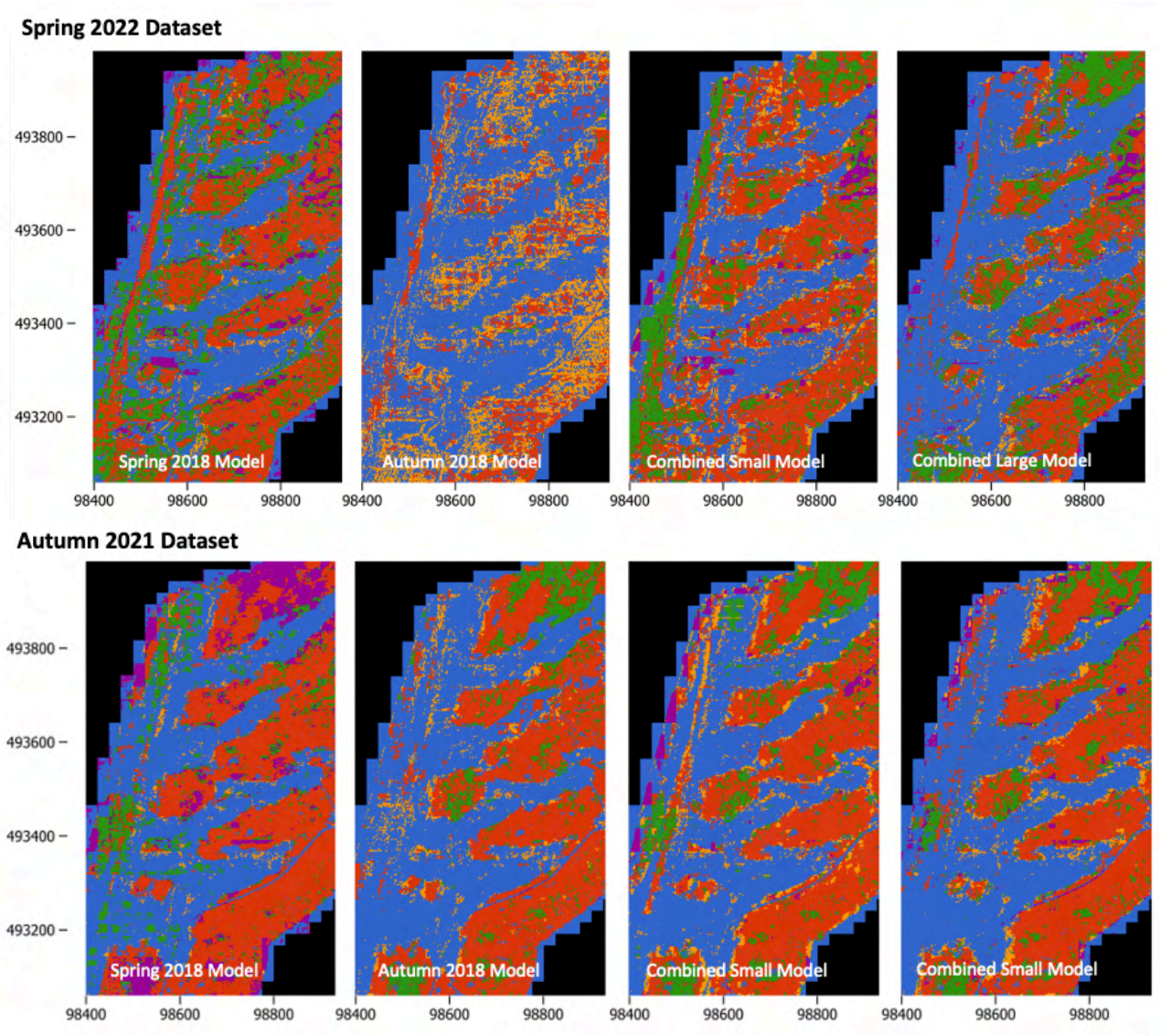


Figure 18: Stitched orthomosaics generated by each model fitted to the spring 2022 and autumn 2021 datasets.

Secondly, embryonic dunes (represented by the diagonal line starting at around 98400:493200 and ending at 984550:493900 in Figure 18) contain several predictions as 'Other' by each of the combined models in the autumn 2021 dataset. Curiously, for the combined small model and spring 2018 models applied to the spring 2022 dataset, we see many predictions for shrubs in that area, which is ecologically virtually impossible. This could be due to the textures in the sand from vehicles driving in that area, as well as shells on the beach (red rectangle of Figure 19), which appear to be consistent with the appearance of shrubs in spring (Figure 20). The combined small model managed to predict vehicle tracks (starting at 98450:493340 to 98470:493420) as 'Other', while the spring 2018 model did not. This is caused by most of the 'Other' habitat in the spring 2018 training dataset containing water (which is clear, and thus appears green), while for autumn 2018 this comprised mostly roads. As a result, the combined small model can make this distinction, despite never having been trained to recognize such patterns in sand as road before.

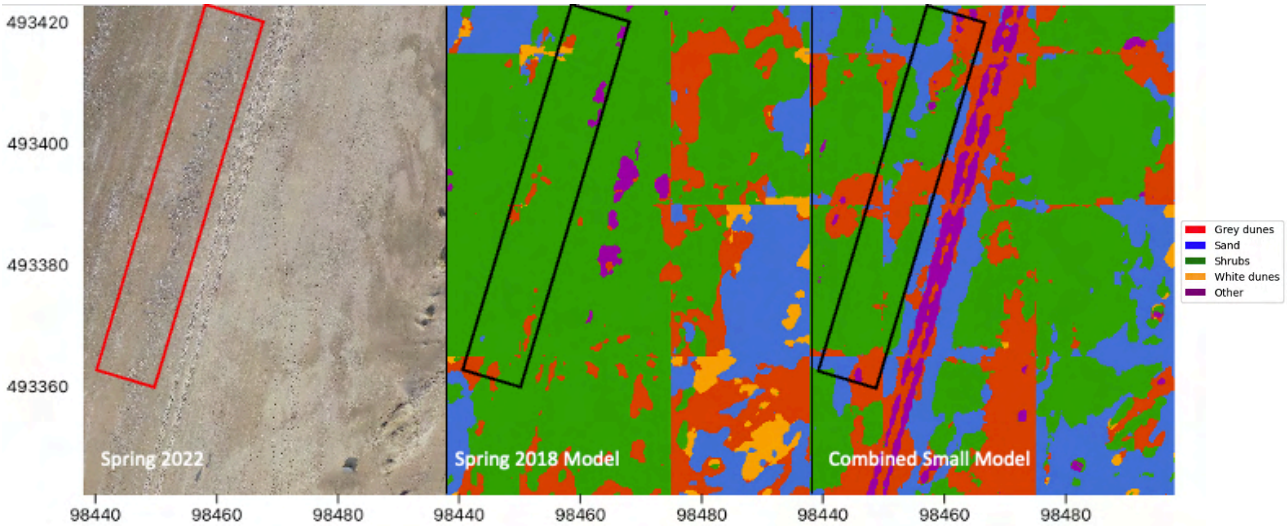


Figure 19: Predictions of spring 2022 data influenced by appearance of shrubs changing between seasons. The red/black rectangles represent either predicted or trained shrub regions.

Figure 20: Differences in appearance of shrubs between spring and autumn. The red rectangles highlight how the predictions of shrubs in the above figure could have come about.

5. Conclusion and Discussion

The purpose of this project was to investigate how drone images captured in different phenological phases or seasons affect the differentiability of habitat types in the Dutch dune environment. To test this, four neural network models were trained on different datasets (representing spring 2018, autumn 2018, a combined dataset, and a dataset comprising 50% of each of the spring 2018 and autumn 2018 datasets), and various analyses carried out.

Using a model that was trained on a specific season to predict a habitat map of its corresponding season yielded the best results, while the opposite was true for season-mismatched predictions. The combined models generally performed as well as, if not better than, seasonally matched models. Additionally, there were minimal performance differences between combined large and combined small models, which illustrates that increasing training data volume does not necessarily improve performance. Finally, the degree of model performance degradation from one period to another can vary quite drastically – for autumn 2018 to 2021 this was less than spring 2018 to 2022. However, lighting and weather effects could have impacted these results and to an extent confounded the impact of season alone. This brings into question the roles played by weather and lighting conditions and the importance of incorporating these conditions into training data.

The results of other studies (Kattenborn et al., 2019; Cruz et al., 2023) are similar, however their experimental setups were different. They both used digital elevation models (DEMs) together with RGB orthomosaics as inputs into their models, while in this study only RGB orthomosaics were used. Additionally, other studies achieved higher F1 scores of around 84% (Kattenborn et al., 2019). Finally, Cruz et al. (2023) also considered the role played by seasons and found that combined models performed better than season-specific models. For the current project, DEMs were not considered relevant because the AOI had been artificially disturbed (through the digging of notches in the dunes), which destroys relationships that could exist between elevation and habitat type. The higher accuracies of other studies stem from several factors. These include larger training set sizes, different approaches (feature extraction rather than HITLML, the former of which is much more time-consuming) and the combination of the use of DEMs with undisturbed natural sites.

This study has several implications. Firstly, it seems that having a more diverse training dataset, rather than an outright larger one, has a larger impact on model performance. Thus, when data collection resources are limited, it is more beneficial to annotate a greater variety of data rather than focussing on outright volume. Secondly, weather conditions have a rather substantial impact on model performance and should be factored into selection of training data samples, or perhaps even through data augmentation. This could be incorporated through casting shadows based on DEMs (sun), as well as blurring parts of the image (wind). Finally, the use of HITLML has benefits, through faster annotation times, however there are also drawbacks associated with it (most prominently, longer post-processing times). For future projects, sufficient resources should be allocated to best capitalize on these benefits and address the drawbacks (e.g., including team members with sufficient programming skills).

This study also highlights the need for future research using neural networks for remote sensing applications. The use of ensemble modelling (i.e., combining predictions of several models to create a final prediction) could be better understood in its application to this subject area. For instance, by comparing a single combined model with a set of ensemble models trained on different splits of that dataset (e.g., by weather conditions and/or season). Another factor worth looking into in this specific study area is, instead of using DEMs as previously discussed, the distance to the sea per pixel could be incorporated to potentially reduce ecologically impossible predictions (such as shrubs close to the sea).

Overall, this research suggests that the combination of CNNs and HITLML are promising tools for uses in automated habitat mapping. Care should be taken in sampling a sufficiently diverse training dataset (across seasons as well as weather conditions) to make the best use of human and computational resources, for the most reliable and useful results.

References

Bishop, C.M. (2006). *Pattern Recognition and Machine Learning*. Springer. https://link.springer.com/book/9780387310732

Buscombe, D., & Goldstein, E. B. (2022). A Reproducible and Reusable Pipeline for Segmentation of Geoscientific Imagery. *Earth and Space Science*, *9*(9), e2022EA002332. https://doi.org/10.1029/2022EA002332

Buscombe, D., Goldstein, E. B., Sherwood, C. R., Bodine, C., Brown, J. A., Favela, J., Fitzpatrick, S., Kranenburg, C. J., Over, J. R., Ritchie, A. C., Warrick, J. A., & Wernette, P. (2022). Human-in-the-Loop Segmentation of Earth Surface Imagery. *Earth and Space Science*, *9*(3), e2021EA002085. https://doi.org/10.1029/2021EA002085

Cohen, Jacob. (1960). *A Coefficient of Agreement for Nominal Scales*. https://journals.sagepub.com/doi/10.1177/001316446002000104

Creative Commons—Attribution 4.0 International—CC BY 4.0. (2023). Retrieved 16 June 2023, from https://creativecommons.org/licenses/by/4.0/

Cruz, C., O'Connell, J., McGuinness, K., Martin, J. R., Perrin, P. M., & Connolly, J. (2023). Assessing the effectiveness of UAV data for accurate coastal dune habitat mapping. *European Journal of Remote Sensing*, *56*(1), 2191870. https://doi.org/10.1080/22797254.2023.2191870

Diakogiannis, F. I., Waldner, F., Caccetta, P., & Wu, C. (2020). ResUNet-a: A deep learning framework for semantic segmentation of remotely sensed data. *ISPRS Journal of Photogrammetry and Remote Sensing*, *162*, 94–114. https://doi.org/10.1016/j.isprsjprs.2020.01.013

Eaton-Rosen, Z., Bragman, F., Ourselin, S., & Cardoso, M. J. (2018). *Improving Data Augmentation for Medical Image Segmentation*. https://openreview.net/forum?id=rkBBChjiG

He, K., Zhang, X., Ren, S., & Sun, J. (2015). *Deep Residual Learning for Image Recognition* (arXiv:1512.03385). arXiv. https://doi.org/10.48550/arXiv.1512.03385

Hesp, P. A. (1991). Ecological processes and plant adaptations on coastal dunes. *Journal of Arid Environments*, *21*(2), 165–191. https://doi.org/10.1016/S0140-1963(18)30681-5

Katal, N., Rzanny, M., Mäder, P., & Wäldchen, J. (2022). Deep Learning in Plant Phenological Research: A Systematic Literature Review. *Frontiers in Plant Science*, *13*. https://www.frontiersin.org/articles/10.3389/fpls.2022.805738

Kattenborn, T., Eichel, J., & Fassnacht, F. E. (2019). Convolutional Neural Networks enable efficient, accurate and fine-grained segmentation of plant species and communities from high-resolution UAV imagery. *Scientific Reports*, *9*(1), Article 1. https://doi.org/10.1038/s41598-019-53797-9

Kattenborn, T., Leitloff, J., Schiefer, F., & Hinz, S. (2021). Review on Convolutional Neural Networks (CNN) in vegetation remote sensing. *ISPRS Journal of Photogrammetry and Remote Sensing*, *173*, 24–49. https://doi.org/10.1016/j.isprsjprs.2020.12.010

Kumar, S., & Hebert, M. (2006). Discriminative Random Fields. *International Journal of Computer Vision*, *68*(2), 179–201. https://doi.org/10.1007/s11263-006-7007-9

LeCun, Y., Bottou, L., Bengio, Y., & Ha, P. (1998). *Gradient-Based Learning Applied to Document Recognition*.

Mosqueira-Rey, E., Hernández-Pereira, E., Alonso-Ríos, D., Bobes-Bascarán, J., & Fernández-Leal, Á. (2023). Human-in-the-loop machine learning: A state of the art. *Artificial Intelligence Review*, *56*(4), 3005–3054. https://doi.org/10.1007/s10462-022-10246-w

Natura2000. (2008a). *Duinen met Hippophaë rhamnoides (H2160)*. Natura2000. https://www.natura2000.nl/sites/default/files/profielen/Habitattypen_profielen/Profiel_habitattype_2160.pdf

Natura2000. (2008b). *Embryonale wandelende duinen (H2110)*. Natura2000. https://www.natura2000.nl/sites/default/files/profielen/Habitattypen_profielen/Profiel_habitattype_2110.pdf

Natura2000. (2008c). *Vastgelegde kustduinen met kruidvegetatie ("grijze duinen") (H2130)*. Natura2000. https://www.natura2000.nl/sites/default/files/profielen/Habitattypen_profielen/Profiel_habitattype_2130.pdf

Natura2000. (2008d). *Wandelende duinen op de strandwal met Ammophila arenaria ("witte duinen") (H2120)*. Natura2000. https://www.natura2000.nl/sites/default/files/profielen/Habitattypen_profielen/Profiel_habitattype_2120.pdf

Oldeland, J., Revermann, R., Luther-Mosebach, J., Buttschardt, T., & Lehmann, J. R. K. (2021). New tools for old problems—Comparing drone- and field-based assessments of a problematic plant species. *Environmental Monitoring and Assessment*, *193*(2), 90. https://doi.org/10.1007/s10661-021-08852-2

Pearse, G. D., Watt, M. S., Soewarto, J., & Tan, A. Y. S. (2021). Deep Learning and Phenology Enhance Large-Scale Tree Species Classification in Aerial Imagery during a Biosecurity Response. *Remote Sensing*, *13*(9), Article 9. https://doi.org/10.3390/rs13091789

Podareanu, D., Codreanu, V., Aigner, S., Leeuwen, C., & Weinberg, V. (2019). *Best Practice Guide—Deep Learning*. https://doi.org/10.13140/RG.2.2.31564.05769

Ronneberger, O., Fischer, P., & Brox, T. (2015). U-Net: Convolutional Networks for Biomedical Image Segmentation. In N. Navab, J. Hornegger, W. M. Wells, & A. F. Frangi (Eds.), *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015* (pp. 234–241). Springer International Publishing. https://doi.org/10.1007/978-3-319-24574-4_28

Ruessink, B. G., Arens, S. M., Kuipers, M., & Donker, J. J. A. (2018). Coastal dune dynamics in response to excavated foredune notches. *Aeolian Research*, *31*, 3–17. https://doi.org/10.1016/j.aeolia.2017.07.002

Ruessink, G. (2023). *Topographic data and orthomosaics of the Noordwest Natuurkern project* (v1.1.1) [Data set]. Zenodo. https://doi.org/10.5281/zenodo.7970837

Sokolova, M., & Lapalme, G. (2009). A systematic analysis of performance measures for classification tasks. *Information Processing & Management*, *45*(4), 427–437. https://doi.org/10.1016/j.ipm.2009.03.002

Sudre, C. H., Li, W., Vercauteren, T., Ourselin, S., & Cardoso, M. J. (2017). *Generalised Dice overlap as a deep learning loss function for highly unbalanced segmentations* (Vol. 10553, pp. 240–248). https://doi.org/10.1007/978-3-319-67558-9_28

Vitousek, S., Buscombe, D., Vos, K., Barnard, P. L., Ritchie, A. C., & Warrick, J. A. (2023). The future of coastal monitoring through satellite remote sensing. *Cambridge Prisms: Coastal Futures*, *1*, e10. https://doi.org/10.1017/cft.2022.4

Zhang, Z., Liu, Q., & Wang, Y. (2018). Road Extraction by Deep Residual U-Net. *IEEE Geoscience and Remote Sensing Letters*, *15*(5), 749–753. https://doi.org/10.1109/LGRS.2018.2802944

Appendix 1: Configuration file

All models were trained using the same configuration files, however the only difference between files is the name given to the model. The spring 2018 model config file is presented here as an example.


```
{
  "TARGET_SIZE": [512,512],
  "MODEL": "resunet",
  "NCLASSES": 5,
  "KERNEL":9,
  "STRIDE":2,
  "BATCH_SIZE": 6,
  "FILTERS":6,
  "N_DATA_BANDS": 3,
  "DROPOUT":0.1,
  "DROPOUT_CHANGE_PER_LAYER":0.0,
  "DROPOUT_TYPE":"standard",
  "USE_DROPOUT_ON_UPSAMPLING":false,
  "DO_TRAIN": true,
  "LOSS":"dice",
  "PATIENCE": 10,
  "MAX_EPOCHS": 100,
  "VALIDATION_SPLIT": 0.5,
  "RAMPUP_EPOCHS": 20,
  "SUSTAIN_EPOCHS": 0.0,
  "EXP_DECAY": 0.9,
  "START_LR":  1e-7,
  "MIN_LR": 1e-7,
  "MAX_LR": 1e-4,
  "FILTER_VALUE": 0,
  "DOPLOT": true,
  "ROOT_STRING": "spring_v2_resunet_512",
  "USEMASK": false,
  "AUG_ROT": 5,
  "AUG_ZOOM": 0.05,
  "AUG_WIDTHSHIFT": 0.05,
  "AUG_HEIGHTSHIFT": 0.05,
  "AUG_HFLIP": true,
  "AUG_VFLIP": false,
  "AUG_LOOPS": 10,
  "AUG_COPIES": 5,
  "SET_GPU": "0",
  "WRITE_MODELMETADATA": false,
  "DO_CRF": false,
  "LOSS_WEIGHTS": false,
  "MODE": "all",
  "SET_PCI_BUS_ID": true,
  "TESTTIMEAUG": true,
  "WRITE_MODELMETADATA": true,
  "OTSU_THRESHOLD": true,
  "TF_GPU_ALLOCATOR" : "cuda_malloc_async",
  "CLEAR_MEMORY" : true
}
```

Appendix 2: Scripts Used

The following scripts were used in this project, and have been made available in the following link: https://github.com/murson/ads_thesis

Name	Description
cohens_kappa.ipynb	Calculate Cohen's kappa inter-annotator reliability between image labels.
download_images.ipynb	Downloads and clips all source images.
retiling_and_file_mgt.ipynb	Retiles source images, as well as performs file operations (renaming, sampling, etc).
img_comparison.ipynb	Creating the visualization in Figure 4.
model_histories.ipynb	Plot model training histories in Figure 13.
confusion_matrix.ipynb	Creation of confusion matrices, f1 score visualization, testing & training dataset land cover analysis.
stitching.ipynb	Stitching of predicted labels into orthomosaics.
img_analysis.ipynb	Analysis of which labels contain black pixels, and then recolouring those as purple, as well as fixing the .npz file produced by DashDoodler to this effect.