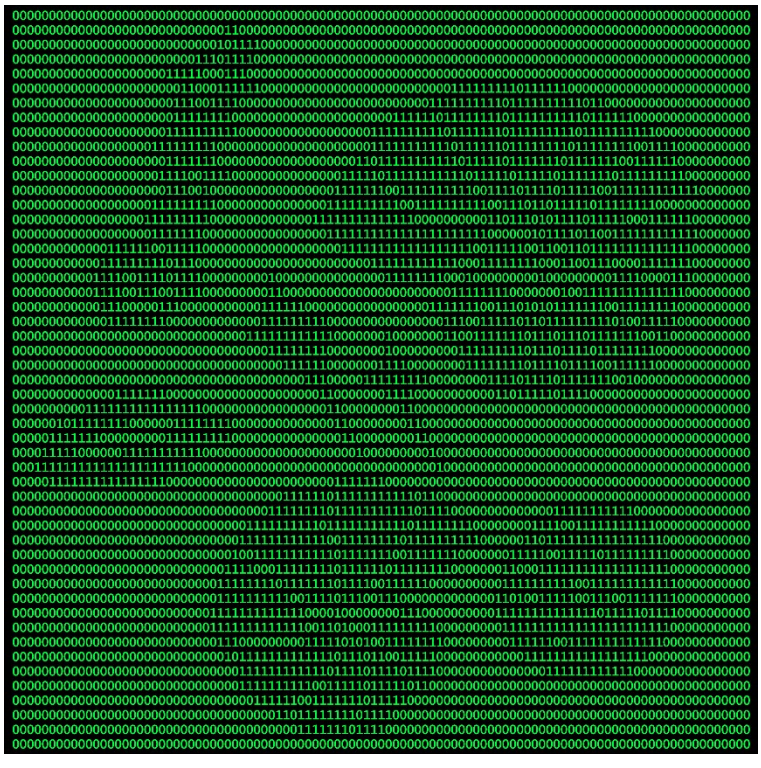
UNIVERSITY OF GOTHENBURG

DEPARTMENT OF BIOLOGICAL AND ENVIRONMENTAL SCIENCES

IDENTIFICATION OF ENVIRONMENTALLY RELEVANT BENTHIC FORAMINIFERA FROM THE SKAGERRAK FJORDS BY DEEP LEARNING IMAGE MODELING



Marko Plavetić

Degree project for Master of Science (120 HEC) with a major in Environmental Science

ES2510, Degree project in Environmental Science, 60 HEC

Second cycle

Semester/year: 2023 Spring

Supervisors: Irina Polovodova Asteman, Department of Marine Sciences;
Allison Yi Hsiang, University of Stockholm, Department of Geological Sciences;
Mats Josefson, Oral Product Development, AstraZeneca

Examiner: Lennart Bornmalm, Department of Marine Sciences

Table of Contents

List of abbreviations

ANN – Artificial Neural Network

AP - Average Precision

API - Application Programming Interface

CNN – Convolutional Neural Network

COCO – Common Objects in Context

$ExpH_{bc}$ – Hill's number diversity index

EQS – Ecological quality status

Foram-AMBI – Foraminiferal AZTI Marine Biotic Index

GPU – graphical processing unit

FOBIMO - FOraminiferal BIo-MOnitoring

IoU – Intersection over Union

mAP - mean Average Precision

mAP@.5 - mean Average Precision at IoU 0.5 threshold

mAP@.5:.95 – mean Average Precision at IoU 0.5 to 0.95 threshold

ML - Machine learning

MSFD - Marine Strategy Framework Directive

NQI_f – Norwegian Quality Index (using Foraminifera)

P - Precision

R - Recall

SML - Supervised Machine Learning

TOC – Total organic carbon

WFD - Water Framework Directive

YOLO – You Only Look Once

Abstract

Over the several past decades, there has been increasing interest in using foraminifera as environmental indicators for coastal marine environments. As compared to macrofauna, which are currently used in environmental studies, foraminifera offer several distinct advantages as bioindicators, including short generation times, a high number of individuals per small sample volume, hard and durable tests with high preservation potential, and low cost of sample extraction. One of the main problems with foraminifera identification is reliance on manual identification and expert judgement, which is a tedious and slow process prone to errors and subjectivity. Deep learning, a subfield of machine learning, has emerged as a promising solution to this challenge, since a neural network can learn to recognize subtle differences in shell morphology that may be difficult for the human eye to distinguish. Benthic foraminifera mounted on microslides from several Skagerrak fjords including Gullmar Fjord, Hakefjord, and Idefjord were imaged using a Nikon SMZ-10 stereomicroscope and DeltaPix DP450 microscope camera. Images were then processed in Roboflow API, where individual foraminifera were labelled and classified. This resulted in 3003 images and 22 138 labelled individuals. Using the labeled images, a dataset was created to be used for deep learning training. YOLO (You Only Look Once) v7 model implemented in the PyTorch framework was used in this work, which has demonstrated state-of-the-art speed and performance for object detection as of the time of writing. Models were trained using a Nvidia RTX A4000 GPU (graphical processing unit). The models are able to distinguish 29 species, while preliminary results show a 90.3% mAP (mean average precision) and 78.8% mAP on the best and the worst performing models, respectively. Even though the imaging and labelling was done in a short amount of time, the results look promising and show that even a relatively small dataset can be used for training a reliable deep learning species identification model.

Keywords: **benthic foraminifera, deep learning, environmental monitoring, YOLOv7**

1. Introduction

Coastal areas are an ever-changing and subject to environmental pressures from both the land and the sea. As such, it is important to have reliable, fast, and easy to use bioindicators to assess the environmental state and health of coastlines. Benthic foraminifera can be used for monitoring of the environmental state and for assessing the reference (pre-disturbance) conditions of a given environment. However, due to the long and tedious process of manual identification and presence of other accepted bioindicators (e.g., macrofauna), foraminifera have not yet been widely accepted in environmental monitoring by governmental organizations.

1.1 Foraminifera

Foraminifera are a diverse group of cosmopolitan marine single-celled protozoans. Depending on the part of the water column they inhabit, one can differentiate between planktonic and benthic foraminifera.

Planktonic foraminifera are floating organisms that live in the upper part of the water column. They have a test (shell) of globular shape made from calcium carbonate. There are about 50 extant species, and their test size is typically smaller than 1 mm.

Benthic foraminifera, on the other hand, as the name implies, belong to the benthos, organisms dwelling on the ocean floor. Unlike planktonic foraminifera, benthic foraminifera have a diverse range of body plans and test sizes, and their tests can be made from calcium carbonate, or out of glued together particles of surrounding sediment. There are about 10 000 extant species of benthic foraminifera, and their test sizes range from 0.5 mm up to 20 cm (Boersma 1998).

1.2 Benthic foraminifera in environmental monitoring

In the year 2000, the European Union member states, European Commission and Norway, agreed to implement a new framework concerning water environments. The Water Framework Directive (WFD) aims to better manage, preserve, and protect European water environments (WFD 2000). The directive produced a framework for the long-term protection of all water resources and was later supported by the Marine Strategy Framework Directive (MSFD 2008). Ecological assessment is based on the status of biological, physicochemical, and hydro morphological quality elements. At the time of implementation of MSFD, the biological elements which are used as indicators were phytoplankton, macroalgae, benthic macroinvertebrates, and fish (Borja *et al.* 2009).

According to Schönfeld *et al.* (2012), good bioindicators are specific to certain habitats and should have fast turnover rates. Foraminifera fit those criteria and offer several key advantages compared to macrofaunal organisms, including short generation times, high number of individuals per small sample volume, hard and durable tests with high fossilization potential, and low cost of sample extraction (Alve 1995; Alve *et al*. 2009). Benthic foraminifera have been used as proxies for different kinds of pollutants, including heavy metals and hydrocarbons in a wide range of marine environments. Effects of pollution can be visible in benthic foraminiferal population by observing changes in: faunal distribution and diversity, local disappearance, development of abnormal tests, and increase of opportunistic species (Alve 1991; Yanko *et al*. 2003). Furthermore, shells are preserved after death in most species of foraminifera, and thus can inform on reference conditions of preindustrial times.

Due to non-standardized methods of sample acquisition, preparation, and data processing, benthic foraminifera have remained a marginal environmental monitoring tool. In recent years there has been a push towards more standardizing monitoring methods, and the first such attempt was done by the FOBIMO initiative (FOraminiferal BIo-MOnitoring). FOBIMO standardized sampling devices, sampling intervals, sampling depth, sample processing, and sample acquisition (Schönfeld *et al.* 2012). After standardizing sample acquisition and processing, scientists started working on the development of foraminiferal diversity and sensitivity indices. Some examples of these include Foram-AMBI, Exp (H_{bc}), NQI_f etc. Sensitivity indexes like Foram-AMBI are used for determining the sensitivity of species to environmental stressors, and in Foram-AMBI they are given five categories depending on their tolerance to total organic carbon (TOC). Benthic foraminiferal species are assigned into one of five ecological groups:

Group I (EGI): "sensitive species" are sensitive to TOC enrichment. Their relative abundance is highest at the lowest TOC values and drops to zero as organic carbon concentration increases.

Group II (EGII): "indifferent species" are indifferent to the initial stages of organic carbon enrichment and never dominate the assemblage. They occur in low relative abundance over a broad range of organic carbon concentrations but are absent at very high concentrations.

Group III (EGIII): "tolerant species" are able to endure excess organic carbon enrichment. They may occur at low TOC; their highest frequencies are stimulated by organic carbon enrichment, but they are absent at very high organic carbon concentrations. This group has been termed "third-order opportunistic species".

Group IV (EGIV): "second-order opportunistic species" show a clear positive response to organic carbon enrichment with maximum abundances between the maxima of EGIII and EGV.

Group V (EGV): "first-order opportunistic species" exhibit a clear positive response to excess organic carbon enrichment with maximum abundances at a higher stress level induced by organic load than species belonging to EGIV. At even higher TOC concentrations, foraminifera are not able to survive.

Foraminiferal diversity indices on the other hand, allow for monitoring the present and past human environmental impact, based on species diversity of a given study area (O'Brien *et al.* 2021).

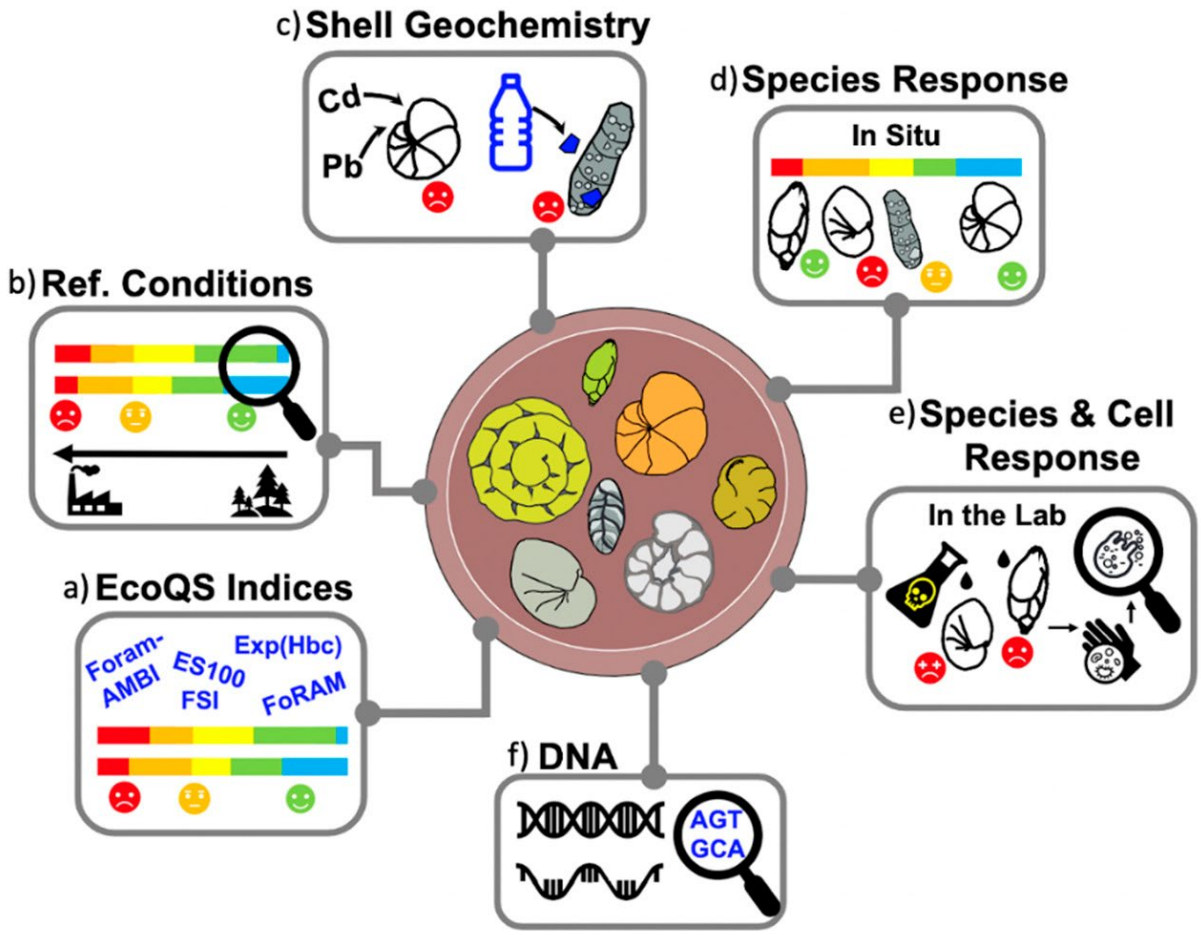


Figure 1. A general overview of foraminiferal applications in biomonitoring studies. a) EcoQS indices based on species sensitivity and diversity; b) reconstruction of preindustrial reference conditions; c) accumulation of pollutants within foraminiferal tests; d) species response to environmental pressures in situ; e) simulated species response in the laboratory; and f) genomic methods such as eDNA (O'Brien *et al.* 2021).

1.3 Aim

Benthic foraminifera can be used for monitoring environmental state and for assessing the reference conditions of a given environment (Fig. 1), but due to long and tedious process of manual picking and species identification, they have not yet been widely used in environmental monitoring studies. This thesis aims to reduce the time necessary for identification and increase the accuracy of identifications by the use of machine learning, specifically deep learning convolutional neural networks.

1.4 Benthic foraminifera used in this thesis

A total of 59 species were identified in the samples from the fjords on the west coast of Sweden including Gullmar Fjord, Idefjord and Hakefjord. Of those 29 species were chosen for training the machine learning model in this study. Gullmar Fjord is a fjord situated in the Bohuslän Province, about 80 km north of the city of Gothenburg. The fjord is about 30 km long and between 1-2 km wide, with a maximum water depth of 119 meters. Idefjord is a fjord that is located at the national border between Norway and Sweden. It is about 150 km north of the city of Gothenburg, and it is about 25 km long and approximately 1 km wide, with a maximum depth of 48 meters. Finally,

Hakefjord is the most southern fjord of the Orust-Tjörn fjord system, located to the east of island Tjörn. It is about 16 km long and 2-5 km wide with a maximum depth of 35 meters.

The most important foraminifera species for the afore-mentioned fjords are the native Skagerrak-Kattegat (S-K) fauna species including *Cassidulina laevigata, Textularia earlandi, Bulimina marginata, Liebusella goesi, Hyalinea balthica* and *Nonionellina labradorica* (Nordberg *et al*. 2000). These species are commonly recorded in high abundances in the adjacent Skagerrak and Kattegat straits. The S-K fauna species usually prevail at sites with salinities >30 psu below the pycnocline (Filipsson & Nordberg 2004). Recently an invasive species *Nonionella* sp. T1 has been discovered in the Skagerrak (Polovodova Asteman & Schönfeld 2015; Deldicq *et al.* 2019) and prefers nitrate-rich sediment at shallower water (30 - 40 m) depth with salinities of 32-34 psu (Choquel *et al.* 2021). Another important species is *Stainforthia fusiformis*, which has an ability to outcompete the native fauna with an opportunistic lifestyle under hypoxic conditions (Alve 2003). In general, S-K fauna showed a general tendency to decline as a response to bottom water hypoxia as a result of reduced bottom water exchanges in Gullmar Fjord.

Below are summarized the most important ecological preferences of the aforementioned species.

The species *H. balthica* is considered to be a cold to temperate water species in the North Atlantic on muddy bottoms (Ross 1984). Another limiting factor in the distribution of this hyaline foraminifera species are nutritional preferences and corresponding competition with more opportunistic species, which may explain why it was missing in parts of the record (Polovodova Asteman & Nordberg 2013).

The hypoxia intolerant *C. laevigata* becomes opportunistic under hyperoxic conditions with rapid growth and reproductive rates. This species is found in sediment depths of up to 14-15 cm in the Skagerrak (Murray 2003).

The two agglutinated species of the S-K fauna are opportunistic and widely spread and abundant species. *Textularia earlandi* and *L. goesi* have an omnivorous lifestyle and are both agglutinated (Polovodova Asteman & Nordberg 2013). *Textularia earlandi* increased in an unfed experiment after 2 years (Alve 2010) and its distribution is suggested to be controlled by sediment depth rather than food availability (Duffield *et al.* 2014). For the best stabilization of the test *L. goesi* uses different sized particles depending on the chamber and sediment (Hari *et al.* 2020).

The detritivore *B. marginata*, a detritivore species, is a part of the infauna of the Skagerrak dwelling down to 13-15 cm sediment depth. The foraminiferal surface water assemblage varies regionally while the deep-water assemblages in Scandinavia are relatively constant with abundant specimens of *C. reniforme*, *Elphidium sp*. and *N. labradorica* (Murray 2006). *Nonionellina labradorica* occurs simultaneously with *Nonionella* sp. T1 in Oslo fjord and is generally a widespread species (Deldicq *et al*. 2019). The species is common in salinities >34 psu in cold Scandinavian fjords and other cold-water regions (Murray 2006).

Invasive *Nonionella* sp. T1

The first record of *Nonionella* sp. T1 in Gullmar Fjord was from 2011, though its first appearance can be traced in the sediment core to sometime around 1985, afterwards it spread northwards to Oslofjord. The reference species with the closest morphological resemblance is *N. stella* from the San Pedro Basin, although specimens from Oslofjord and California are genetically not the same as is confirmed by DNA studies (Deldicq *et al.* 2019). Both *N. stella* and *N.* sp. T1 can be recognized morphologically via the expansion of the last chamber, which covers the umbilicus and resembles a hand. *Nonionella* sp. T1 is thought to have similar benefits from hypoxic conditions as the opportunistic *S. fusiformis*, however when comparing two sites with documented hypoxia and not, *Nonionella* sp. T1 was more abundant at well-oxygenated site (Choquel *et al.* 2021). This is likely explained by the *Nonionella* sp. T1 ability to perform a complete denitrification (Choquel *et al*. 2021), which requires presence of oxygen.

Opportunistic *Stainforthia fusiformis*

Stainforthia fusiformis is categorized as a ubiquitous and infaunal species with a thin test and an opportunistic lifestyle in the North Sea at salinities of >28 psu (Alve, 2003). The low oxygen tolerance of *S. fusiformis* enables the species to thrive and outcompete the natural S-K fauna under low oxygen conditions. The success stems from the ability to store copious amounts of nitrate and perform a complete denitrification (Risgaard-Petersen *et al.* 2006). This species' abundance and dominance over the native foraminifera species of the Gullmar Fjord is indicative of its hypoxic conditions. Comparison between the native S-K fauna and *S. fusiformis* can be used to reconstruct paleoenvironmental conditions (Filipsson & Nordberg 2004).

Other relevant species Ten more species were selected for identification in this study and those are described below.

● *Bolivina pseudopunctata* was recorded to occur under similar environmental conditions as the opportunistic *S. fusiformis*. It has a thinner test and is relatively small in size. *Bolivina* species together with *Textularia* and *Bulimina* species can commonly be found at oxygen deficient sites (Alve 1995; Bernhard & Alve 1996; Nordberg *et al.* 2000)

● *Eggerelloides scaber* requires salinities >24 psu and has no dependence on specific substrate types (Luze *et al.* 1983)

● *Eggerelloides medius'* test is rougher in texture while *E. scabers'* test is more elongated. *E. medius* is found in muddy sediment deeper than 40 m (Murray 2003).

● *Nonionella iridea* flourishes in greater depths within the sediment and displays opportunistic behaviour in the presence of phytodetritus (Duffield *et al.* 2014).

● *Nonionella turgida* is a calcareous species with an increasing concentration in Gullmar Fjord after 1990 (Filipsson & Norberg 2004). This species together with *S. fusiformis* shows an affinity for chlorophyll A in surface layers (Murray 2006)

● *Epistominella vitrea* is a cosmopolite and opportunistic species commonly found in deep sea and it associated with phytodetritus deposition on the sea floor (Murray 2006; Pawlowski *et al*. 2007).

● *Quinqueloculina stalkeri* is agglutinated species, which can be indicative of glaciomarine fjords with muddy sediment (Filipsson & Norberg, 2004; Korsun & Hald 1998).

● *Adercotryma glomerata* is an agglutinated foraminifera often discovered in temperate fjords and areas such as Sweden, Norway, and Scotland (Murray 2006). Its salinity tolerance range is considered as large (28-35 psu) (Polovodova Asteman *et al.* 2011) and it has disappeared from the heavily polluted Idefjord in connection with the maximum effluent discharges from the pulp and paper mill (Polovodova Asteman *et al.* 2015).

● *Brizallina skagerakensis* is a hyaline species that can be found in the outer fjord areas (Murray, 2006) and has been associated with increased primary productivity in the Skagerrak and Norwegian fjords (Duffield *et al*. 2014; Polovodova Asteman *et al.* 2018).

● *Elphidium excavatum* is commonly observed in temperate fjords with little water turbidity. The species lives within the shallow brackish water condition with sandy sediment together with *E. scaber.* Experiments have been conducted on these species revealing their ability to survive 24 h of anoxia (Murray 2006). It has also been shown that the species is sensitive to heavy metal pollution (Lintner *et al.* 2021).

2. Background on machine learning

Machine learning (ML) is a field in computer science devoted to understanding and developing methods that can give machines the ability to "learn". In this background, I will mainly focus on a subset of machine learning, called **supervised machine learning** (SML). The SML methodology involves using a given set of identified and labeled images, commonly referred to as a training set, to construct a machine learning model that learns the correlation between an input image and its corresponding classification. To be able to do this, a machine learning model must be able to determine which parts of the image are relevant to the object detection task at hand. The process of determining which parts in an image are relevant is called feature extraction. Feature extraction is a form of dimensionality reduction in which the complexity of the data is reduced into a set of simpler explanatory variables that are grouped using similarity or distance metrics. After obtaining the appropriate correlation between an input image and classification, the fitted model is used to classify a second dataset called the validation dataset. Validation dataset provides an unbiased evaluation of the model, since the model has not seen those images before while tuning the model's hyperparameters. Finally, a test dataset is a dataset used to evaluate the accuracy of the final model. The test dataset was not used in training or validation stages of model fitting, so it is a completely unbiased measure of the model's performance.

Artificial neural networks (ANNs) are used in machine vision applications. ANNs consist of a cluster of neurons or nodes and connections in between those neurons. When there are numerous neurons connected, the input for the following neuron is the output of the previous neuron. Each connection is assigned an associated weight, based on the relative importance of the input. In a neuron, computation is carried out by determining the summation of its inputs. The computation a neuron makes is known as an activation function. It is the activation function that provides the non-linear modelling ability for ANNs. Some of the common activation functions are: a) sigmoid, which maps the input to a value between 0 and 1; b) Rectified Linear Unit (ReLU), which replaces all negative input values with zero while leaving positive values unchanged; and c) Tanh, which maps the input to a value between -1 and 1. Both, sigmoid and Tanh suffer from a vanishing gradient problem, which is why ReLU is the most popular activation function in machine vision tasks (Krizhevsky *et al.* 2012). The output of the neuron is passed to the next adjacent neuron. Neural networks used in machine vision are typically feed-forward networks, meaning that all connections flow in a single forward direction. This implies that all neurons in a single layer have no connections between each other, but only with neurons in preceding and following layers (Fig. 2). For a long time, the most commonly used type of feed-forward ANN was the multilayer perceptron, and in this type of ANN, every neuron in a fully connected layer has connections to every neuron in the preceding layer (Castro *et al.* 2017).

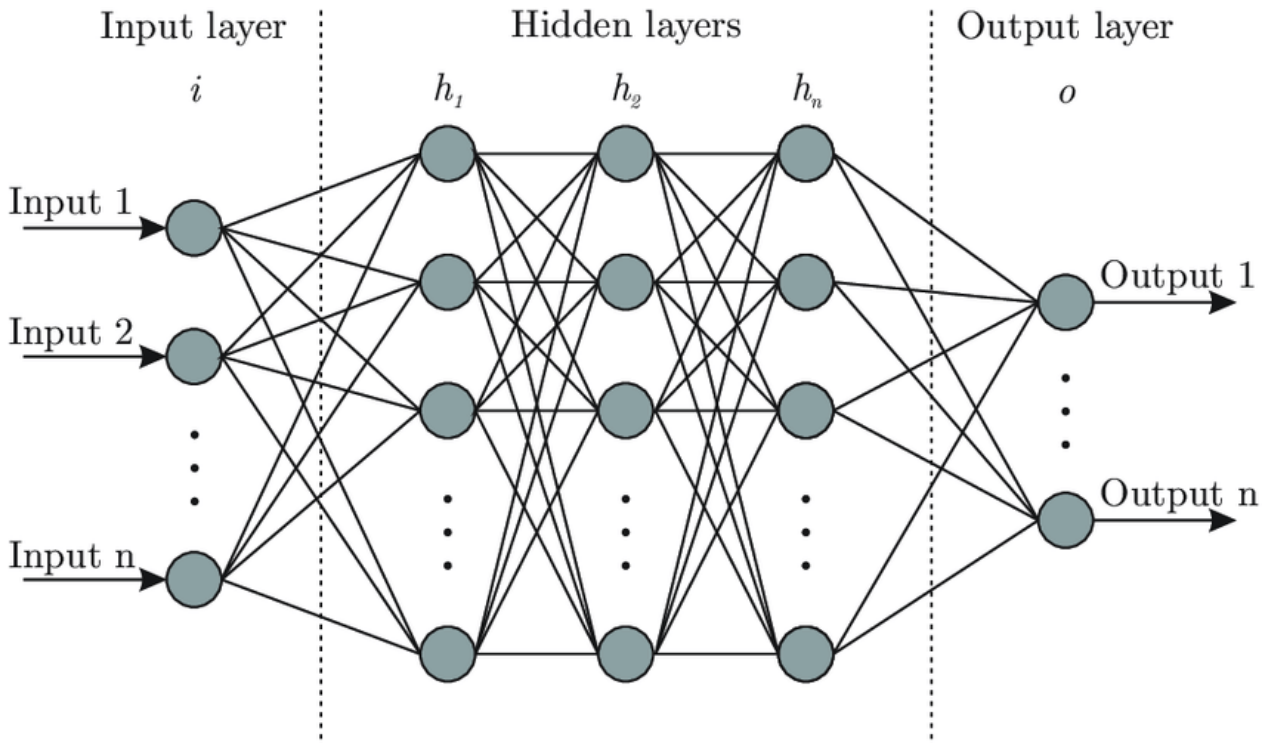


Figure 2. Artificial neural network architecture, which has an input layer, a set of hidden layers and an output layer. In each hidden and output layer, there are neurons interconnected via adaptive weights. These weights are calibrated through the training process (Bre *et al.* 2017).

Convolutional neural networks (CNNs) are currently the state-of-the-art algorithms for image classification and feature extraction. CNNs extend to ANNs by including several layers that perform convolutions, which are used to extract specific features from the image. Each image can be represented by a matrix of values for each pixel. Convolutions use these pixel values to compute new values by using element-wise matrix multiplication with a smaller matrix called a filter or a kernel that operates over the original pixel values (Fig. 3). The sum of the element-wise multiplication with a filter matrix and the original pixel matrix results in a new matrix of convolved features or also known as a feature map. Examples of feature maps are horizontal and vertical edge detection, sharpening, and blurring (Fig. 4). Convolutional operations are inherently linear, but linear functions are limited in their ability to map relationships between the input and the output. Also in CNN, an activation function is used to introduce nonlinearity into the network. After applying an activation layer, a pooling layer is applied to reduce the dimension or in other words to downsample the input image. This removes all unimportant information, while preserving features that are relevant for identification. Common pooling layers are max pooling, where the highest value in a given area of a pixel is retained, and average pooling, where the average value of the pixel in a given area is calculated and retained (Dumoulin & Visin 2016).

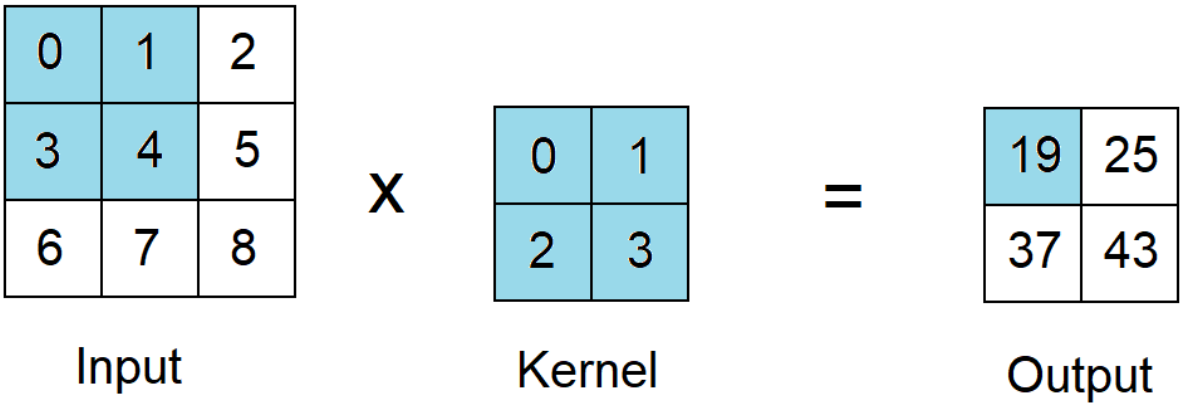


Figure 3. Two-dimensional convolution operation on a 3×3 matrix using a 2×2 kernel. The shaded portions are the first output element as well as the input and kernel elements used for the output computation: 0×0 + 1 × 1 + 3 × 2 + 4 × 3 = 19. Convolution is performed by moving the filter across the input image and summing the values from element-wise multiplication. These sums create a new matrix that corresponds to a convolved feature, which is known as an "activation map" or a "feature map".

A training set is a dataset of labelled images, which provides a correct mapping of pixel values and weights and the final classification. When the CNN is first initialized, all the weights and filters are randomly selected. Then the network takes the input image and performs a forward pass through the network. The total error of the pass is calculated. The network then performs a backpropagation, a process that uses gradient descent to update the weights and filters to minimize the total error. One forward propagation and one backpropagation is called an epoch. Since the training system's memory is limited, we are forced to split a dataset into smaller batches (ideally, we would load the whole dataset into the system, but this is not currently possible due to hardware limitations). In general, a larger batch is better than a smaller one. The number of batches needed to complete one epoch is called the number of iterations. The number of epochs varies based on the dataset and parameters set at the start of the training. By using backpropagation to update weights and kernels, the network learns how to classify the training images. The accuracy of the model is then tested by using a validation set of images that the model has not yet seen, as this gives us an idea of what sort of accuracy to expect when classifying completely new images. The performance of the model is evaluated using the validation accuracy i.e., the proportion of correctly identified objects in an image and the validation loss function, which is a sum of all errors for each image in a validation set (Fei-Fei *et al.* 2007, Lecun *et al.* 2010).

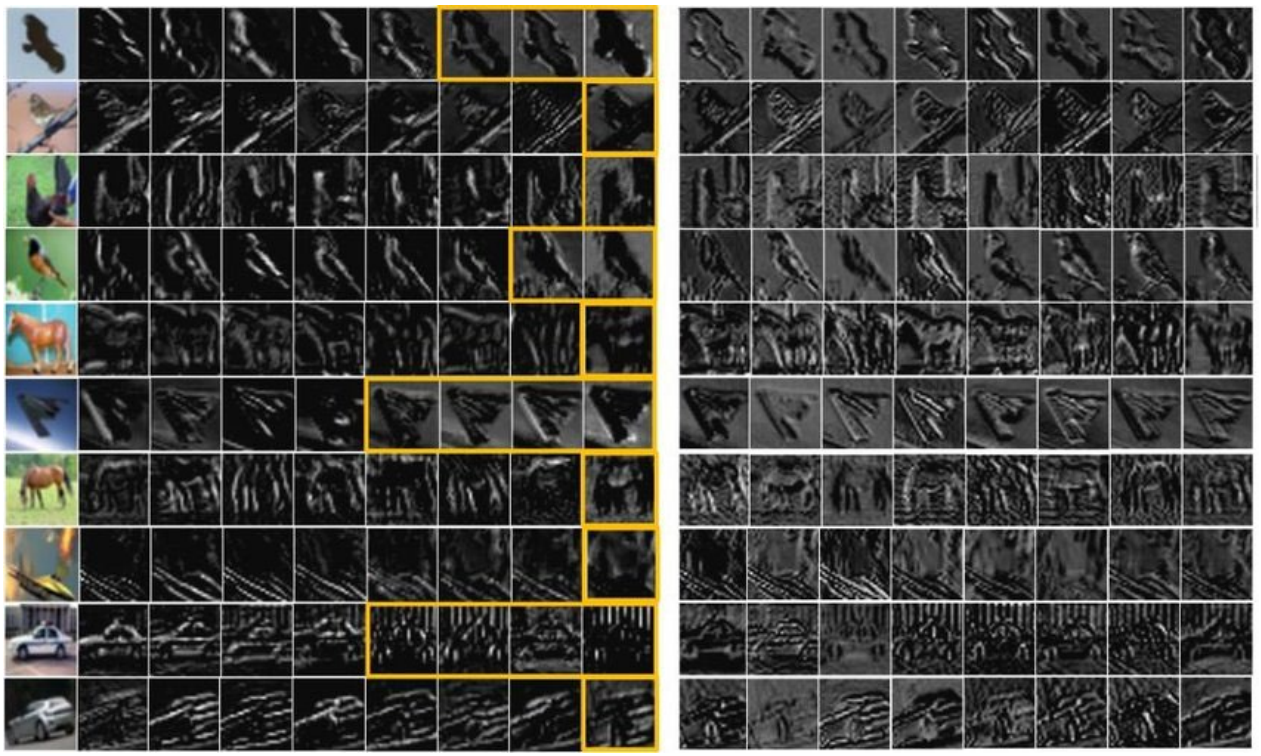


Figure 4. Visualization of a feature map. From left to right: Original image; feature maps of the background; and the feature map of the object (Sun *et al.* 2018).

You Only Look Once or YOLO is a family of machine vision deep learning models used for object detection (Redmon *et al.* 2015). They have consistently been the fastest and most accurate models since their introduction back in 2015. Since then, it has had 8 iterations of which the latest one is YOLOv8. In this project, YOLOv7 was used. YOLOv7 is a single-shot detector, which uses reparametrized convolutions and model scaling for improving the speed of the detector. A single shot detector uses a different approach compared to traditional object detection algorithms. Instead of using a sliding window approach where a smaller window is passed through an image to scan for objects, single shot detector divides the image into an x-by-x grid and have each grid cell be responsible for detecting objects in that region of the image. After identifying objects, an anchor box is assigned to the detected object. During training, anchor boxes are manipulated in size so that they resemble ground truth bounding boxes as close as possible. It has several versions; 2 P5 models and 4 P6 models respectively (Wang *et al*. 2022). P5 models take 640×640 pixels as their input image size, while P6 models take 1280×1280 pixels. If the image is bigger than the input size, a sliding window approach is used to pass through the whole image. In Table 1 the differences between each model are discussed in detail. In this thesis, 3 models were used, YOLOv7, YOLOv7x, and YOLOv7E6E. P5 models were chosen first since they are less resource intensive compared to P6 models, but one P6 model was included for comparison since it performed better on the Common Objects in Context (COCO) database. The COCO database contains 300 thousand images and 1.5 million objects (Lin *et al.* 2014) and is commonly used for training object detection deep learning models.

To reduce the computational cost and train robust models with relatively small datasets and relatively resource limited machines, a so-called transfer learning method can be used to train the model. In transfer learning, the weights of a model previously trained on a different dataset are used for a new task; the frozen top layer is used as a starting point for a new model. This reduces the time required to train the new model because it has already learned low level general features (e.g., edges) from the previous dataset. This method allows a model to be trained by using thousands of images, rather than millions of images (He *et al.* 2015). In the case of YOLOv7 a COCO database was used for obtaining the starting weights before training the model.

Table 1. A comparison of YOLOv7 models performance on the COCO dataset.

Model	Type	Test Size	APtest	AP50test	AP75test	batch 1 fps	batch 32 average time
YOLOv7 P5	P5	640	51.40%	69.70%	55.90%	161 fps	2.8 ms
YOLOv7-X	P5	640	53.10%	71.20%	57.80%	114 fps	4.3 ms
YOLOv7-W6	P6	1280	54.90%	72.60%	60.10%	84 fps	7.6 ms
YOLOv7-E6	P6	1280	56.00%	73.50%	61.20%	56 fps	12.3 ms
YOLOv7-D6	P6	1280	56.60%	74.00%	61.80%	44 fps	15.0 ms
YOLOv7-E6E	P6	1280	56.80%	74.40%	62.10%	36 fps	18.7 ms

In this thesis, a relatively large dataset of benthic foraminifera with associated labels was created with the help of a taxonomic expert. This dataset was then used to train a supervised machine learning object detection classifier by using deep CNNs that can automatically identify benthic foraminifera within an image with accuracies comparable to an identification by human experts.

3. Materials and methods

3.1 Image acquisition

Sediment samples prepared or picked for benthic foraminifera by the Department of Marine Sciences (UGOT) students from Swedish fjords including Gullmar Fjord, Hakefjord, and Idefjord, were imaged on a Nikon SMZ 10 stereo microscope using a DeltaPix DP450 microscope camera (1.92 MP, 1600×1200 resolution) (Fig. 5). Samples were imaged at 30× optical magnification, which gives an optical resolution of 4.16 microns per pixel. A total of 3095 images were produced during a 2-month period. Special care was taken to ensure that all images were imaged at the same exposure time, light angle, and magnification to produce a normalized dataset that could later be augmented, so that all the images to which augmentation was applied, behaved in a predictable manner i.e., for a 20% brightness increase there would be a similar effect on different images to which augmentation was applied. Since the micropaleontological slides included both mounted and non-mounted foraminifera, this ensured that there would be different orientations of the tests, which has been considered beneficial for training a robust machine learning model.

After creating the dataset on which the model was trained on, a second dataset was obtained during the SEEPS II cruise. SEEPS II cruise was done between 17th and 23rd of April 2023, and it's purpose was to detect gas seepage and pockmarks. During the cruise I assisted in sediment core extraction and sediment processing and sampling. After the sediment was processed, some surface samples were imaged and a dataset was created. This dataset has 2011 images, and differs from the training one, as it only has unpicked foraminifera with the surrounding sediment and faecal pelets. This was done in order to assess the models performance on a new dataset, giving independent results.

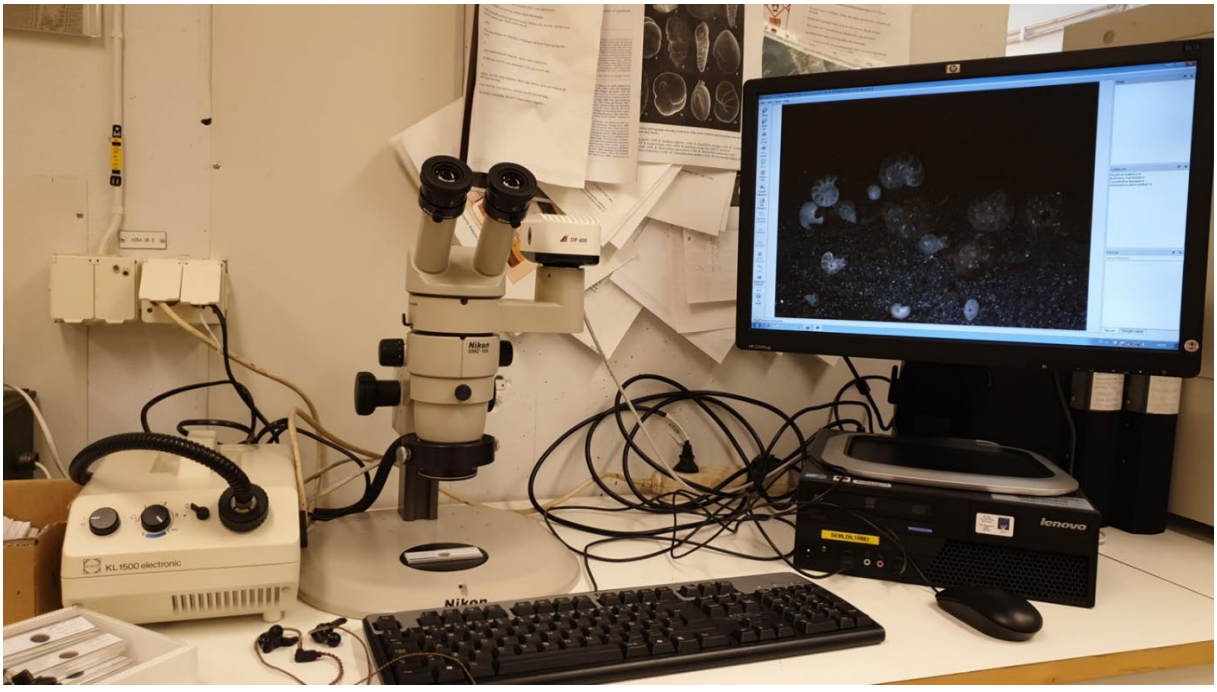


Figure 5. Experimental setup used for image acquisition. Visible in the image is the Nikon SMZ10 stereomicroscope, DeltaPix DP450 microscope camera, and a Lenovo PC used for saving the data.

3.2 Image processing

All images were saved in TIFF format and labeled by using a web-based application programming interface (API) Roboflow (Fig. 6). Roboflow is a machine learning tool that enables the end user to easily label and develop a machine learning dataset. The tool includes several useful features including a content aware selection tool, which speeds up the labelling process by selecting a rough mask of the object selected to label, which significantly reduces the labelling time per individual. A total of 22 138 individual foraminifera were labelled.

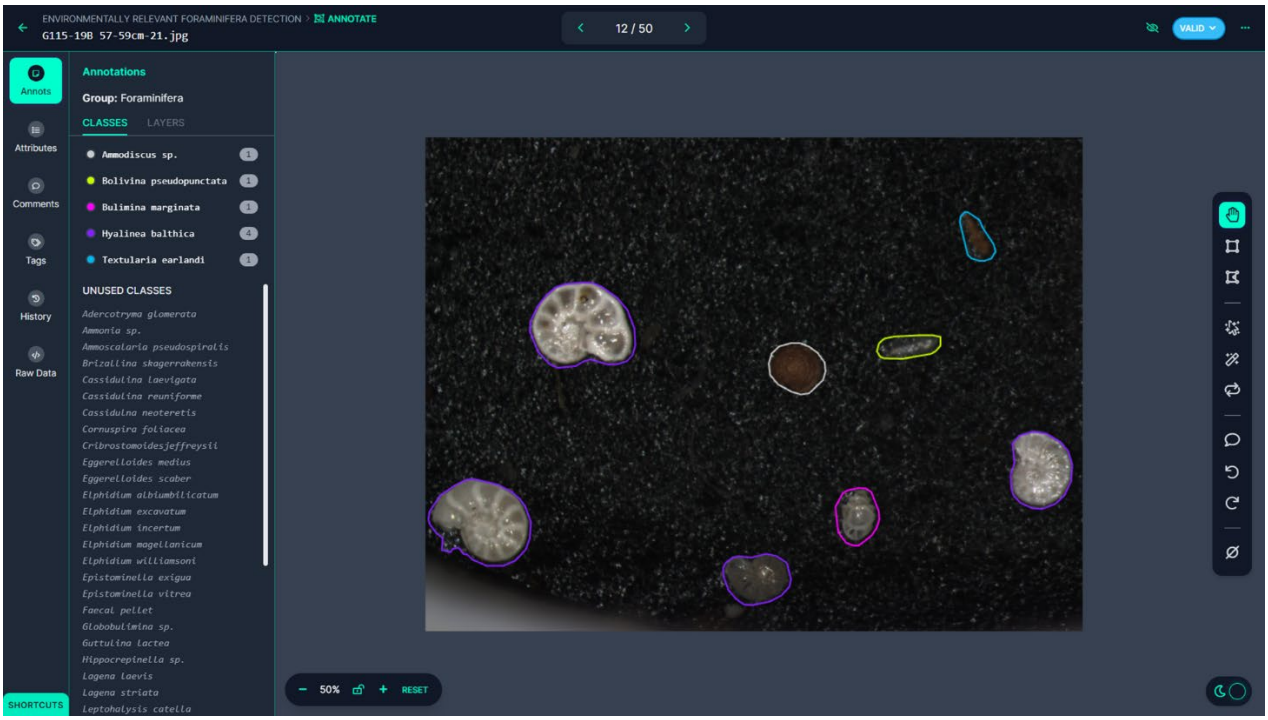


Figure 6. Roboflow interface for classifying individual foraminifera in an image. On the left side of the window, there is a color-coded list of labelled foraminifera in an image on the right. On the rightmost side, there is a tool selection panel, from which a user can label foraminifera either by using an automatic feature extraction tool or by manually drawing a polygon around the foraminifera.

3.3 Dataset creation

After producing a labeled set of 3003 images, a dataset was created. Out of 59 identified species and 22 138 individuals, a total of 29 species were used for creating the dataset (Table 2). In total, 30 species were discarded due to too low individual count, which would reduce the accuracy of the model while training it. The cutoff point was established to be around 70 individuals while training the first models. The dataset was split into training, validation, and test sets in a 70%-20%-10% split. To increase the number of effective images for training, images were augmented in their brightness, contrast, and vibrance, and they were horizontally and vertically flipped. After augmentation, the dataset contained 7089 images and 49 459 individual foraminifera.

Table 2. Number of individual foraminifera in the raw dataset (3003 images, 22 138 individuals). Color mapping is to indicate the viability of each species for being used for training purposes. Here the cutoff point is *Quinqueloculina seminula* at 72 individuals, meaning that species with < 72 labelled individuals were excluded from training dataset. The thick line represents the cutoff point.

Species	Count	Species	Count	Species	Count	Species	Count
Bulimina marginata	**2325**	*Spiroplectammina biformis*	**341**	*Epistominella vitrea*	**68**	*Nonionella iridea*	**9**
Textularia earlandi	**2105**	*Globobulimina sp.*	**335**	*Cornuspira foliacea*	**37**	*Oolina hexagona*	**9**
Stainforthia fusiformis	**1737**	*Brizallina skagerrakensis*	**307**	*Mellonis barleeanum*	**37**	*Lagena striata*	**8**
Hyalinea balthica	**1719**	*Nonionella turgida*	**299**	*Haplophragmoides bradyi*	**31**	*Milliolinella subrotunda*	**8**
Eggerelloides scaber	**1590**	*Uvigerina peregrina*	**215**	*Cribrostomoides sp.*	**26**	*Trifarina angulosa*	**7**
Nonionella sp. T1	**1507**	*Liebusella goesi*	**211**	*Guttulina lactea*	**23**	*Trochammina rotaliformis*	**7**
Elphidium excavatum	**1343**	*Leptohalysis catella*	**203**	*Elphidium albiumbilicatum*	**21**	*Milliammina fusca*	**6**
Ammonia sp.	**1279**	*Quinqueloculina stalkeri*	**187**	*Cassidulina neoteretis*	**15**	*Psammosphaera bowmanni*	**6**
Bolivina pseudopunctata	**1221**	*Ammoscalaria pseudospiralis*	**169**	*Cribrostomoides jeffreysii*	**15**	*Hormosinella gracilis*	**4**
Cassidulina laevigata	**804**	Inner Organic Lining (IOL)	**169**	*Textularia bocki*	**15**	*Lagena laevis*	**4**
Ammodiscus sp.	**774**	*Elphidium magellanicum*	**134**	*Elphidium williamsoni*	**12**	*Recurvoides trochamminiforme*	**4**
Adercotryma glomerata	**725**	*Eggerlloides medius*	**97**	*Elphidium macellum*	**11**	*Lagena mollis*	**2**
Nonionellina labradorica	**664**	*Hippocrepinela acuta*	**87**	*Pullenia osloensis*	**10**	*Epistominella exigua*	**1**
Cibicides lobatulus	**549**	*Pyrgo williamsoni*	**73**	*Elphidium incertum*	**9**		
Reophax sp	**396**	*Quinqueloculina seminula*	**72**	*Glandulina laevigata*	**9**		

3.4 ML model training

After creating an augmented dataset, images were used to train three different YOLOv7 models, listed in order of increasing size, and include YOLOv7, YOLOv7x and YOLOv7E6E. The main difference between those models is the number of layers in a model, and the different input size of images; the first two take 640 x 640 pixel images, while the last one takes 1280 x1280 pixel images. The models were trained on a workstation comprising of an Intel i7 9700K (8 cores, 8 threads@3.60 GHz), 32 GB of RAM and a Nvidia RTX A4000 graphics card running Kubuntu 22.04. PyTorch version 2.0.0 was used for training with CUDA version 11.7. Models were trained on 350 epochs. Training times were 18 hours for the smallest model, 25 hours for the medium sized model and 120 hours for the largest model.

3.5 Introduction to object detection metrics

In object detection, accuracy of the training model is determined by the overall precision (P) and recall (R) of the model. Precision measures the proportion of objects that are correctly classified in a model, while recall denotes the proportion of the objects that are retrieved in a picture. They are expressed in mathematical formulas as such:

$$Precision(P) = \frac{TP}{TP + FP} \quad , \quad Recall(R) = \frac{TP}{TP + FN}$$

with TP standing for True Positive, FP for False Positive and FN for False Negative. In object detection however, there is also another important metric, which measures the overlap of the object location of the ground truth versus the model detection (Davis & Goadrich 2006). Intersection over Union (IoU) measures the overlapping percentage between the ground truth and models detection bounding box (Fig. 7). The IoU measure will be considered a good match if the overlap between the two bounding boxes exceeds a certain threshold. By using precision, recall and IoU, new indicators can be computed such as Average Precision (AP) and mean Average Precision (mAP). Average Precision is computed as the area under the PR curve, where PR curve is a plot of the precision (y-axis) and the recall (x-axis) for different confidence thresholds. F1 score is a measure that combines the precision and recall scores of a model. Based on a harmonic mean of precision and recall, F1 score can be computed by using the following formula:

$$F1 = 2 * \frac{P * R}{P + R}$$

Since the F1 score is an average of precision and recall, it means that F1 score gives equal weight to precision and recall, meaning that a model with a high precision and recall will obtain a high F1 score, a model with low precision and recall will obtain a low F1 score, and lastly a model with a low precision and high recall or vice versa will obtain a medium F1 score. This characteristic makes it suitable for unbalanced datasets with high discrepancy in individual counts between classes.

Finally, mAP is calculated as the average of the AP values over all classes (Padilla *et al.* 2021). Mean average precision can be calculated at different thresholds of IoU. Common mAP thresholds are 0.5 and 0.5 to 0.95. This is depicted as mAP@.5 and mAP@.5:.95 in literature.

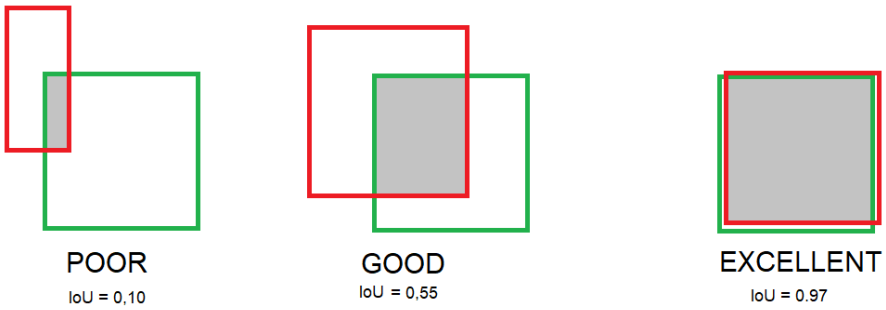


Figure 7. Illustration of the IoU measure. The green bounding box is the ground truth, and the red is the detected bounding box. IoU depends on the threshold defined by the user, so a threshold can be higher or lower than the ones depicted in the figure.

4. Results

After performing the augmentations on the original dataset, the augmented dataset contained 7089 images and 49 459 images. This dataset was then used to train the machine learning models. In total there were 58 attempts while training in the YOLOv7 architecture. Out of those 58 attempts, there were 23 successful attempts and 35 failed ones, which failed due to memory overflow. The default name for each attempt is 'exp' in YOLOv7 architecture, so those were the names used for the models. Models exp 10 and exp 50 used a 16-image batch size, while exp 56 and 58 used 8 image batch size. E6E models exp 51, 52 and 53 used only a 2-image batch size due to memory limitations. The best performing models from the training attempts are presented in Table 3.

Table 3. A performance comparison of trained models during this project. The best performing model is highlighted in green.

Model	P	R	mAP@.5	mAP@.5:.95	model architecture
exp 10	0.868	0.855	0.889	0.696	YOLOv7@640
exp 50	0.877	0.882	0.903	0.699	YOLOv7x@640
exp 51	0.797	0.849	0.867	0.661	YOLOv7E6E@640
exp 52	0.862	0.863	0.895	0.690	YOLOvE6E@1280
exp 53	0.833	0.853	0.870	0.680	YOLOv7E6E@640
exp 56	0.845	0.874	0.887	0.695	YOLOv7x@640
exp 58	0.827	0.883	0.888	0.691	YOLOv7@640

As can be seen in Table 3, the best performing model in this study is exp 50. Each of the models was trained on the same dataset, with the same cutoff point (72 labelled individuals) for the species used as seen in Table 2. The models differed in the size of the input image and learning rate drop off rate, which was faster for the lower performing models compared to exp 50.

4.1 Training curves

There are three different types of loss shown in Figure 8: box loss, objectness loss and classification loss. The box loss represents how well the algorithm can locate the center of an object and how well the predicted bounding box covers an object. Objectness is a measure of the probability that an object exists in a proposed region of interest. If the objectness is high, this means that the image window is likely to contain an object. Classification loss gives an idea of how well the algorithm can predict the correct class of a given object (Alexe *et al.* 2012).

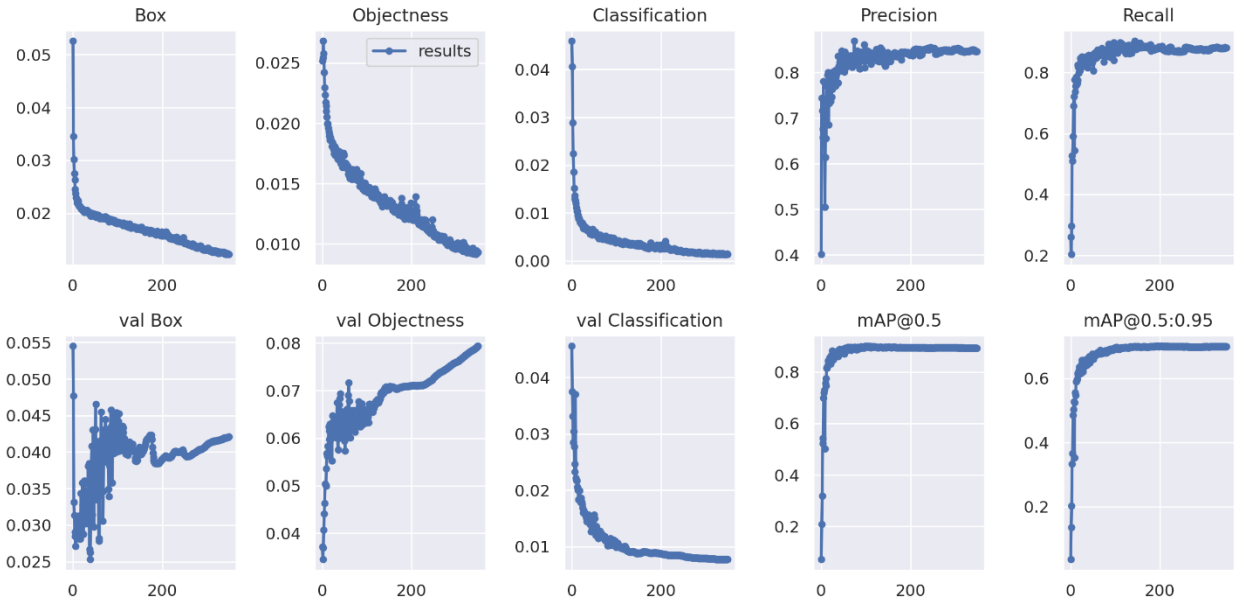


Figure 8. Results showing box loss, objectness loss, classification loss, precision, recall and mean average precision (mAP) over the training epochs for the training and validation set of model exp 50 in this study.

4.2 Confusion matrix

A confusion matrix is a matrix that summarizes the performance of the machine learning model on a set of test data (Fig. 9). It shows how many predictions are correct and incorrect per class. It helps in identifying the classes that are being confused by the model as other classes.

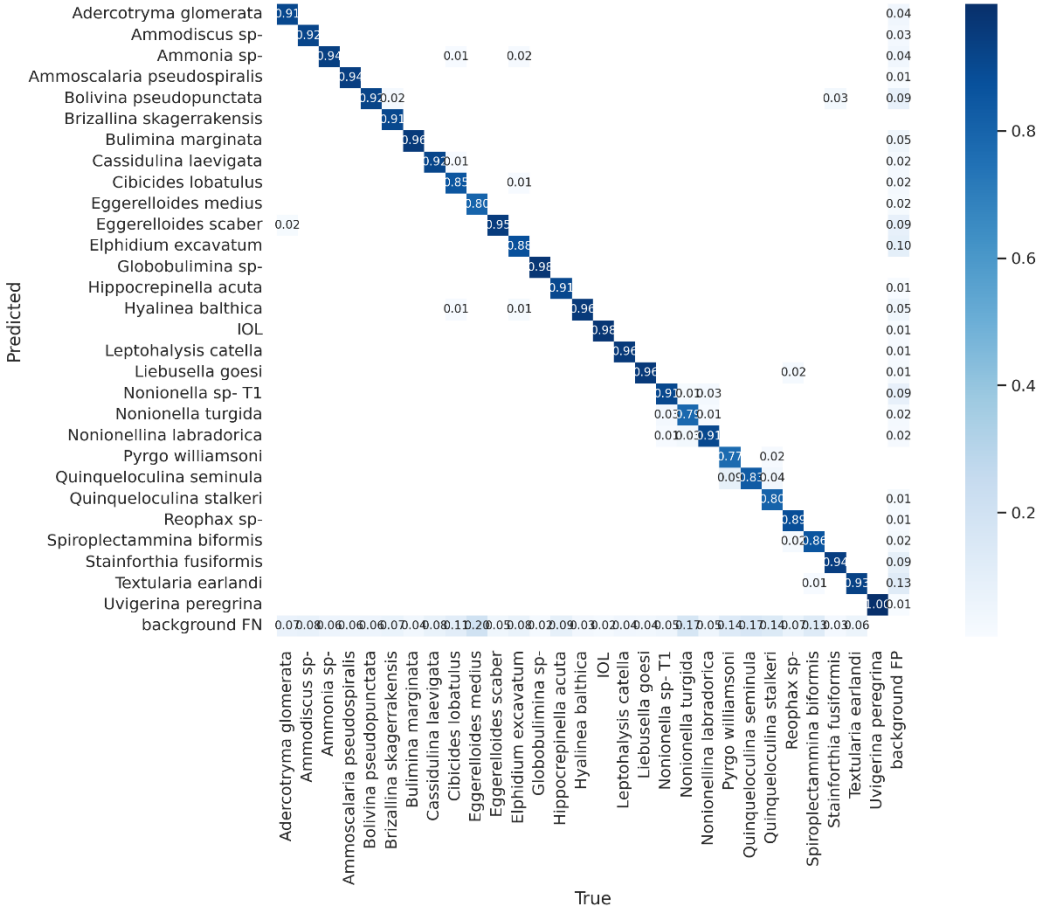


Figure 9. Confusion matrix of the best performing model exp 50. On the left-hand side are the predicted labels made by the model, and on the bottom of the matrix are the ground truth labels. The shade of the blue indicates the probability of the model to correctly identify the given species (only values >0 are shown). It is visible that the model has false negatives of background in all species, while mistaking only a small percentage of one foraminifera species with other foraminifera species.

4.3 F1 curve

F1 scores can be plotted across different confidence thresholds. We can adjust the threshold in order to maximize F1 score. In this case, the confidence threshold for obtaining the maximum F1 score is 0.540.

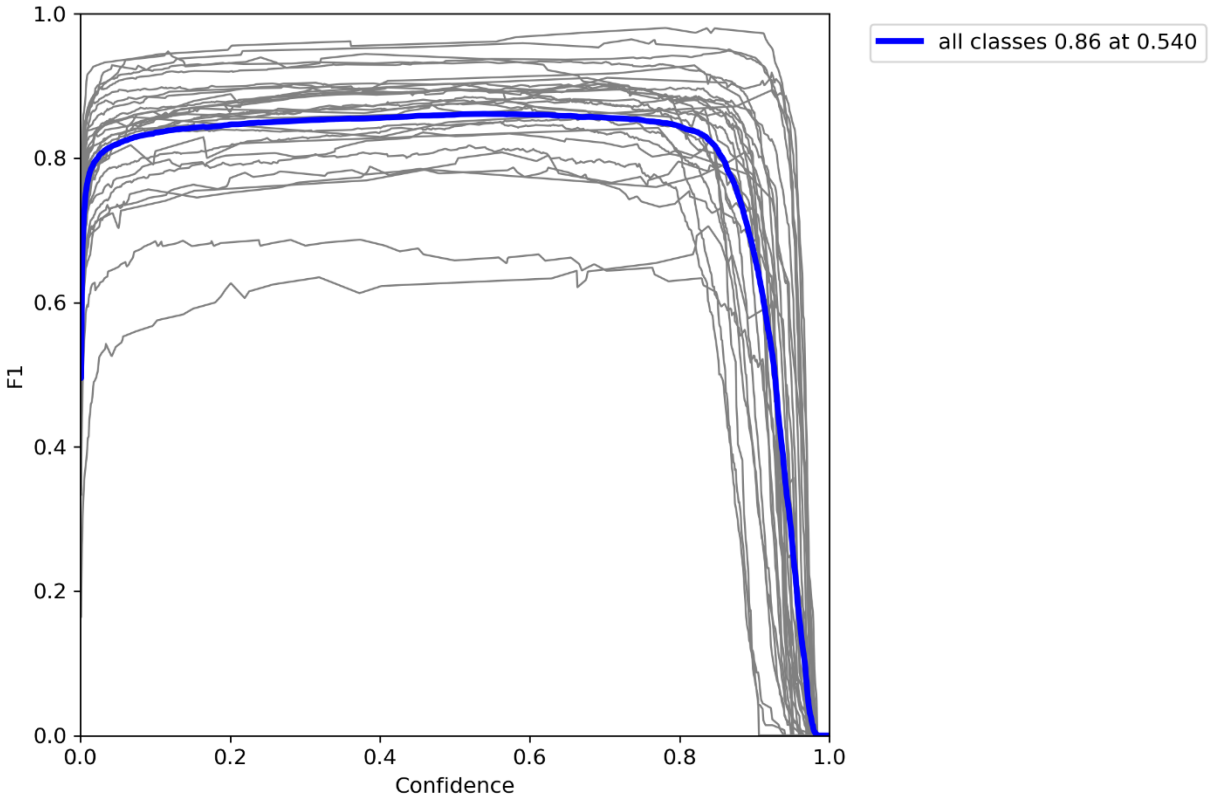


Figure 10. F1 score plot of model exp 50.

4.4 PR curve

Precision-Recall curve is a curve that combines precision and recall in a single visualization. For every threshold, a P and R value is calculated and plotted. It is desired that an algorithm has both a high precision and a high recall value. A high area under the curve represents both high recall and high precision, where high precision relates to a low false positive rate, and high recall relates to a low false negative rate.

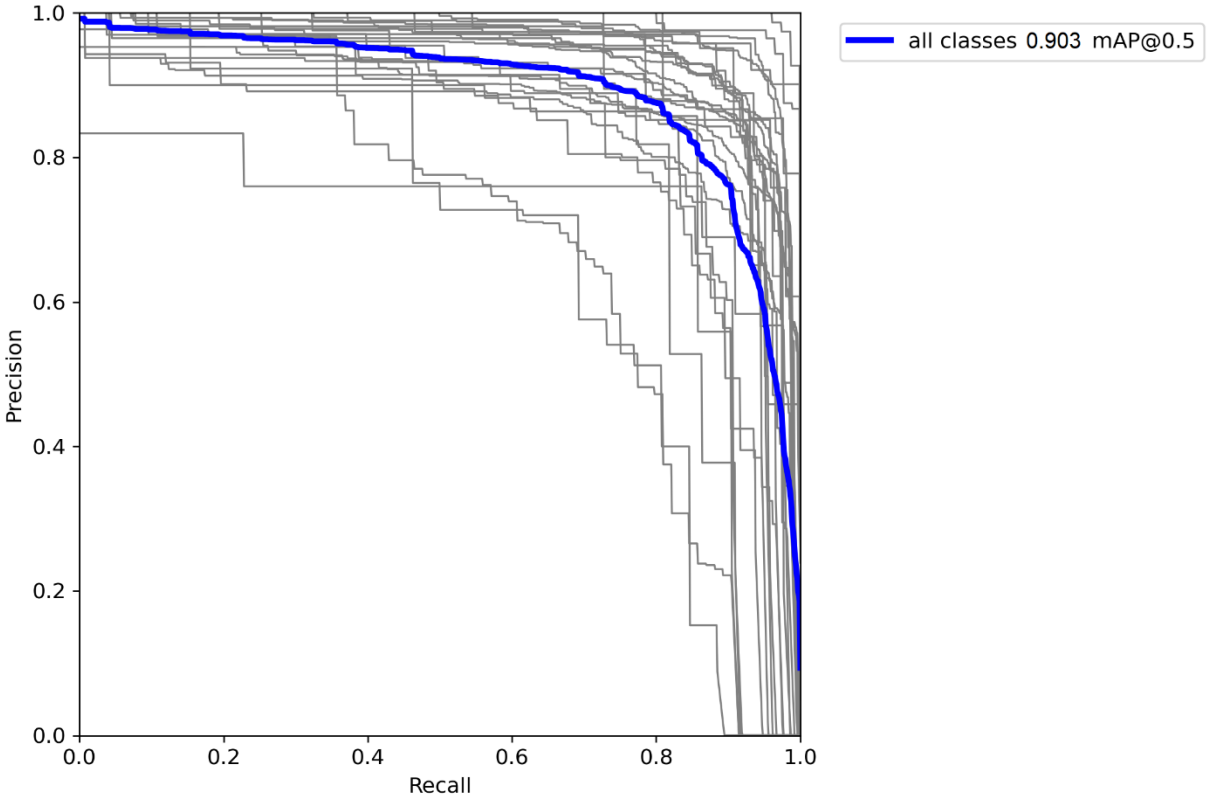


Figure 11. PR plot of model exp 50.

4.5 P and R curves

Precision and recall curves are useful for determining if a model is behaving properly while training it. Typically, as you increase the confidence threshold the precision will go up, and the recall will go down, as is visible in Figure 12 and 13. Since precision and recall are inversely related, other metrics such as F1 score, and PR curve give a more balanced outlook on the model's performance.

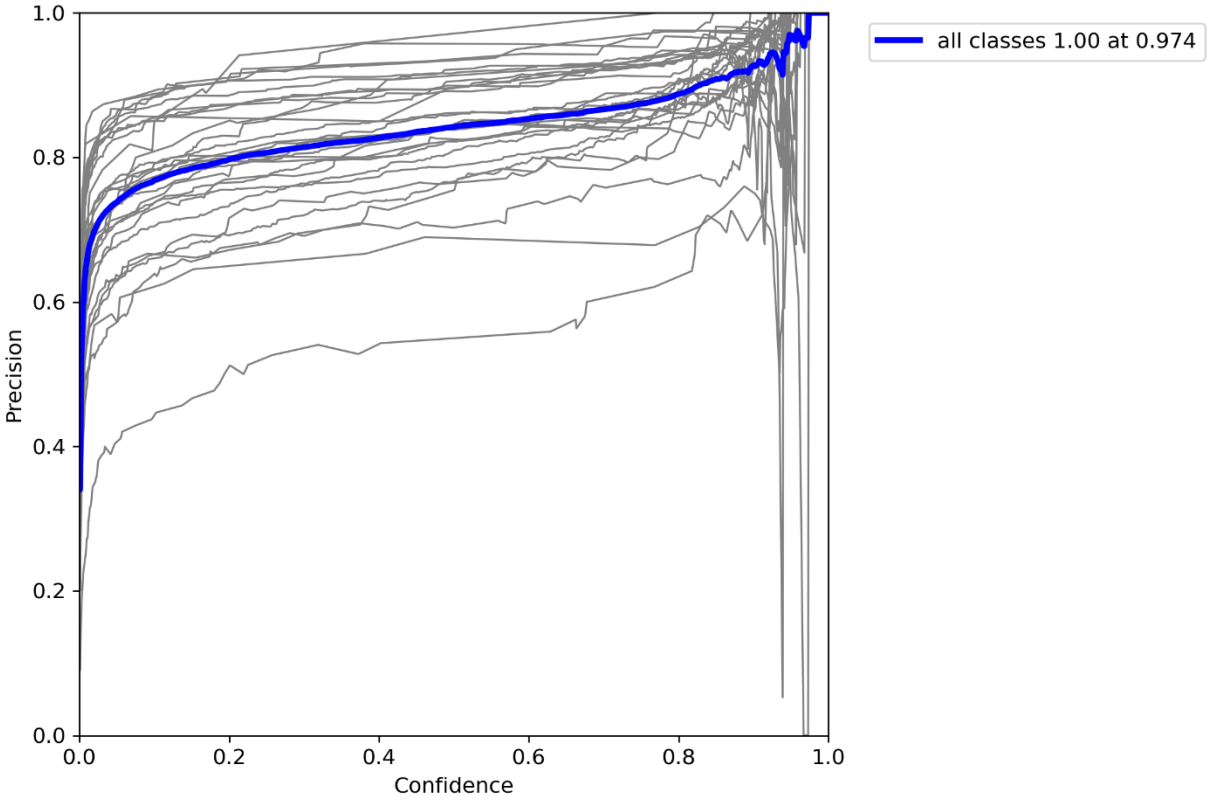


Figure 12. P plot of model exp 50.

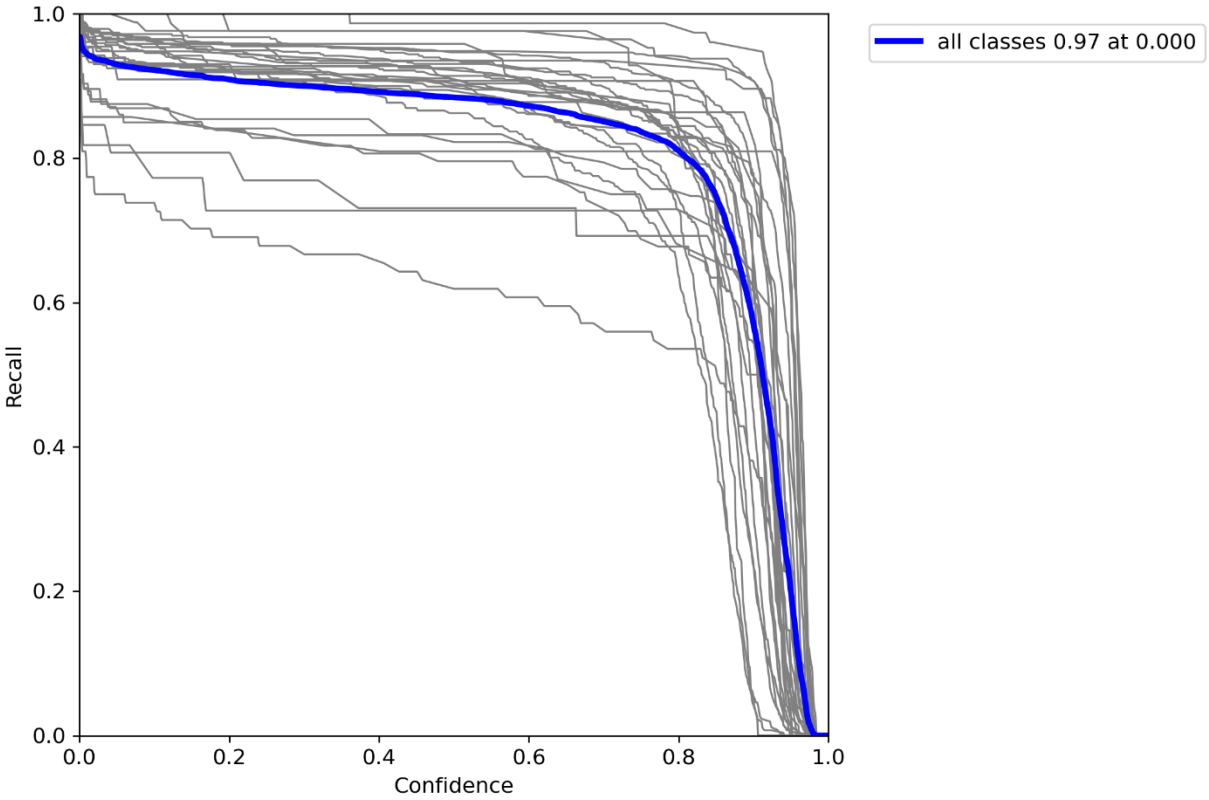


Figure 13. R plot of model exp 50.

4.6 Output image

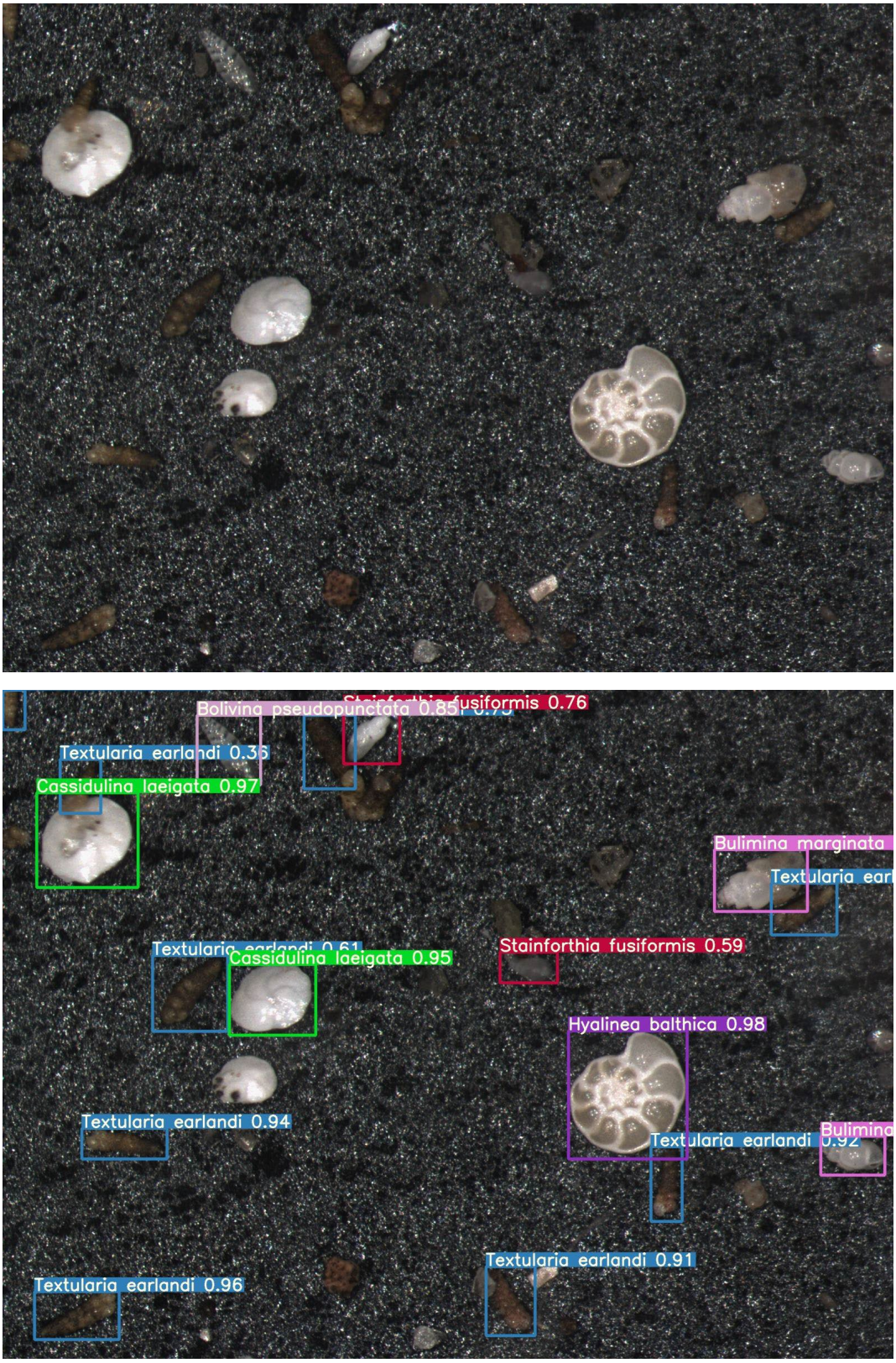


Figure 14. Input image of foraminifera on the top of the figure is processed by a model and creates a new image with bounding boxes and species name. The number next to the species is the confidence level of the model between 0-1.

5. Discussion

5.1 Species performance comparison

Model exp 50 was evaluated on a per species basis and obtained the following results. As can be seen in Table 4, performance of the model varies considerably, based on the identified species. Generally, it appears, that common denominators of poor performance of the model are:

a) agglutinated species, which the model has difficulty of distinguishing against the black background.
b) large and voluminous (e.g., globular to spherically shaped) species, which require multiple focus points in order to distinguish details needed for accurate identification.
c) Species with a low number of individuals in the training set (lower than 100 individuals).

Table 4. Performance of model exp 50 on individual species level.

Species	P	R	mAP@.5	mAP@.5:.95
Adercotryma glomerata	0.862	0.888	0.883	0.574
Ammodiscus sp.	0.893	0.895	0.915	0.713
Ammonia sp.	0.905	0.902	0.953	0.715
Ammoscalaria pseudospiralis	0.957	0.919	0.971	0.843
Bolivina pseudopunctata	0.779	0.855	0.844	0.508
Brizallina skagerrakensis	0.916	0.907	0.944	0.796
Bulimina marginata	0.920	0.947	0.964	0.780
Cassidulina laevigata	0.908	0.880	0.910	0.731
Cibicides lobatulus	0.768	0.796	0.838	0.712
Eggerlloides medius	0.553	0.731	0.691	0.522
Eggerelloides scaber	0.862	0.931	0.946	0.765
Elphidium excavatum	0.818	0.880	0.888	0.619
Globobulimina sp.	0.939	0.976	0.989	0.905
Hippocrepinela acuta	0.687	0.898	0.750	0.579
Hyalinea balthica	0.925	0.949	0.975	0.867
IOL	0.843	0.976	0.947	0.550
Leptohalysis catella	0.850	0.948	0.878	0.607
Liebusella goesi	0.828	0.935	0.955	0.846
Nonionella sp. T1	0.852	0.918	0.937	0.703
Nonionella turgida	0.709	0.619	0.699	0.519
Nonionellina labradorica	0.899	0.887	0.925	0.797
Pyrgo williamsoni	0.901	0.810	0.887	0.800
Quinqueloculina seminula	0.976	0.727	0.856	0.755
Quinqueloculina stalkeri	0.778	0.833	0.824	0.611
Reophax sp	0.827	0.904	0.923	0.709
Spiroplectammina biformis	0.842	0.822	0.825	0.631
Stainforthia fusiformis	0.826	0.933	0.916	0.610
Textularia earlandi	0.794	0.914	0.881	0.635
Uvigerina peregrina	0.937	0.987	0.996	0.883

Since some of the foraminifera used in this dataset were only imaged on mounted micropaleontological slides, this can also be a contributing factor to a poor model performance,

since it allows the model to see only one orientation of the foraminifera. Background choice is also an important factor to be discussed, since as is seen in Figure 9, ML model confused the background as a FP or FN for every species between 1 to 13% at a time. Since the low contrast of the background could explain the FP and FN performance, a different color of the background could be beneficial for reducing the error rate. Furthermore, in Figure 9, it is visible that the model confuses foraminifera in the same genus, which is comparable to human taxonomists.

5.2 Comparison with other foraminifera ML models

To the best of my knowledge, there are no YOLO based foraminifera ML models reported and published to date. Therefore, I shall compare the model performance to different ML model architecture. Within those models, only a part deals with benthic foraminifera, and for that reason I will include the planktonic foraminifera ones as well for completion. Firstly, I will give a brief overview of foraminifera ML models used to date, and afterwards I will compare the performance of those models to the model used in this thesis.

Hsiang *et al*. (2019) used a VGG16 ML model trained on a dataset of 35 species of planktonic foraminifera obtaining an 87.4% precision. Their project also resulted in a published open access dataset *Endless Forams*, which contains 34 640 individual planktonic foraminifera. Further, Marchant *et al*. (2020) used their custom modified version of ResNet50 ML model, which was pretrained on ImageNet database. They used a custom Base-Cyclic CNN, which adapts to the input image size, and reduces the time needed to learn image features, since it learns the features from multiple orientations at the same time. On *Endless Forams* database, their network obtained a 90,3% precision, while in their custom dataset, which included both planktonic and benthic foraminifera, it resulted in an 89.8% precision. In another study by Mitra *et al*. (2019), the authors used a combination of ResNet50 and VGG16 ML models on a custom planktonic foraminifera dataset and obtained 80% precision. Their lower precision results could be attributed to a small dataset of foraminifera (1437 individuals). Next, Johansen & Sørensen (2020) use a VGG16 ML model, that detects the particles on a micropaleontological tray, and classifies it either as a sediment particle, benthic or planktonic foraminifera. For such a simple task VGG16 proved to be highly effective, and the model obtained 98.5% precision. Karaderi *et al*. (2022) use the *Endless Forams* database and implement a custom ResNet50 ML model obtaining 92% accuracy. Nanni *et al.* (2023) use the same dataset as Mitra *et al.* (2019), but they use a different pipeline for preprocessing the images. They use RGB preprocessing on original greyscale images to improve ResNet50 performance, and obtain an 89% precision rate. When ML models are compared to taxonomic experts, they outperform taxonomic experts by a 10-15% margin.

Now comparing my YOLOv7x exp 50 model to the aforementioned models, it can be seen that it performs comparably well, but the difference is that the YOLO family of models are object detection models, which can be used for more applications than just classification-based models.

For example, object detection models can be used in imaging unpicked foraminifera samples and still obtain a species identification. This allows for shorter processing times of sediment samples, as compared to manual picking of foraminifera. The performance of model exp 50 can be seen in the appendix, where one can see the sediment ignorance of the model. The SEEPS 2 dataset unfortunately hasn't yet been fully labeled so accurate metrics are not available, but preliminary results show higher than 85% precision on the new images. Another application could be real-time foraminifera detection from a webcam or mobile phone.

The use case of such applications could be teaching future taxonomists or used by non-experts, who-do not have the means to learn foraminifera identification from a trained expert, or to be used as a teaching aid in a university environment. During the course of this thesis, one such application was developed for Android based devices, but it is not yet available for use, since the YOLO models are non-optimized for ARM processors. Work on this application is going and it is anticipated that it will be published later this year.

5.3 Implications for future studies

One limitation in the method described in this project is that each individual foraminifera specimen is only represented by one hyperfocal image. This is a problem since some foraminifera have distinguishing features at different focal planes, so for further research, one should strive to take images at different focal planes and Z-stack them for better performance. Another drawback is that the background used for imaging the foraminifera is low contrast compared to the foraminifera, so one could change the background into some other high contrast color. A personal suggestion would be chroma key green (i.e., the shade used for green screens), since it has a proven track record for distinguishing foreground from background in the movie industry for decades.

Better performance could also result from higher quality microscopes and cameras, since the optical resolution and megapixel count of the camera could have affected the model training performance for small individuals, which on 1.92 MP can be only a few dozen pixels in size. A higher megapixel count of the camera would allow for better resolving power of certain identifying features in foraminifera, which would allow for them to be identified to species level. For some genus such as *Ammonia* this is crucial, since identifying features are really small and hard to spot even for human taxonomic experts (Pavard *et al.* 2021).

An important observation was made while preparing the training dataset: one should always try to use the same imaging settings and setup while imaging, to reduce the number of variables that can affect the model performance. One should use augmentation to artificially add different lighting conditions to the dataset. This is best used when preparing a relatively large dataset in a short amount of time. Another time-saving method would be to employ an automatic microscopic table and autofocus system which would dramatically reduce the time necessary to image the microscopic samples.

In order to fully test the performance of the newly trained model, a good practice would be to test the model on a different imaging setup to see how the model performs on different hardware.

6. Conclusion

This thesis addresses the two main tasks of foraminifera detection and classification using deep learning.

For the detection, besides showing that deep learning-based approaches can be used successfully to detect benthic foraminifera on light microscope images, the chosen model has demonstrated extremely fast detection times on the scale of milliseconds. This study, compared to other works, has included a relatively high count of benthic foraminifera species used in machine learning dataset, and obtained favorable results in picked and unpicked sediment samples.

Lastly, this study does not aim to eliminate human expertise from the taxonomic identification process. Instead, an ML based approach could be used as a labor-saving device to go through the bulk of the dataset, and later a person could validate and if needed correct the identifications. By reducing the time needed to identify foraminifera, one could focus more on analysis of the ecological interactions between the species.

Acknowledgements

I would like to express my gratitude to everyone who has contributed to the completion of this thesis.

First, I would like to thank all my supervisors. Prof. Irina Polovodova Asteman, thank you for your guidance and expertise in foraminifera identification and for the patience for going through all the images I collected and checking the labels of foraminifera. Dr. Allison Hsiang, thank you for the guidance and advice for data acquisition and creating a good dataset for machine learning. Dr. Mats Josefson and AstraZeneca thank you for all the helpful comments of my code and for the hardware I used for imaging the foraminifera and for the computer for training the machine learning model.

I thank my dear friend Ana Crnogorac for designing and drawing the cover image for this thesis.

I thank all my friends and my family for the support given while writing this thesis.

Finally thank you to all the students who have picked the foraminifera from the cores, for if there weren't you this thesis would not be possible to produce.

References

Alexe B, Deselaers T, Ferrari V. (2012). Measuring the Objectness of Image Windows. IEEE transactions on pattern analysis and machine intelligence, doi 10.1109/TPAMI.2012.28.

Alve E. (1995). Benthic foraminiferal responses to estuarine pollution. Journal of Foraminiferal Research - J FORAMIN RES 25: 190–203.

Alve E. (1991). Benthic foraminifera in sediment cores reflecting heavy metal pollution in Sorfjord, western Norway. The Journal of Foraminiferal Research 21: 1–19.

Alve E. (2003). A common opportunistic foraminiferal species as an indicator of rapidly changing conditions in a range of environments. Estuarine, Coastal and Shelf Science 57: 501–514.

Alve E. (2010). Benthic foraminiferal responses to absence of fresh phytodetritus: A two-year experiment. Marine Micropaleontology 76: 67–75.

Alve E, Lepland A, Magnusson J, Backer-Owe K. (2009). Monitoring strategies for re-establishment of ecological reference conditions: Possibilities and limitations. Marine pollution bulletin 59: 297–310.

Bernhard JM, Alve E. (1996). Survival, ATP pool, and ultrastructural characterization of benthic foraminifera from Drammensfjord (Norway): response to anoxia. Marine Micropaleontology 28: 5–17.

Boersma A. (1998). 2 - Foraminifera. In: Haq BU, Boersma A (ed.). Introduction to Marine Micropaleontology (Second Edition), pp. 19–77. Elsevier Science B.V., Amsterdam.

Borja A, Miles A, Occhipinti-Ambrogi A, Berg T. (2009). Current status of macroinvertebrate methods used for assessing the quality of European marine waters: Implementing the Water Framework Directive. Hydrobiologia 633: 181–196.

Bre F, Gimenez J, Fachinotti V. (2017). Prediction of wind pressure coefficients on building surfaces using Artificial Neural Networks. Energy and Buildings, doi 10.1016/j.enbuild.2017.11.045.

Castro W, Oblitas J, Santa-Cruz R, Avila-George H. (2017). Multilayer perceptron architecture optimization using parallel computing techniques. PLoS ONE, doi 10.1371/journal.pone.0189369.

Choquel C, Geslin E, Metzger E, Filipsson H, Risgaard-Petersen N, Launeau P, Giraud M, Jauffrais T, Jesus B, Mouret A. (2021). Denitrification by benthic foraminifera and their contribution to N-loss from a fjord environment. Biogeosciences 18: 327–341.

Davis J, Goadrich M. (2006). The Relationship Between Precision-Recall and ROC Curves. Proceedings of the 23rd International Conference on Machine Learning, ACM, doi 10.1145/1143844.1143874.

Deldicq N, Alve E, Schweizer M, Asteman IP, Hess S, Darling K, Bouchet VMP. (2019). History of the introduction of a species resembling the benthic foraminifera *Nonionella stella* in the Oslofjord (Norway): Morphological, molecular and paleo-ecological evidences. Aquatic Invasions 14: 182–205.

Directive 2000/60/EC of the European Parliament and of the Council of 23 October 2000 establishing a framework for Community action in the field of water policy, Official Journal L 327 , 22/12/2000 P. 0001 – 0073

Directive 2008/56/EC of the European Parliament and of the Council of 17 June 2008 establishing a framework for community action in the field of marine environmental policy (Marine Strategy Framework Directive) (Text with EEA relevance), OJ L 164, 25.6.2008, p. 19–40

Duffield CJ, Edvardsen B, Eikrem W, Alve E. (2014). Effects of different potential food sources on upper-bathyal benthic foraminifera: An experiment with propagules. The Journal of Foraminiferal Research. 44. 427-444. 10.2113/gsjfr.44.4.416.

Dumoulin V, Visin F. (2016). A guide to convolution arithmetic for deep learning. doi 10.48550/ARXIV.1603.07285.

Fei-Fei L, Fergus R, Perona P. (2007). Learning generative visual models from few training examples: An incremental Bayesian approach tested on 101 object categories. Computer Vision and Image Understanding 106: 59–70.

Filipsson HL, Nordberg K. (2004). Climate variations, an overlooked factor influencing the recent marine environment. An example from Gullmar Fjord, Sweden, illustrated by benthic foraminifera and hydrographic data. Estuaries 27: 867–881.

Hari S, Littmann S, Glock N, von Arx J, Coenen T, Roy A-S. (2020). Correlative cathodoluminescence and EDS imaging of the benthic agglutinated foraminifer *Liebusella goesi*. EGU General Assembly Conference Abstracts, p. 20478.

He K, Zhang X, Ren S, Sun J. (2015). Deep Residual Learning for Image Recognition. doi 10.48550/ARXIV.1512.03385.

Hsiang AY, Brombacher A, Rillo MC, Mleneck-Vautravers MJ, Conn S, Lordsmith S, Jentzen A, Henehan MJ, Metcalfe B, Fenton IS, Wade BS, Fox L, Meilland J, Davis C V., Baranowski U, Groeneveld J, Edgar KM, Movellan A, Aze T, Dowsett HJ, Miller CG, Rios N, Hull PM. (2019). Endless Forams: >34,000 Modern Planktonic Foraminiferal Images for Taxonomic Training and Automated Species Recognition Using Convolutional Neural Networks. Paleoceanography and Paleoclimatology 34: 1157–1177.

Johansen TH, Sørensen SA. (2020). Towards detection and classification of microscopic foraminifera using transfer learning. doi 10.48550/ARXIV.2001.04782.

Karaderi T, Burghardt T, Hsiang AY, Ramaer J, Schmidt DN. (2022). Visual microfossil identification via deep metric learning. Pattern Recognition and Artificial Intelligence: Third International Conference, ICPRAI 2022, Paris, France, June 1–3, 2022, Proceedings, Part I, pp. 34–46. Springer,

Korsun S, Hald M. (1998). Modern Benthic Foraminifera off Novaya Zemlya Tidewater Glaciers, Russian Arctic. Arctic and Alpine Research 30: 61.

Krizhevsky A, Sutskever I, Hinton G. E. (2012). ImageNet Classification with Deep Convolutional Neural Networks. Advances in Neural Information Processing Systems 25. Curran Associates, Inc.

Lecun Y, Kavukcuoglu K, Farabet C. (2010). Convolutional networks and applications in vision. Proceedings of 2010 IEEE International Symposium on Circuits and Systems, doi 10.1109/iscas.2010.5537907.

Lin T-Y, Maire M, Belongie S, Bourdev L, Girshick R, Hays J, Perona P, Ramanan D, Zitnick CL, Dollár P. (2014). Microsoft COCO: Common Objects in Context. doi 10.48550/ARXIV.1405.0312.

Lintner M, Lintner B, Wanek W, Keul N, von der Kammer F, Hofmann T, Heinz P. (2021). Effects of heavy elements (Pb, Cu, Zn) on algal food uptake by *Elphidium excavatum* (Foraminifera). Heliyon 7: e08427.

Luze G.F, Mackensen A, Wefer G. (1983) Foraminiferen der Kieler Bucht: 2. Salinitätsansprüche von *Eggerella scabra* (Williamson). Meyniana 35, 55-65.

Marchant R, Tetard M, Pratiwi A, Adebayo M, de Garidel-Thoron T. (2020). Automated analysis of foraminifera fossil records by image classification using a convolutional neural network. Journal of Micropalaeontology 39: 183–202.

Mitra R, Marchitto TM, Ge Q, Zhong B, Kanakiya B, Cook MS, Fehrenbacher JS, Ortiz JD, Tripati A, Lobaton E. (2019). Automated species-level identification of planktic foraminifera using convolutional neural networks, with comparison to human performance. Marine Micropaleontology 147: 16–24.

Murray J. (2003). An illustrated guide to the benthic foraminifera of the Hebridean Shelf, west of Scotland, with notes on their mode of life. Palaeontologia Electronica 5: 31.

Murray J. (2006). Ecology and Applications of Benthic Foraminifera. Palaeogeography, Palaeoclimatology, Palaeoecology 95: 1–426.

Nanni L, Faldani G, Brahnam S, Bravin R, Feltrin E. (2023). Improving foraminifera classification using Convolutional Neural Networks with Ensemble Learning. doi 10.20944/preprints202302.0396.v1.

Nordberg K, Gustafsson M, Krantz A-L. (2000). Decreasing oxygen concentrations in the Gullmar Fjord, Sweden, as confirmed by benthic foraminifera, and the possible association with NAO. Journal of Marine Systems 23: 303–316.

O'Brien PAJ, Polovodova Asteman I, Bouchet VMP. (2021). Benthic Foraminiferal Indices and Environmental Quality Assessment of Transitional Waters: A Review of Current Challenges and Future Research Perspectives. Water, doi 10.3390/w13141898.

Padilla R, Passos WL, Dias TLB, Netto SL, Da Silva EAB. (2021). A comparative analysis of object detection metrics with a companion open-source toolkit. Electronics (Switzerland) 10: 1–28.

Pavard J-C, Richirt J, Courcot L, Bouchet P, Seuront L, Bouchet VMP (2021). Fast and Reliable Identification of *Ammonia* Phylotypes T1, T2 and T6 Using a Stereomicroscope: Implication for Large-Scale Ecological Surveys and Monitoring Programs. Water, 13(24):3563. doi/10.3390/w13243563

Pawlowski J, Bowser SS, Gooday AJ. (2007). A note on the genetic similarity between shallow- and deep-water *Epistominella vitrea* (Foraminifera) in the Antarctic. Deep Sea Research Part II: Topical Studies in Oceanography 54: 1720–1726.

Polovodova Asteman I, Filipsson H, Nordberg K. (2018). Tracing winter temperatures over the last two millennia using a NE Atlantic coastal record. Climate of the Past 14: 1097–1118.

Polovodova Asteman I, Hanslik D, Nordberg K. (2015). An almost completed pollution-recovery cycle reflected by sediment geochemistry and benthic foraminiferal assemblages in a Swedish–Norwegian Skagerrak fjord. Marine pollution bulletin, doi 10.1016/j.marpolbul.2015.04.031.

Polovodova Asteman I, Nordberg K. (2013). Foraminiferal fauna from a deep basin in Gullmar Fjord: The influence of seasonal hypoxia and North Atlantic Oscillation. Journal of Sea Research 79: 40–49.

Polovodova Asteman I, Nordberg K, Filipsson H. (2011). The benthic foraminiferal record of the Medieval Warm Period and the recent warming in the Gullmar Fjord, Swedish west coast. Fuel and Energy Abstracts 81: 95–106.

Polovodova Asteman I, Schönfeld J. (2015). Recent invasion of the foraminifer *Nonionella stella* Cushman & Moyer, 1930 in northern European waters: Evidence from the Skagerrak and its fjords. Journal of Micropalaeontology, doi 10.1144/jmpaleo2015-007.

Redmon J, Divvala S, Girshick R, Farhadi A. (2015). You Only Look Once: Unified, Real-Time Object Detection. doi 10.48550/ARXIV.1506.02640.

Risgaard-Petersen N, Langezaal A, Høgslund S, Schmid M, Jetten M, Op den Camp H, Derksen J, Piña-Ochoa E, Eriksson S, Nielsen LP, Revsbech N, Cedhagen T, Zwaan G. (2006). Evidence for complete denitrification in a benthic foraminifer. Nature 443: 93–96.

Ross CR. (1984). *Hyalinea balthica* and its late Quaternary paleoclimatic implications; Strait of Sicily. Journal of Foraminiferal Research 14: 134–139.

Schönfeld J, Alve E, Geslin E, Jorissen F, Korsun S, Spezzaferri S, Abramovich S, Almogi-Labin A, du Chatelet EA, Barras C, Bergamin L, Bicchi E, Bouchet V, Cearreta A, Di Bella L, Dijkstra N, Disaro ST, Ferraro L, Frontalini F, Gennari G, Golikova E, Haynert K, Hess S, Husum K, Martins V, McGann M, Oron S, Romano E, Sousa SM, Tsujimoto A. (2012). The FOBIMO (FOraminiferal BIo-MOnitoring) initiative-Towards a standardised protocol for soft-bottom benthic foraminiferal monitoring studies. Marine Micropaleontology 94–95: 1–13.

Sun X, Nasrabadi NM, Tran TD. (2018). Supervised Deep Sparse Coding Networks. 2018 25th IEEE International Conference on Image Processing (ICIP), doi 10.1109/icip.2018.8451701.

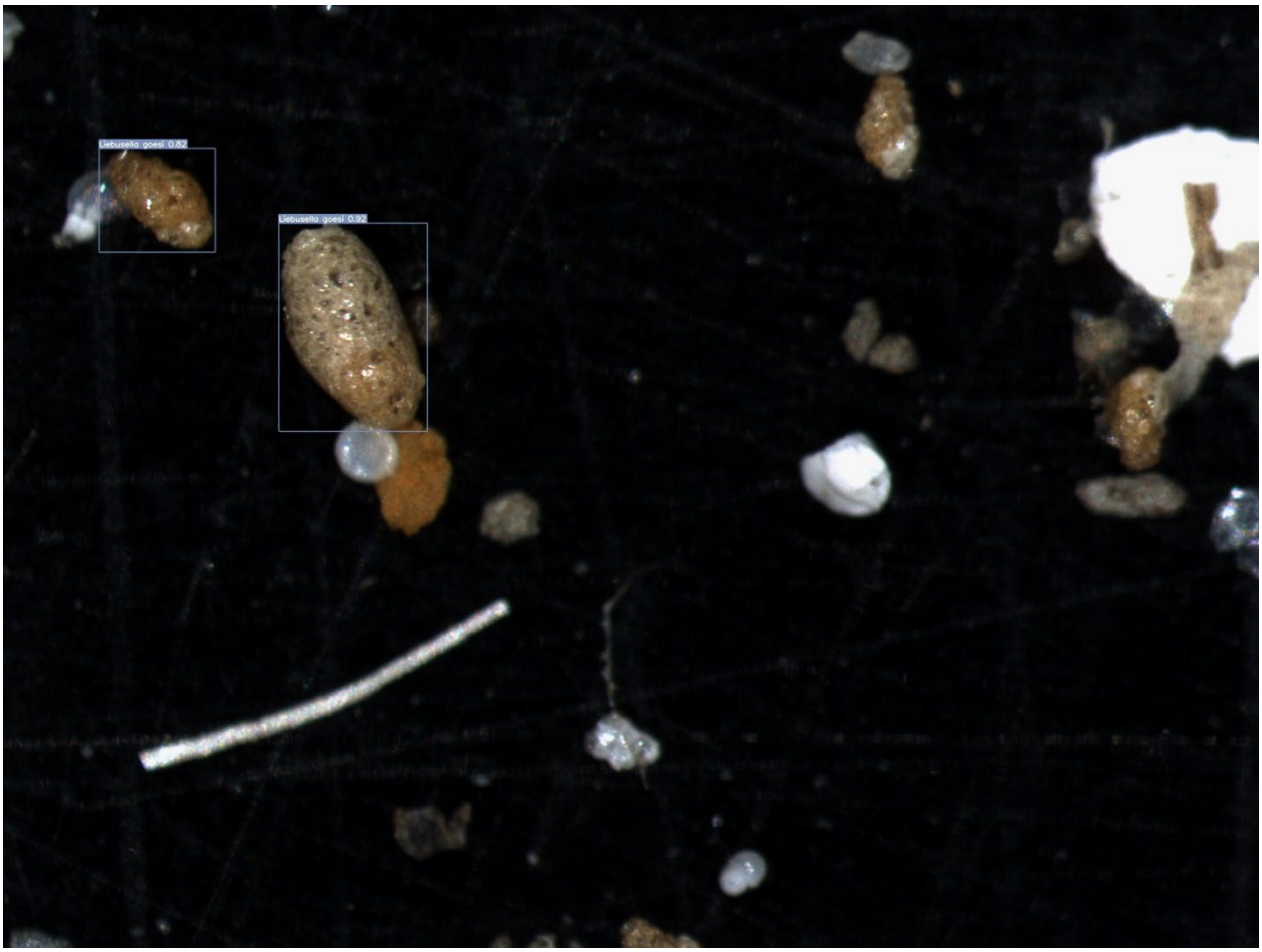
Wang C-Y, Bochkovskiy A, Liao H-YM. (2022). YOLOv7: Trainable bag-of-freebies sets new state-of-the-art for real-time object detectors. doi 10.48550/ARXIV.2207.02696.

Yanko V, Arnold A, Parker W, Gupta B. (2003). Effects of marine pollution on benthic Foraminifera. pp. 217–235.

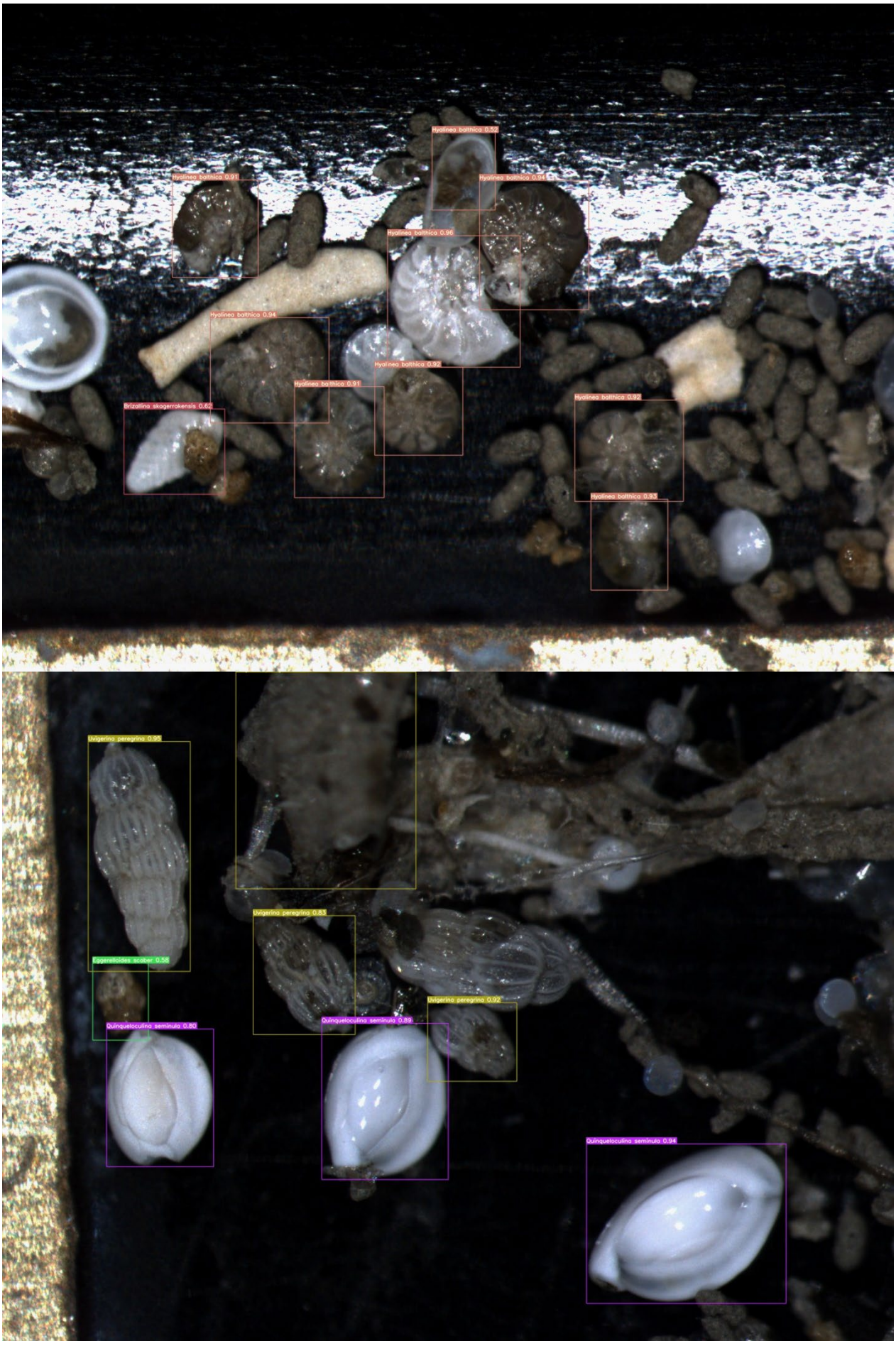
Appendix

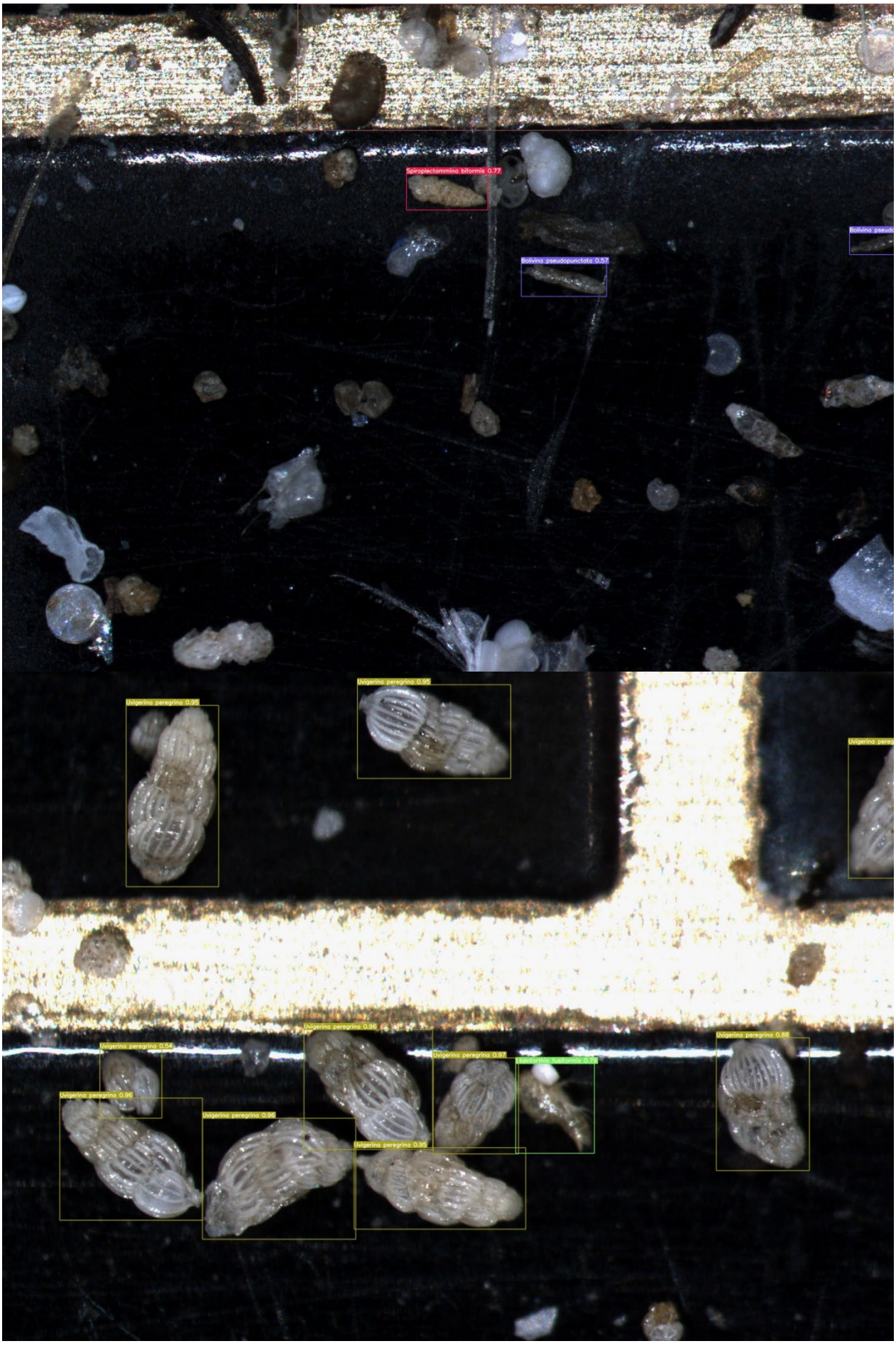
Additional images of model detections

Listed below are examples of images taken during the SEEPS II cruise. They are here to illustrate the performance of the model on unpicked samples and to showcase excellent sediment and faecal pellets ignorance of the model.




Liebusella goesi 0.96
Uvigerina peregrina 0.68
Liebusella goesi 0.96
Globobulimina sp- 0.90
Liebusella goesi 0.97
Liebusella goesi 0.97
Globobulimina sp- 0.87
Eggerellodes scaber 0.55
Liebusella goesi 0.96
Globobulimina sp- 0.92


Hyalinea_balthica 0.52
Hyalinea_balthica 0.91
Hyalinea_balthica 0.94
Hyalinea_balthica 0.96
Hyalinea_balthica 0.94
Hyalinea_balthica 0.92
Hyalinea_balthica 0.91
Brizalina_skagerrakensis 0.62
Hyalinea_balthica 0.92
Hyalinea_balthica 0.93
Uvigerina_peregrina 0.95
Uvigerina_peregrina 0.83
Eggerelloides_scaber 0.58
Uvigerina_peregrina 0.92
Quinqueloculina_seminula 0.80
Quinqueloculina_seminula 0.89
Quinqueloculina_seminula 0.94


Spiroplectammina biformis 0.77
Bolivina pseudopunctata 0.57
Uvigerina peregrina 0.95
Uvigerina peregrina 0.95
Uvigerina peregrina 0.96
Uvigerina peregrina 0.54
Uvigerina peregrina 0.97
Uvigerina peregrina 0.96
Uvigerina peregrina 0.96
Uvigerina peregrina 0.95
Uvigerina peregrina 0.88

exp 10

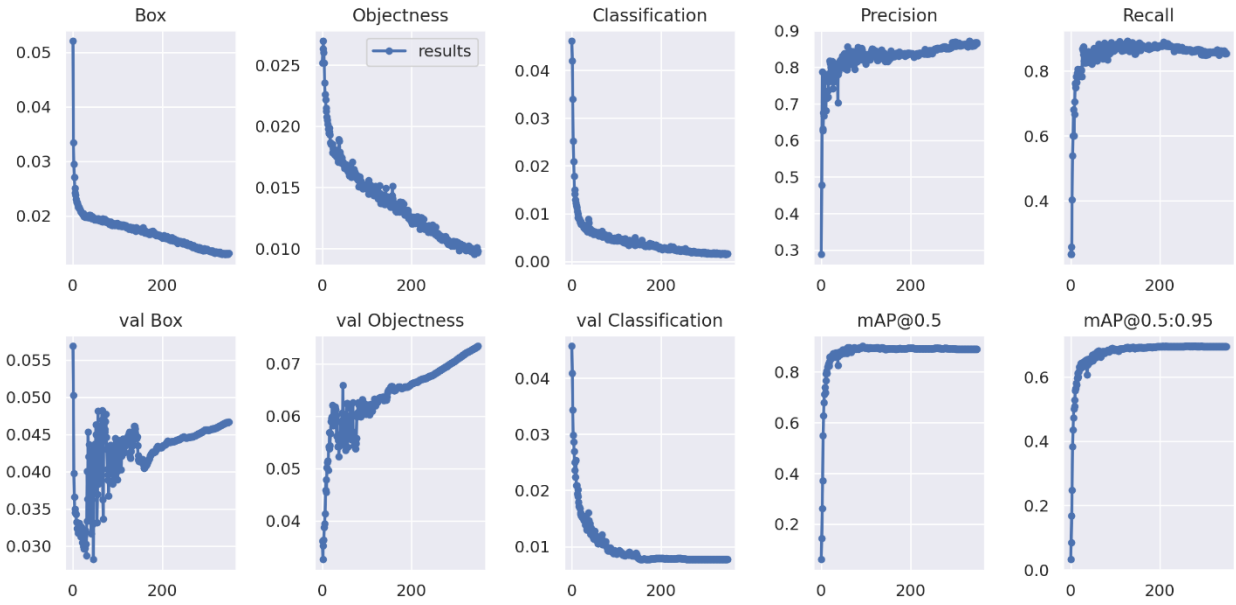


Figure A1. Results of box loss, objectness loss, classification loss, precision, recall and mean average precision (mAP) over the training epochs for the training and validation set of model exp 10.

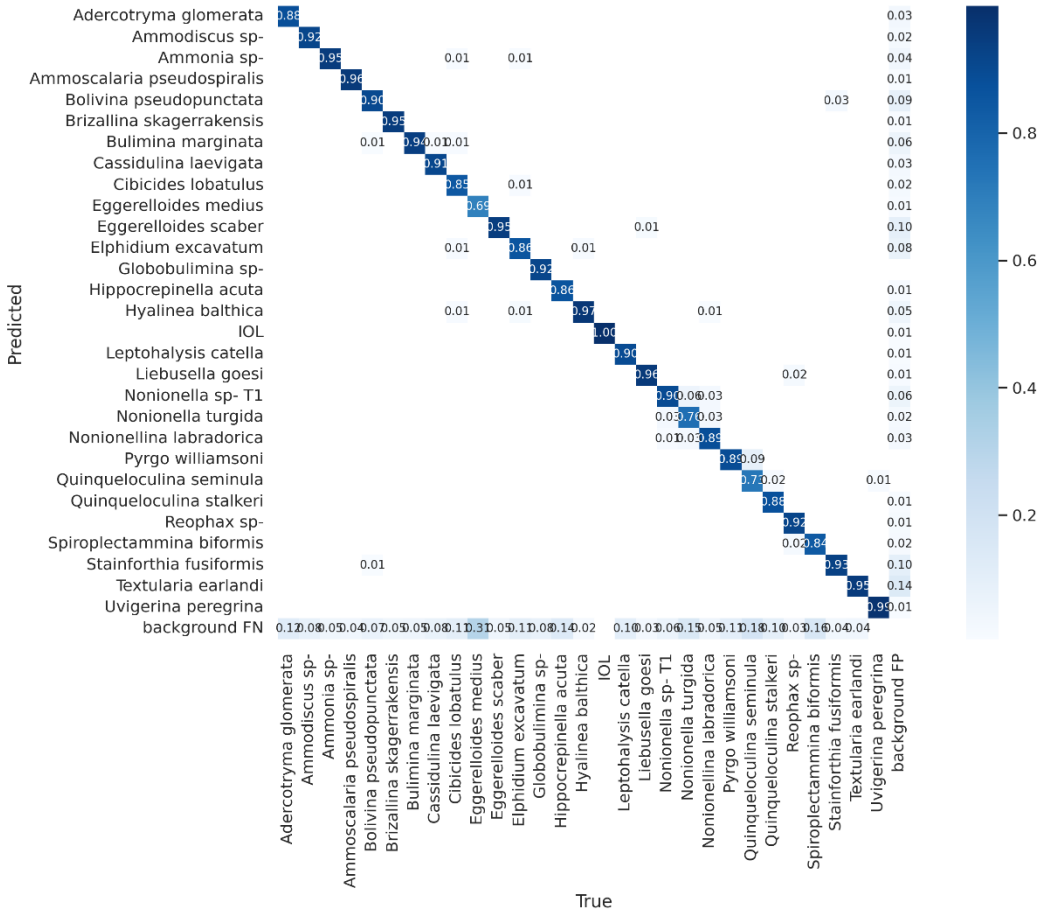


Figure A2. Confusion matrix of exp 10 model. The shade of the blue indicates the probability of the model to correctly identify the given species (only values>0 are shown).

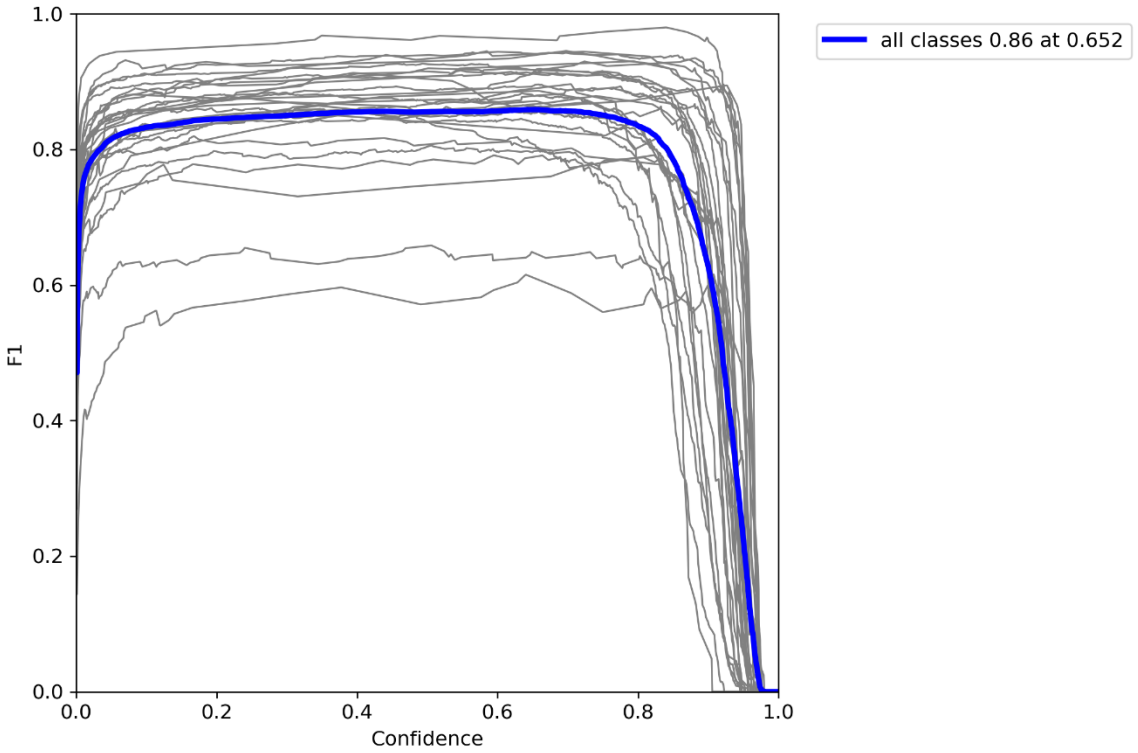


Figure A3. F1 score plot of exp 10 model.

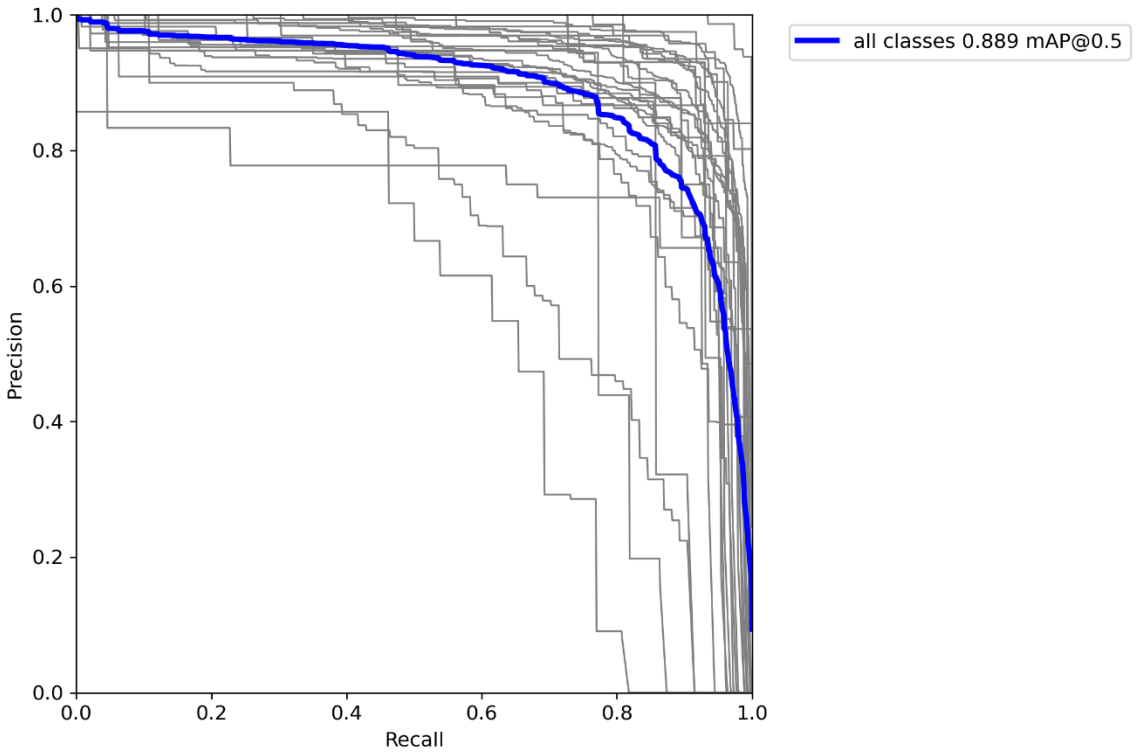


Figure A4. PR plot of exp 10 model.

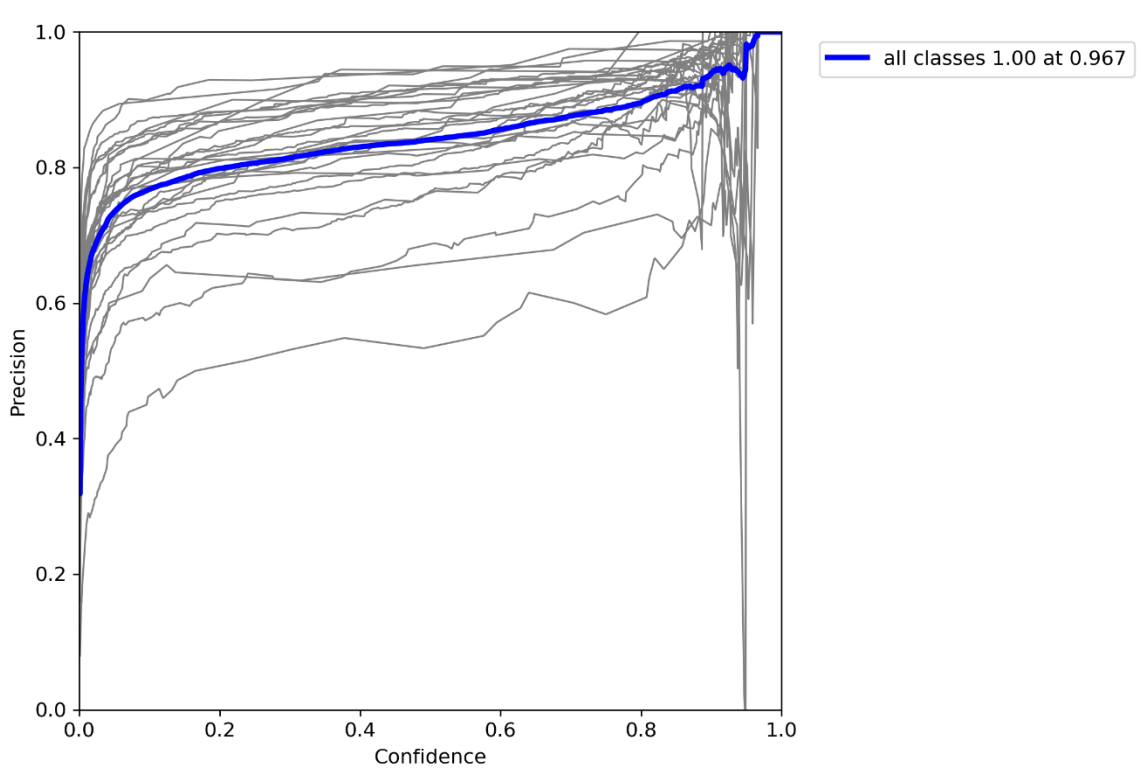


Figure A5. P plot of exp 10 model.

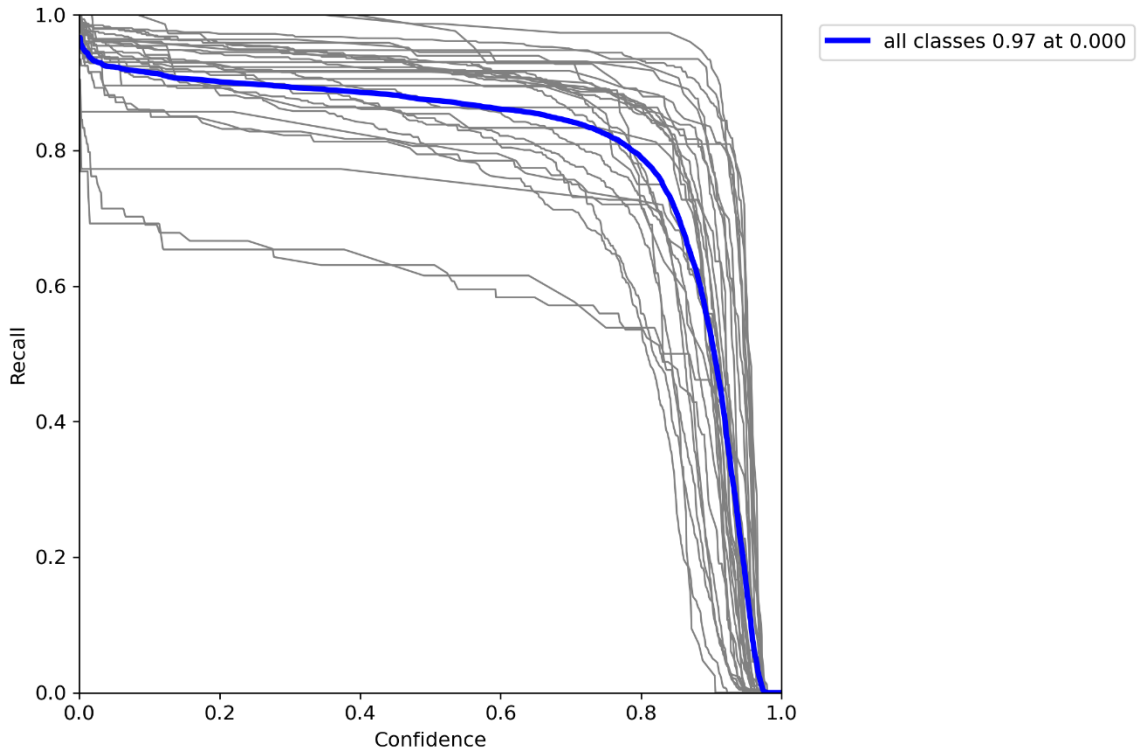


Figure A6. R plot of exp 10 model.

exp 51

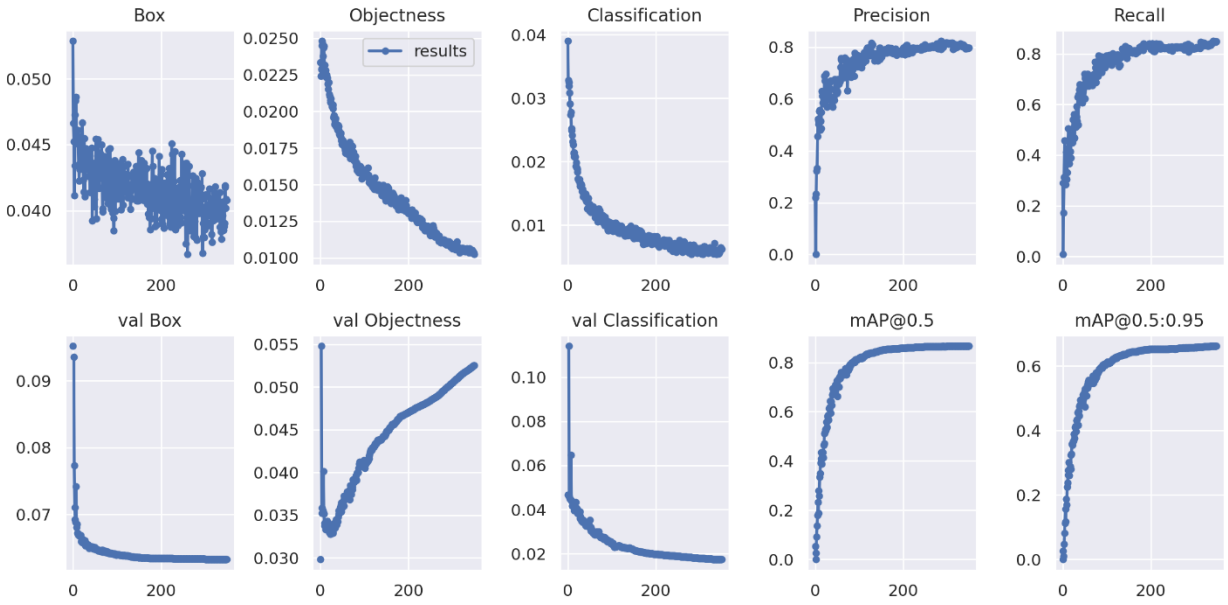


Figure A7. Results of box loss, objectness loss, classification loss, precision, recall and mean average precision (mAP) over the training epochs for the training and validation set of model exp 51.

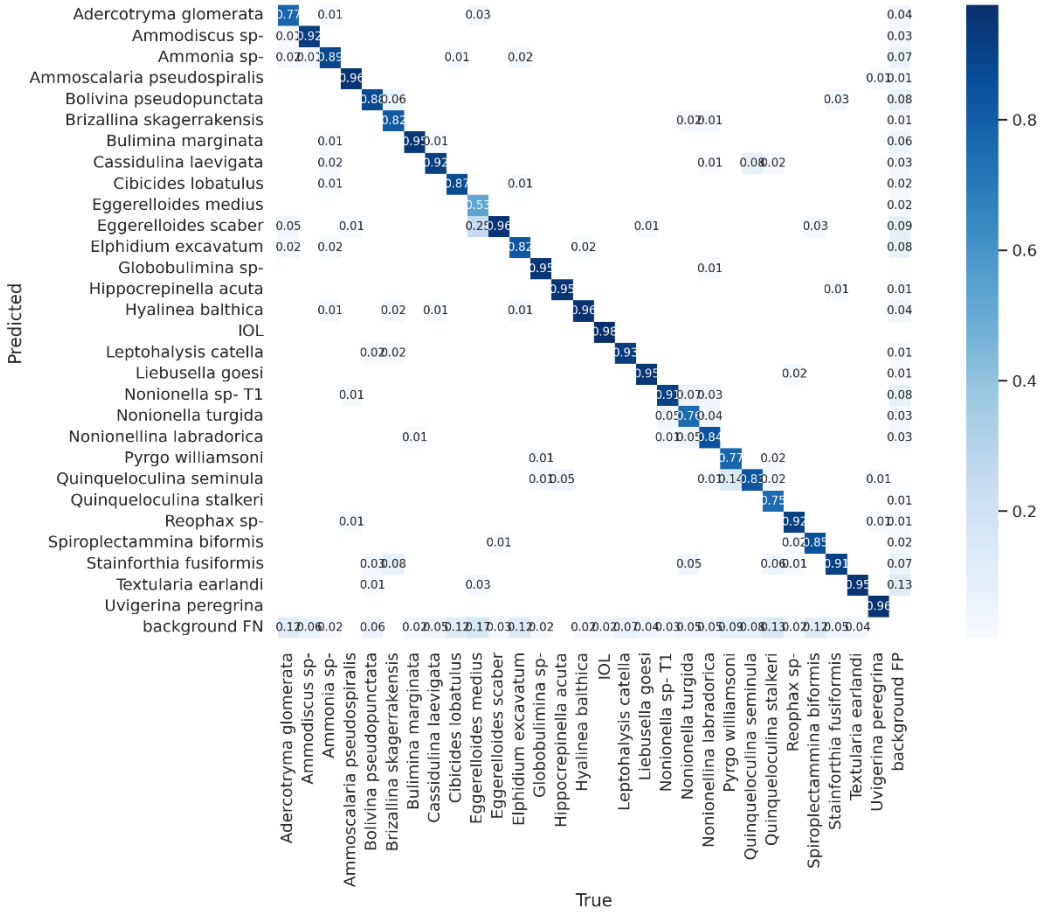


Figure A8. Confusion matrix of exp 51 model. The shade of the blue indicates the probability of the model to correctly identify the given species (only values>0 are shown).

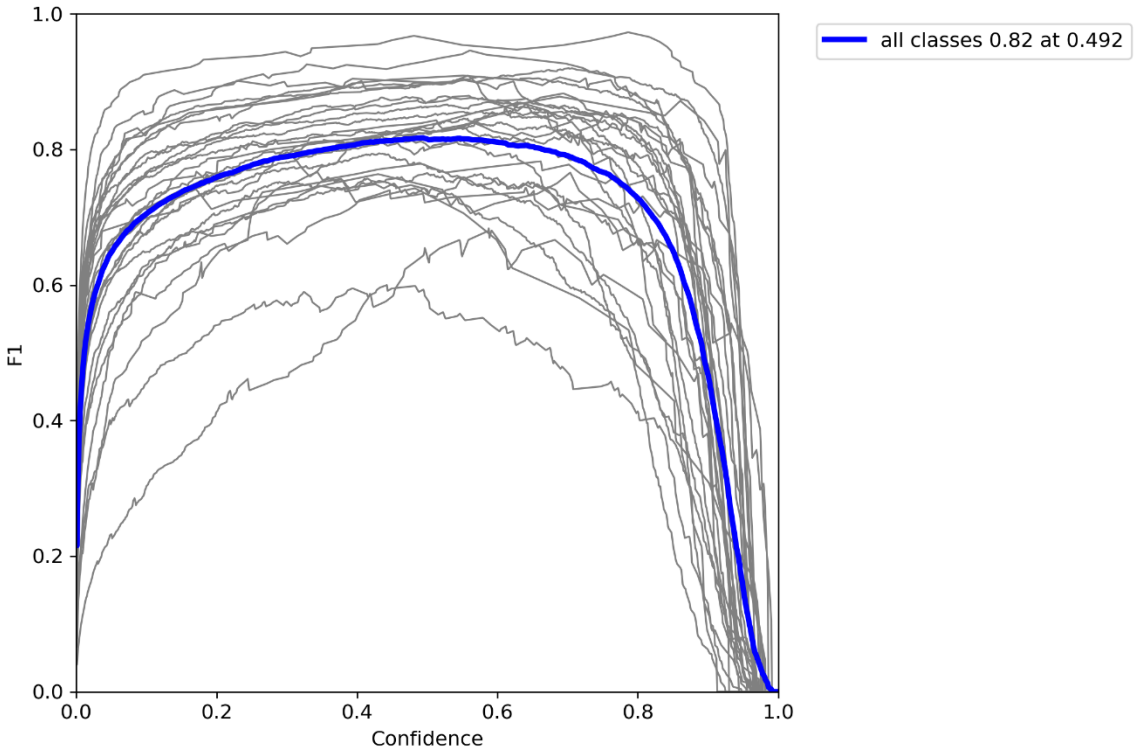


Figure A9. F1 score plot of exp 51 model.

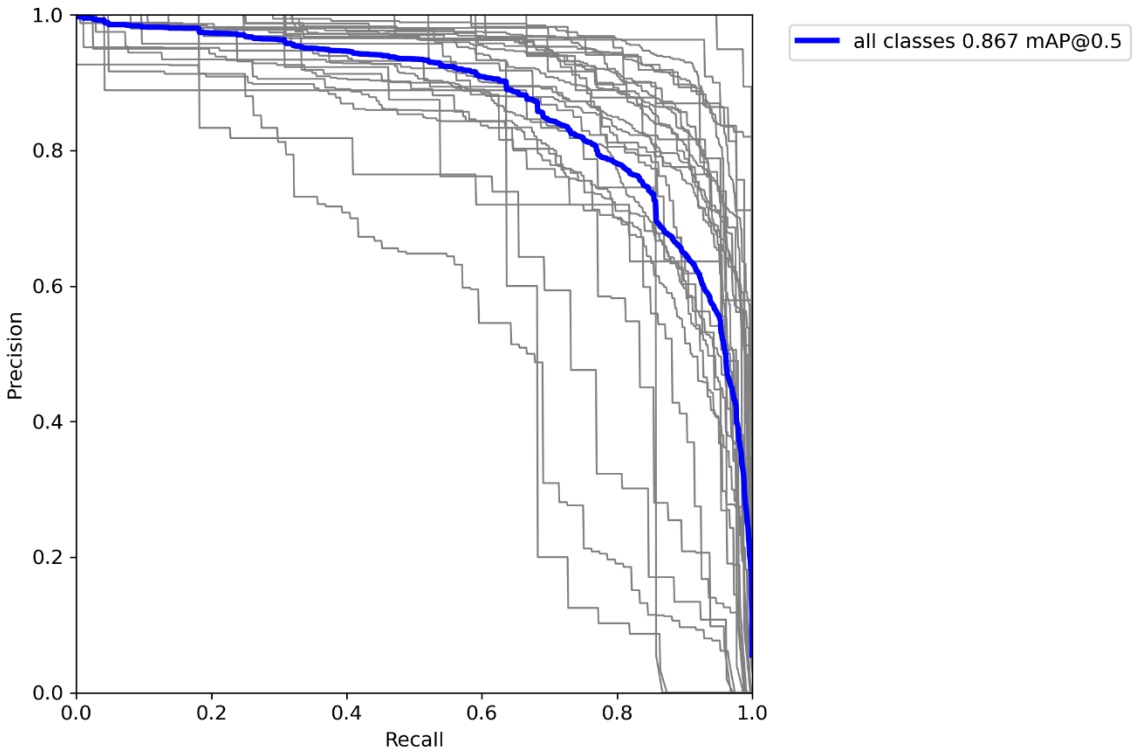


Figure A10. PR plot of exp 51 model.

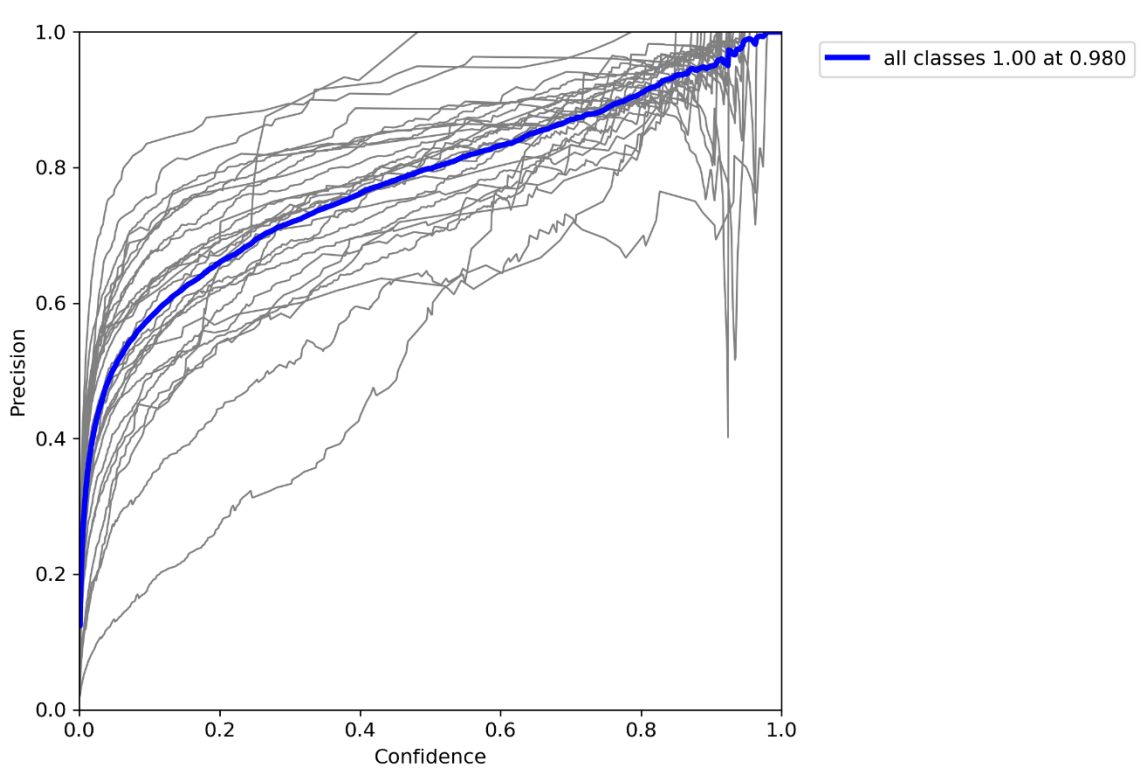


Figure A11. P plot of exp 51 model.

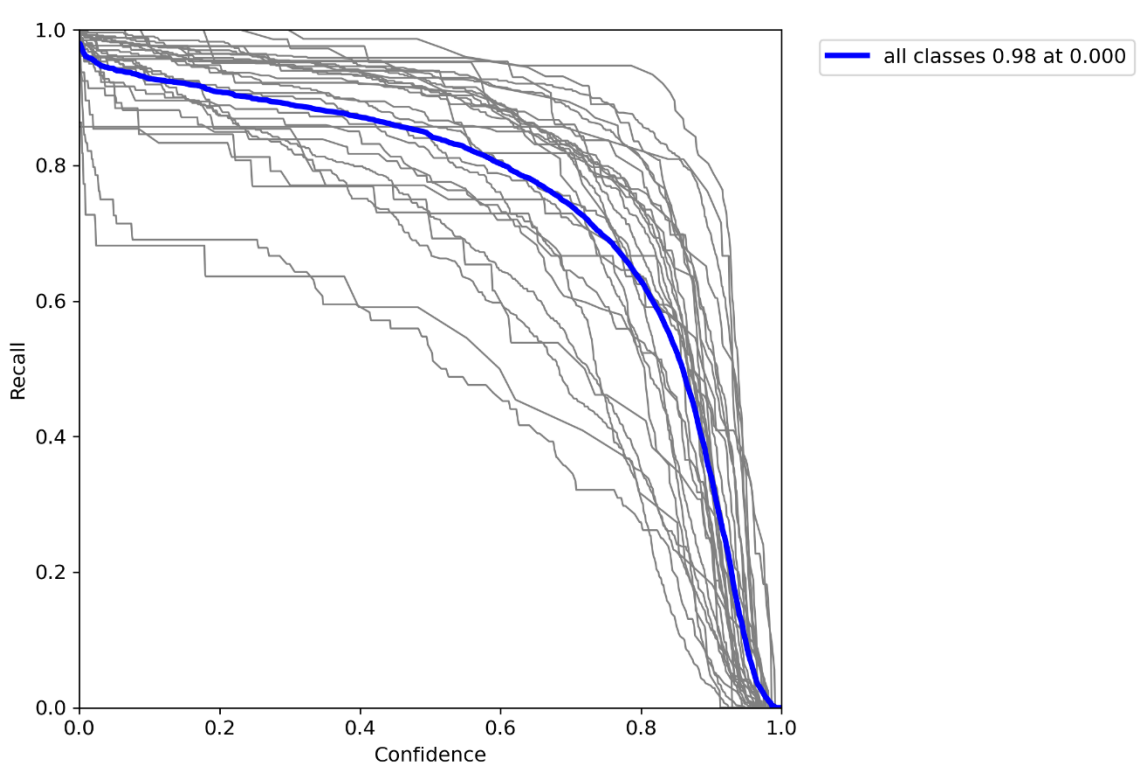


Figure A12. R plot of exp 51 model.

exp 52

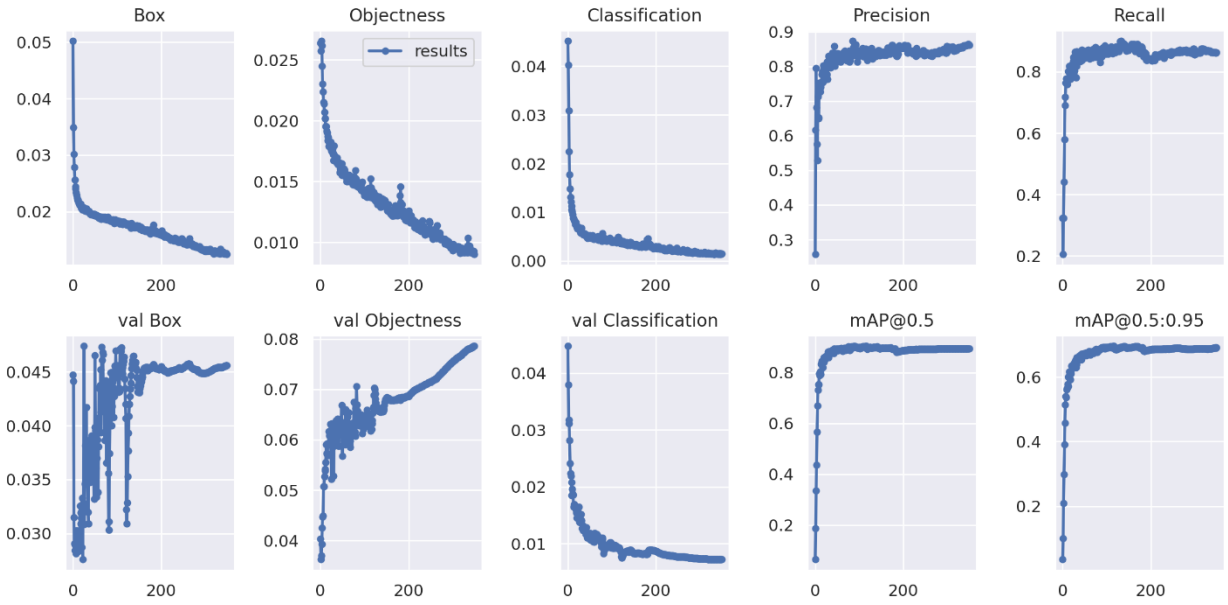


Figure A13. Results of box loss, objectness loss, classification loss, precision, recall and mean average precision (mAP) over the training epochs for the training and validation set of model exp 52.

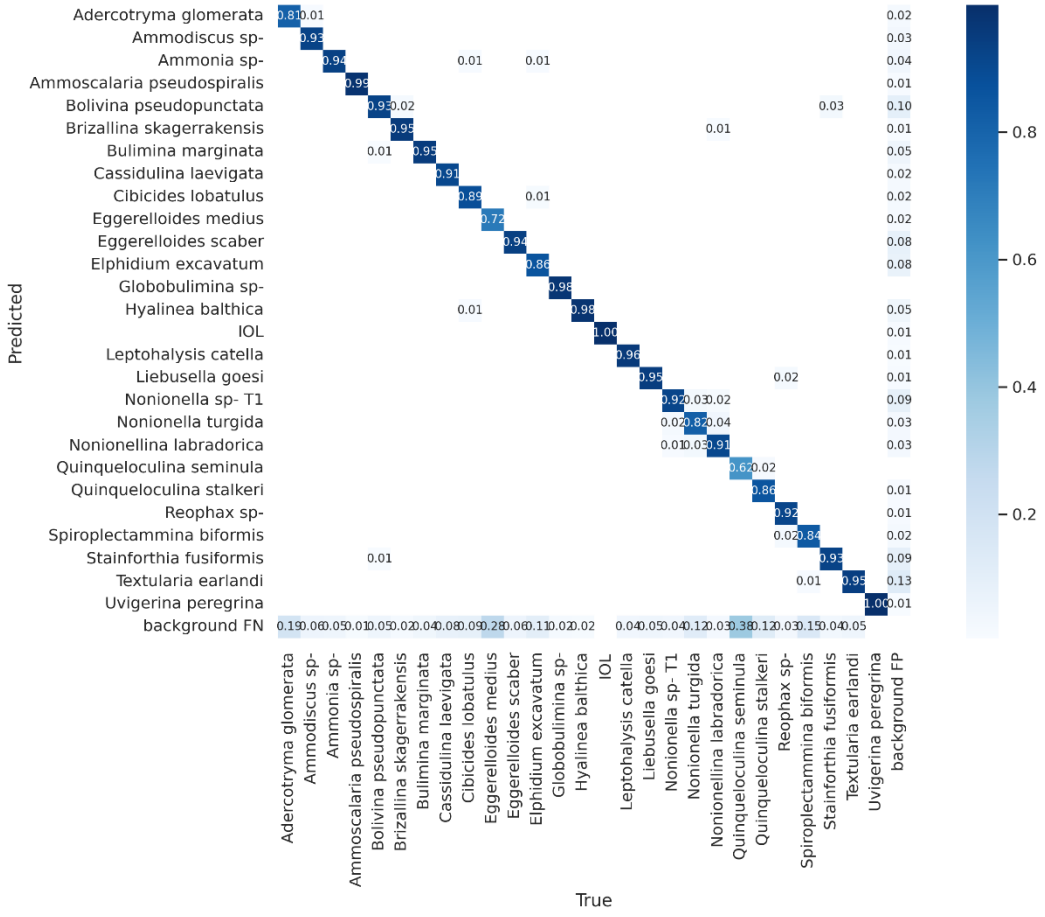


Figure A14. Confusion matrix of exp 52 model. The shade of the blue indicates the probability of the model to correctly identify the given species (only values>0 are shown).

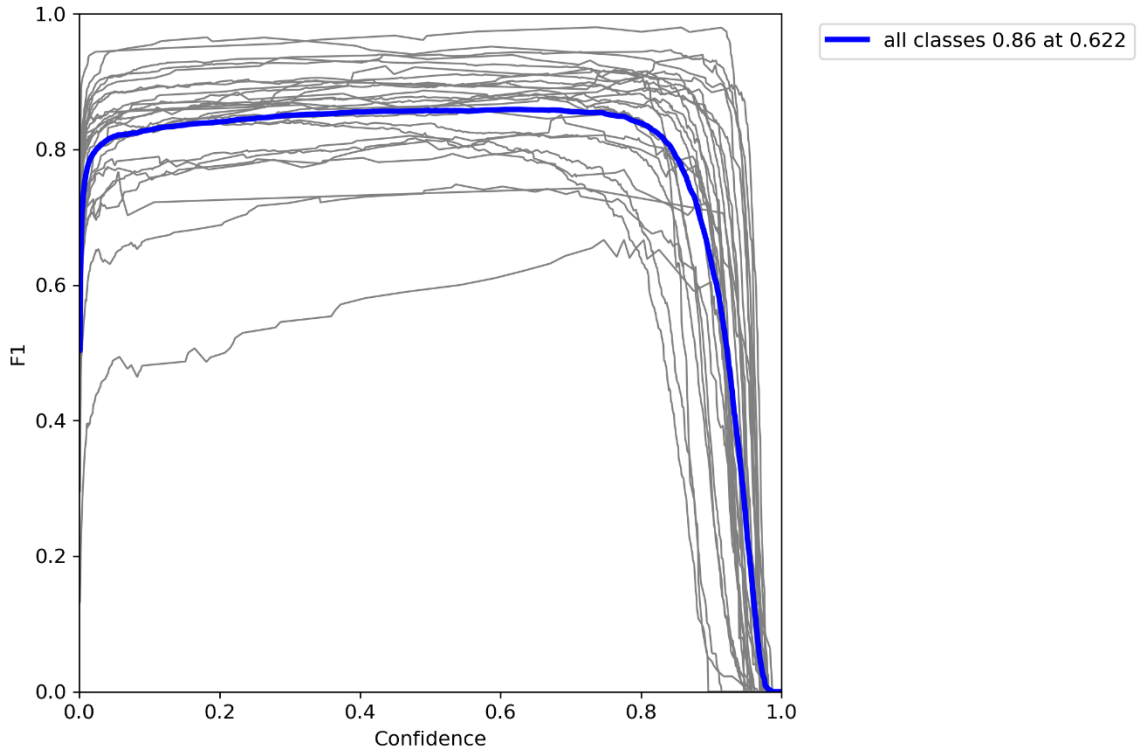


Figure A15. F1 score plot of exp 52 model.

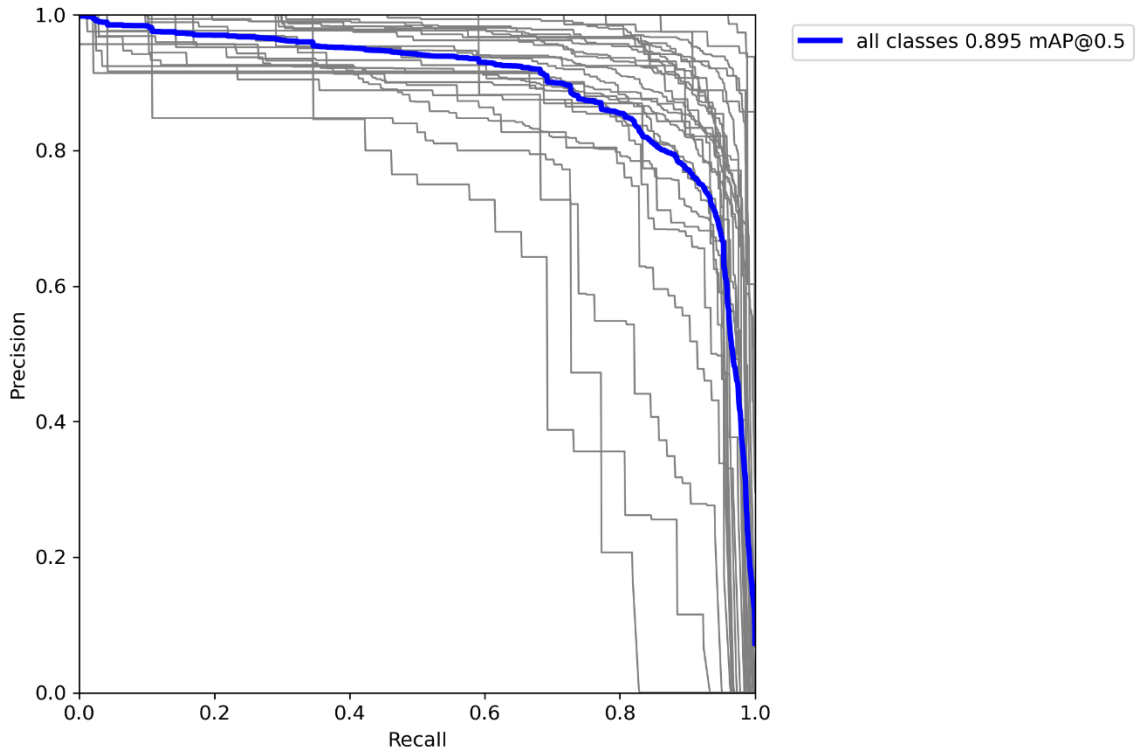


Figure A16. PR plot of exp 52 model.

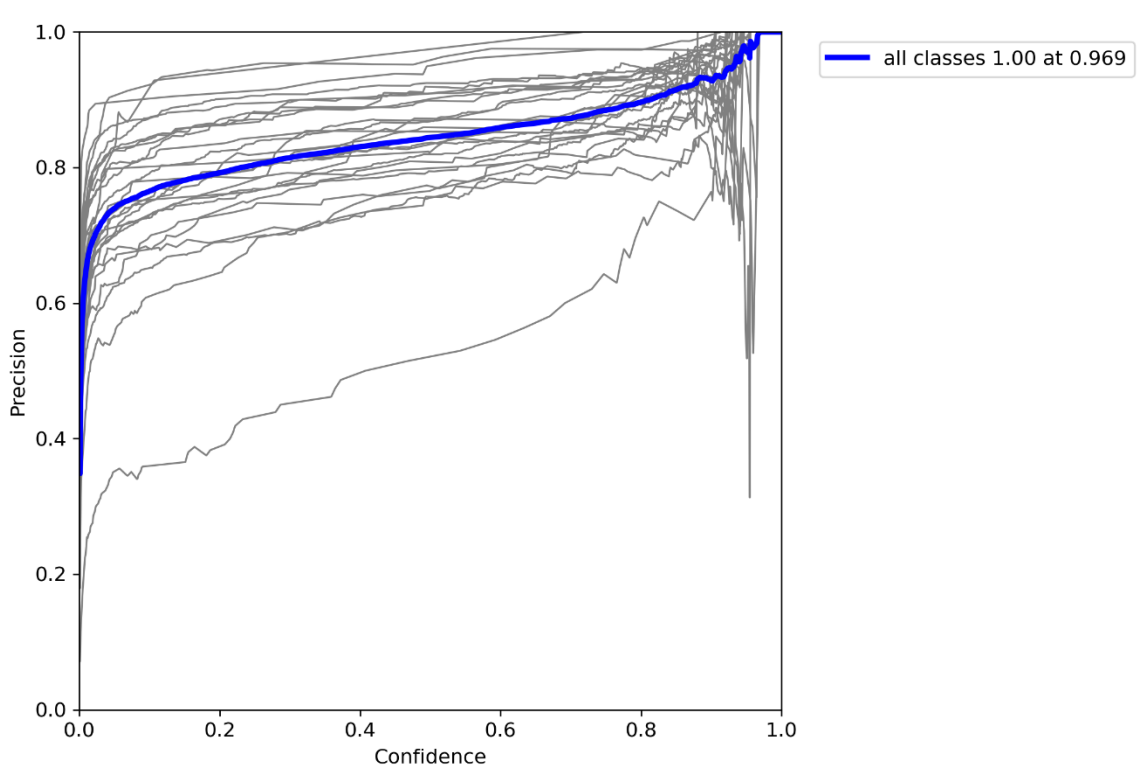


Figure A17. P plot of exp 52 model.

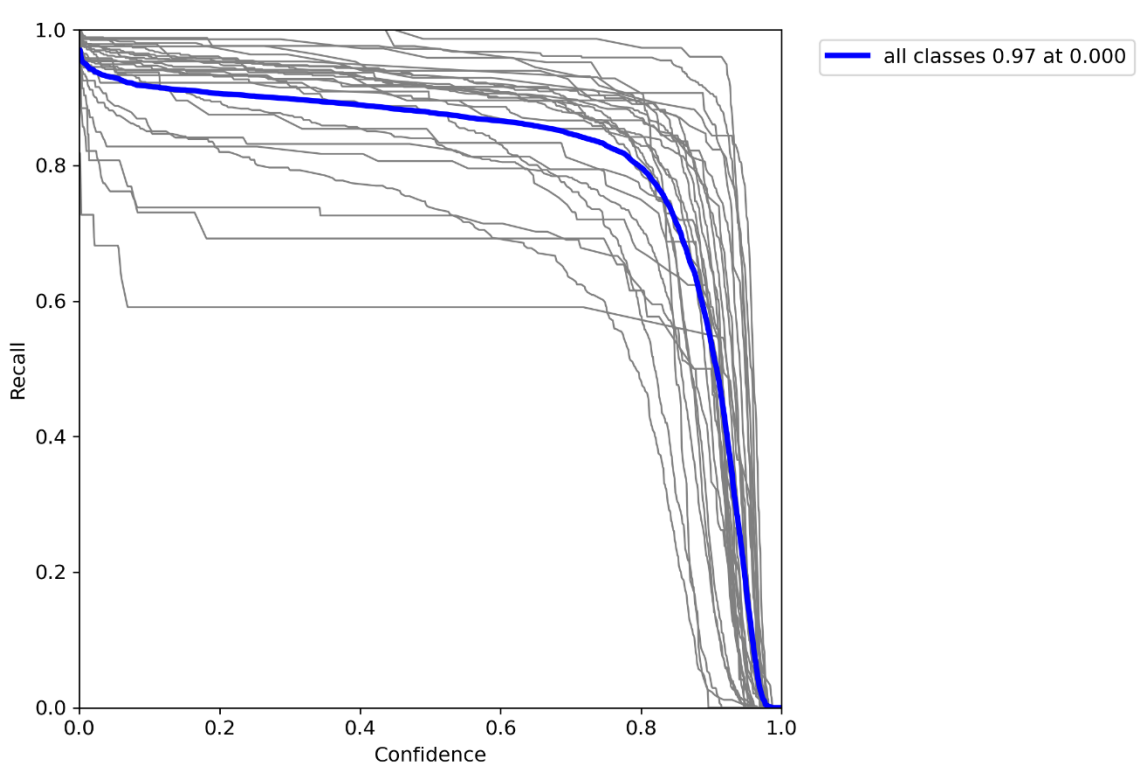


Figure A18. R plot of exp 52 model.

exp 53

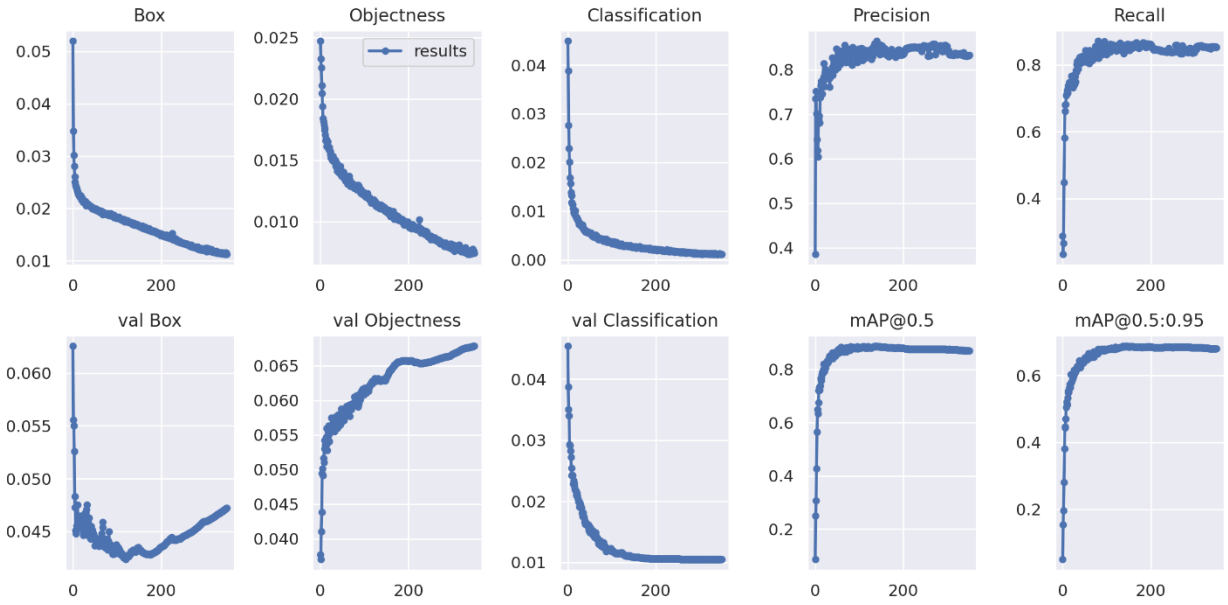


Figure A19. Results of box loss, objectness loss, classification loss, precision, recall and mean average precision (mAP) over the training epochs for the training and validation set of model exp 53.

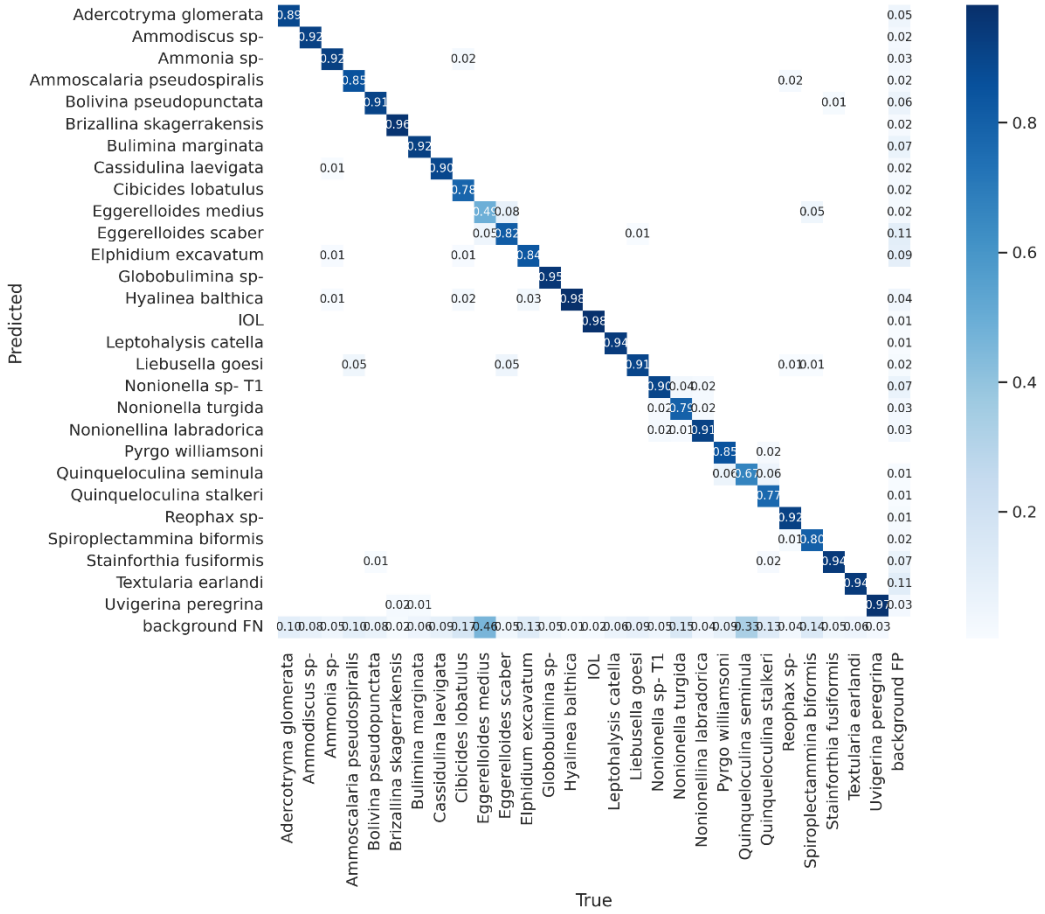


Figure A20. Confusion matrix of exp 53 model. The shade of the blue indicates the probability of the model to correctly identify the given species (only values>0 are shown).

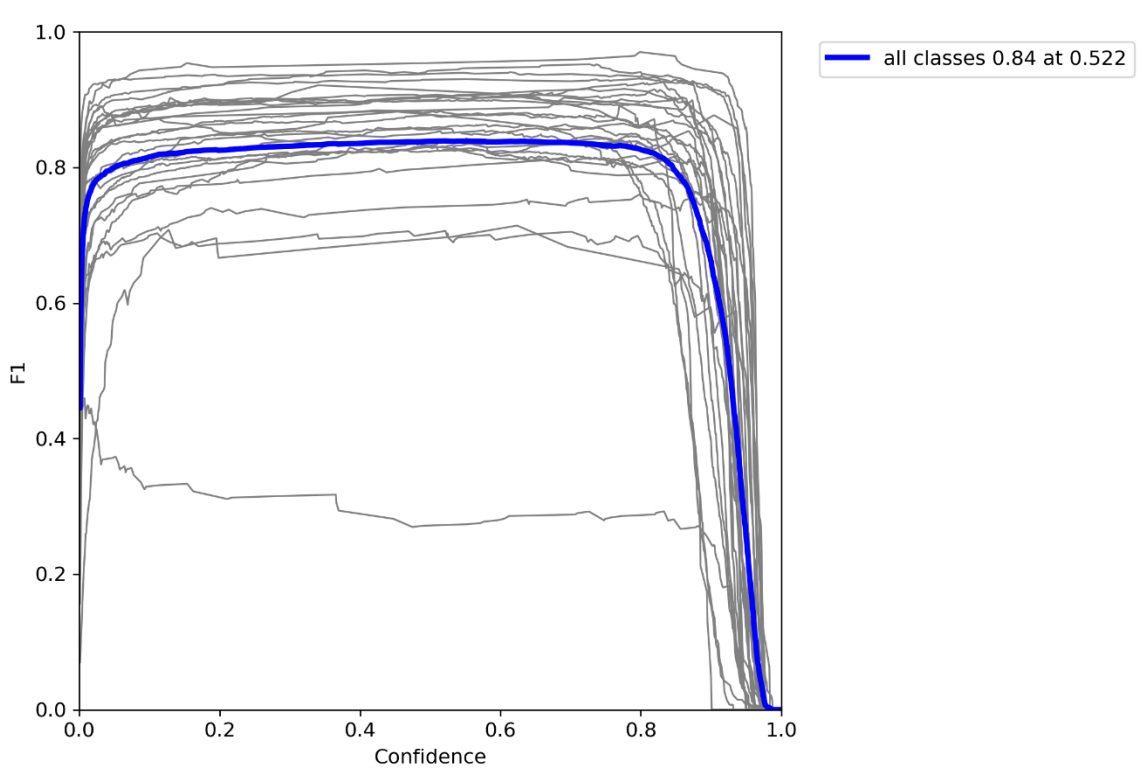


Figure A21. F1 score plot of exp 53 model.

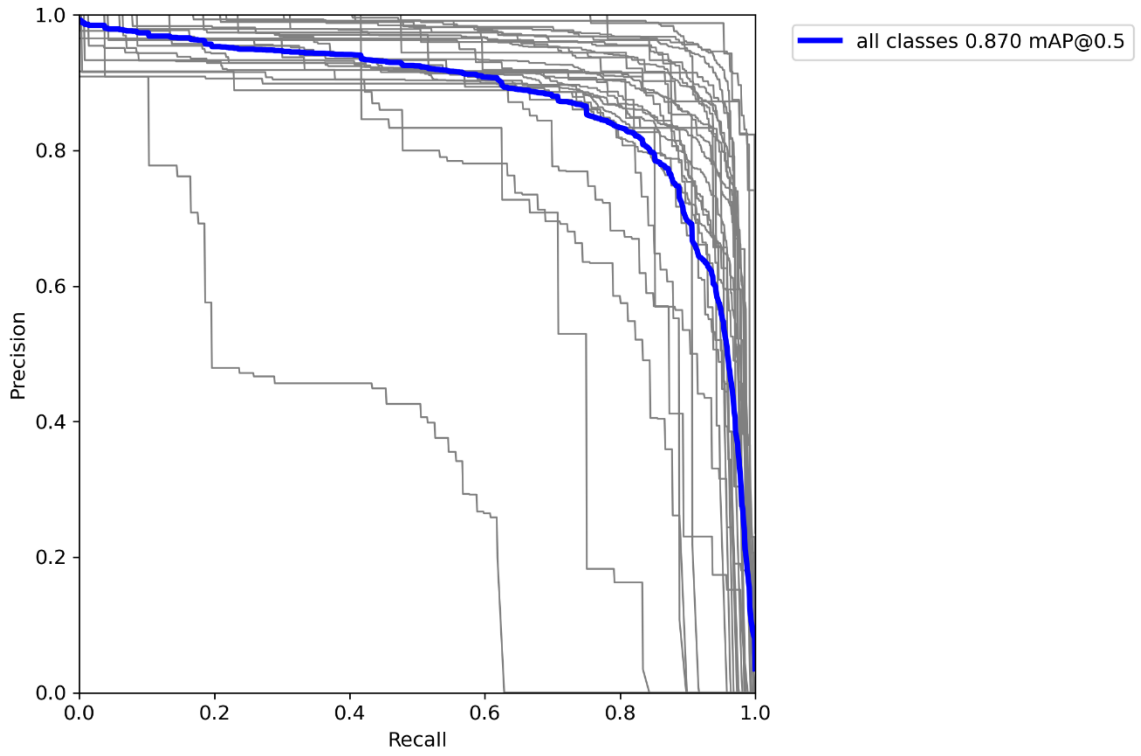


Figure A22. PR plot of exp 53 model.

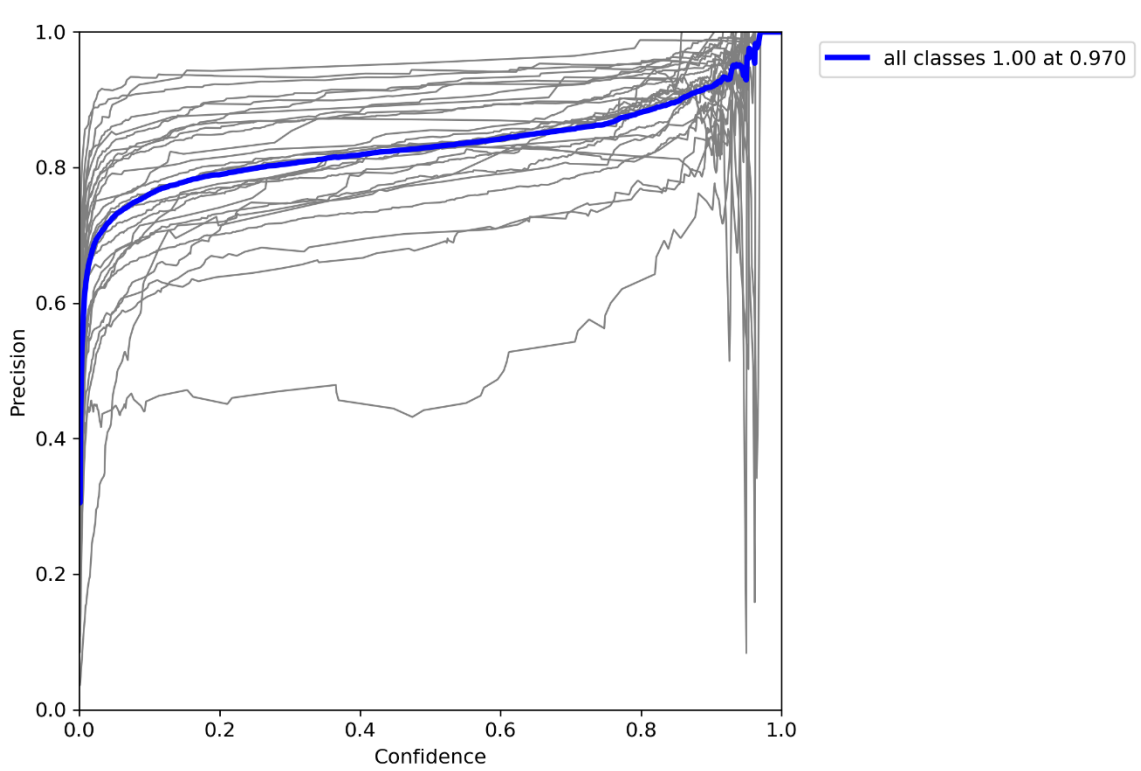


Figure A23. P plot of exp 53 model.

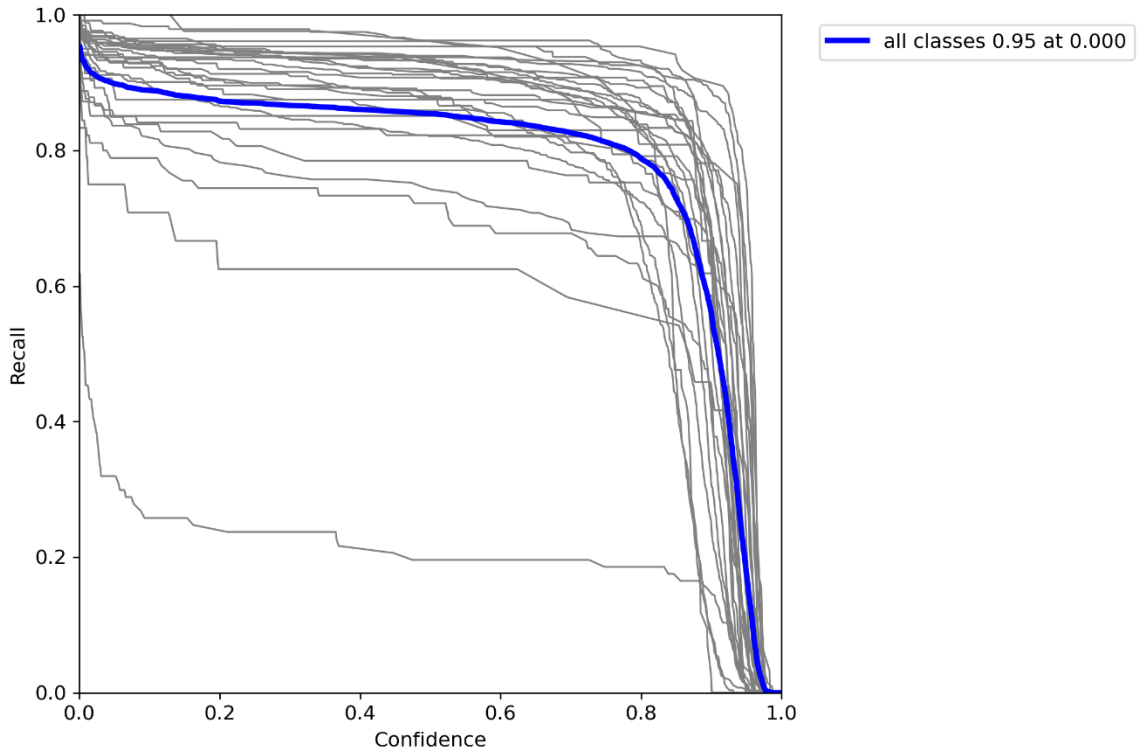


Figure A24. R plot of exp 53 model.

exp 56

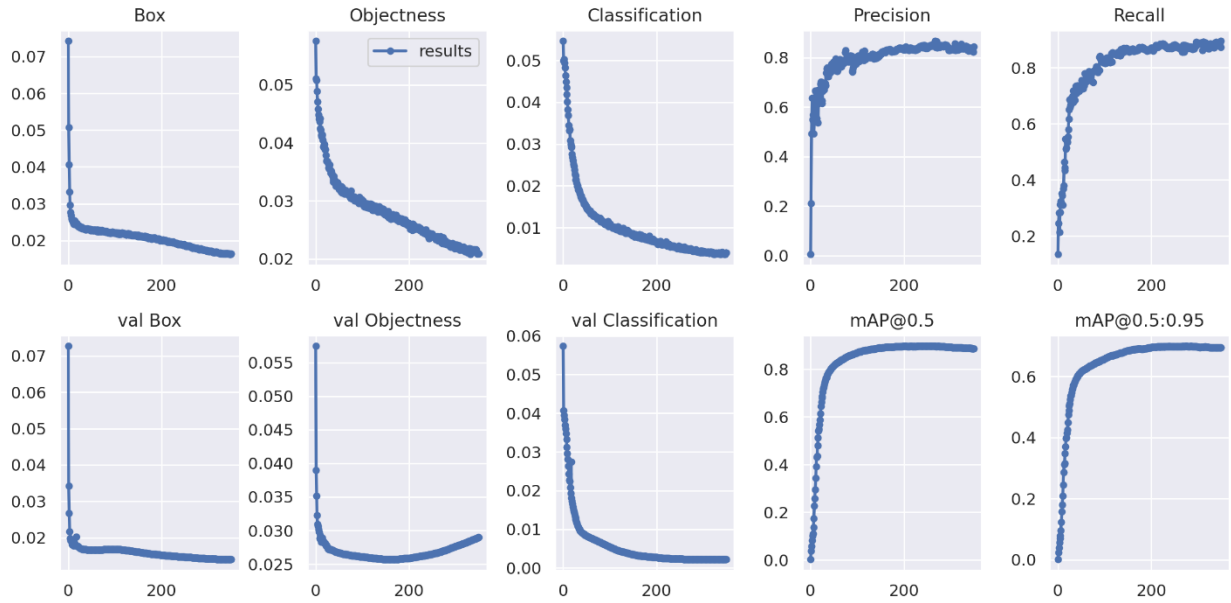


Figure A25. Results of box loss, objectness loss, classification loss, precision, recall and mean average precision (mAP) over the training epochs for the training and validation set of model exp 56.

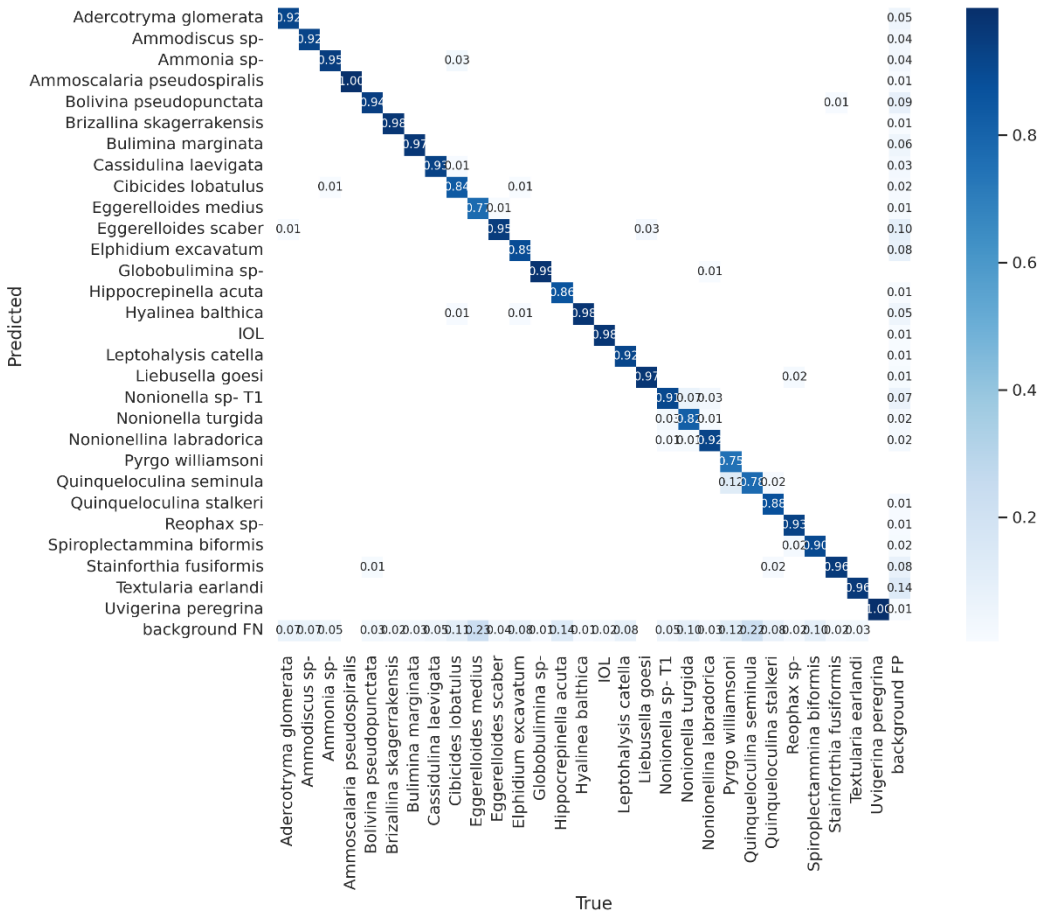


Figure A26. Confusion matrix of exp 56 model. The shade of the blue indicates the probability of the model to correctly identify the given species (only values>0 are shown).

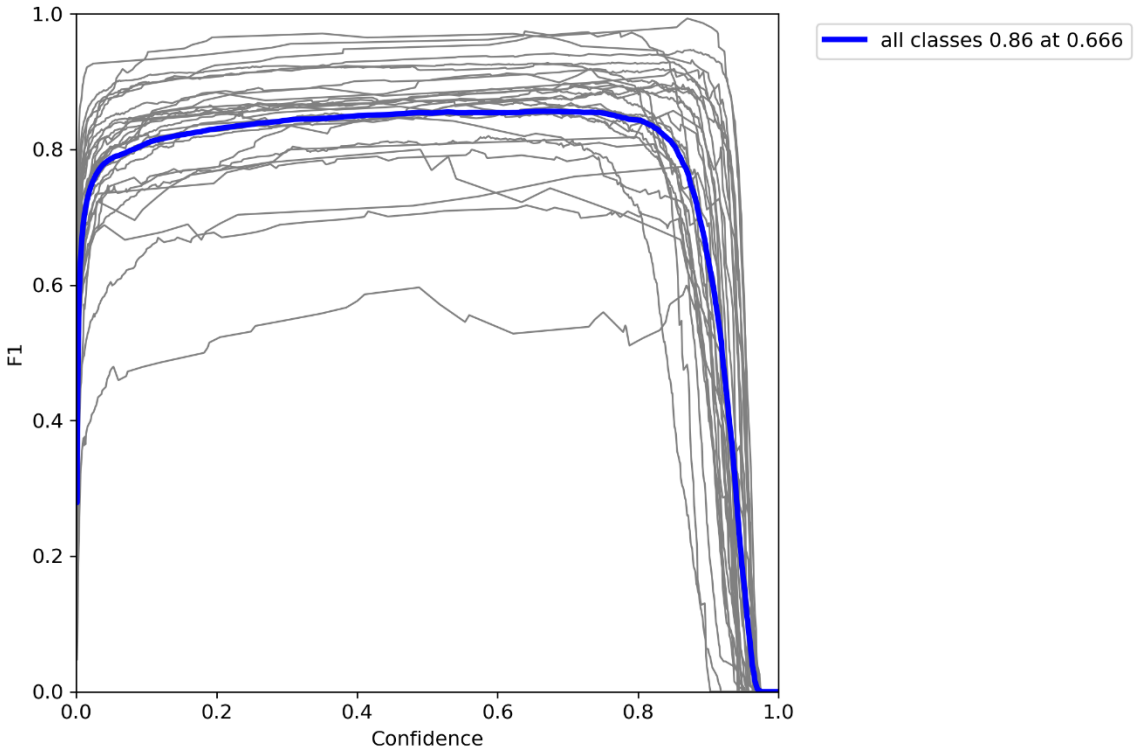


Figure A27. F1 score plot of exp 56 model.

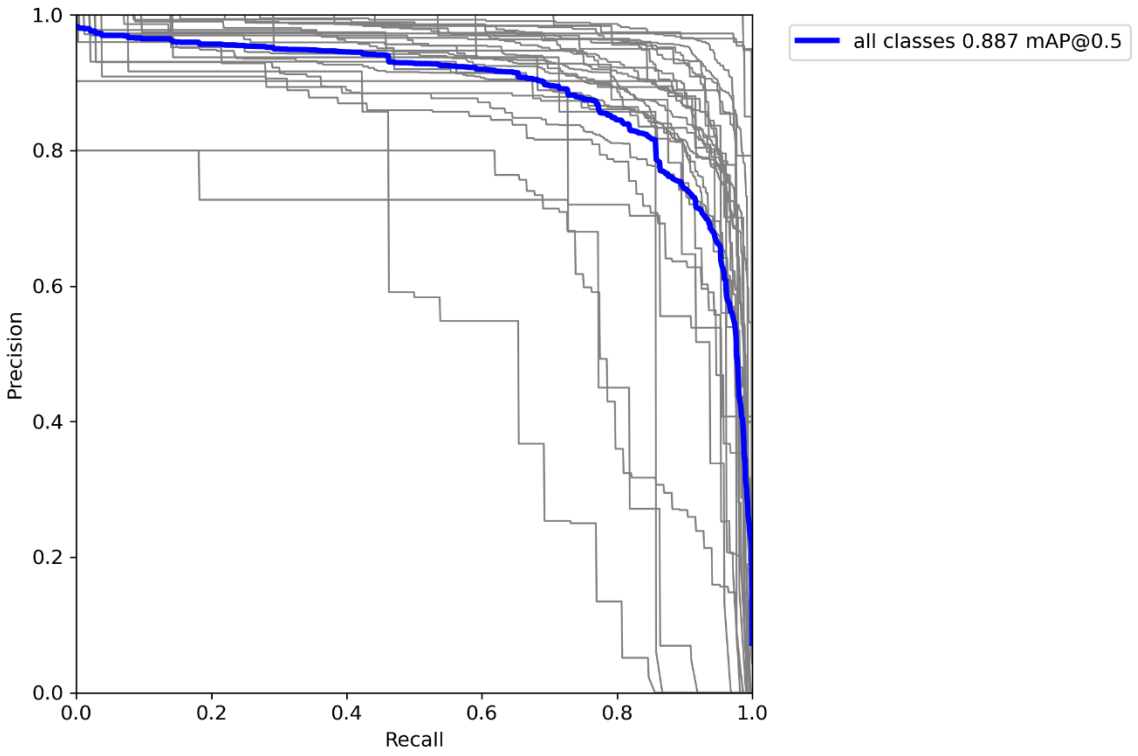


Figure A28. PR plot of exp 56 model.

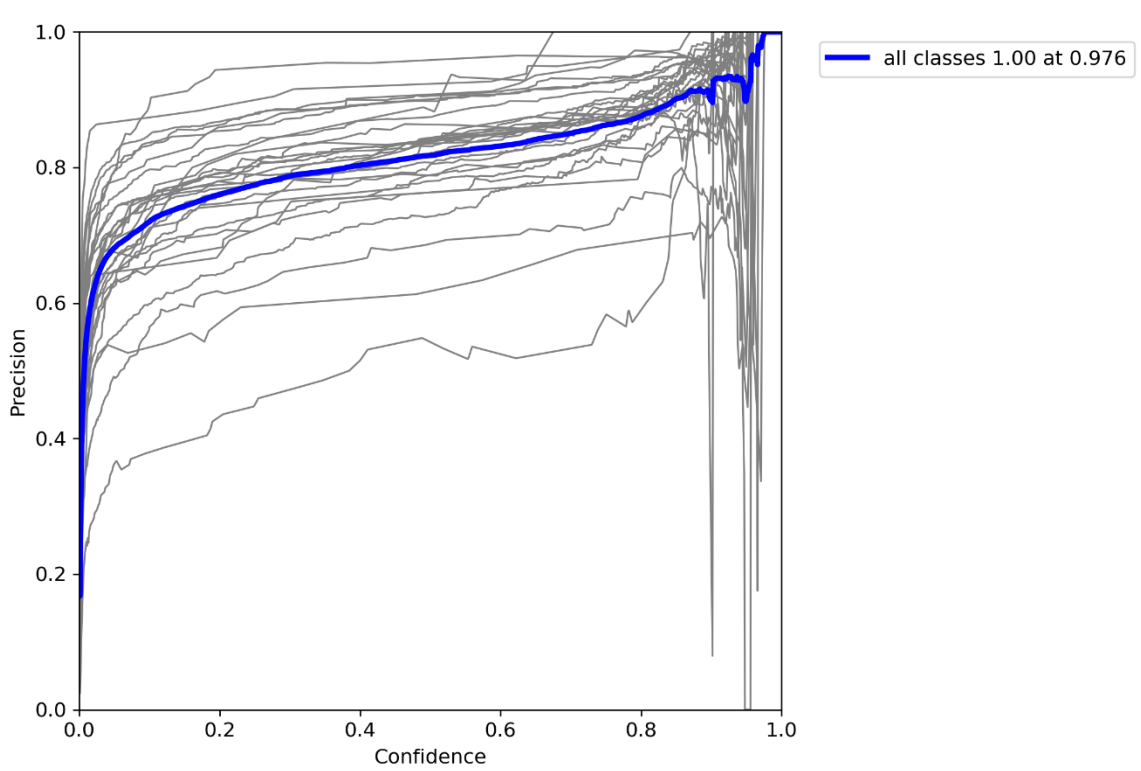


Figure A29. P plot of exp 56 model.

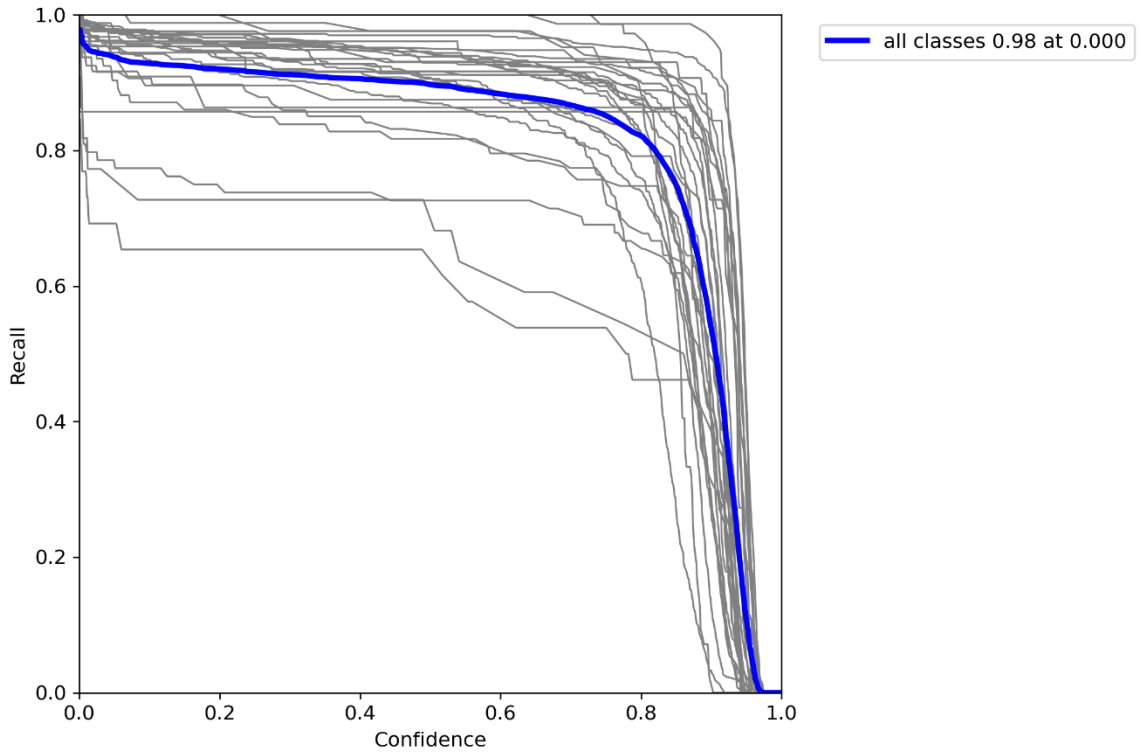


Figure A30. R plot of exp 56 model.

exp 58

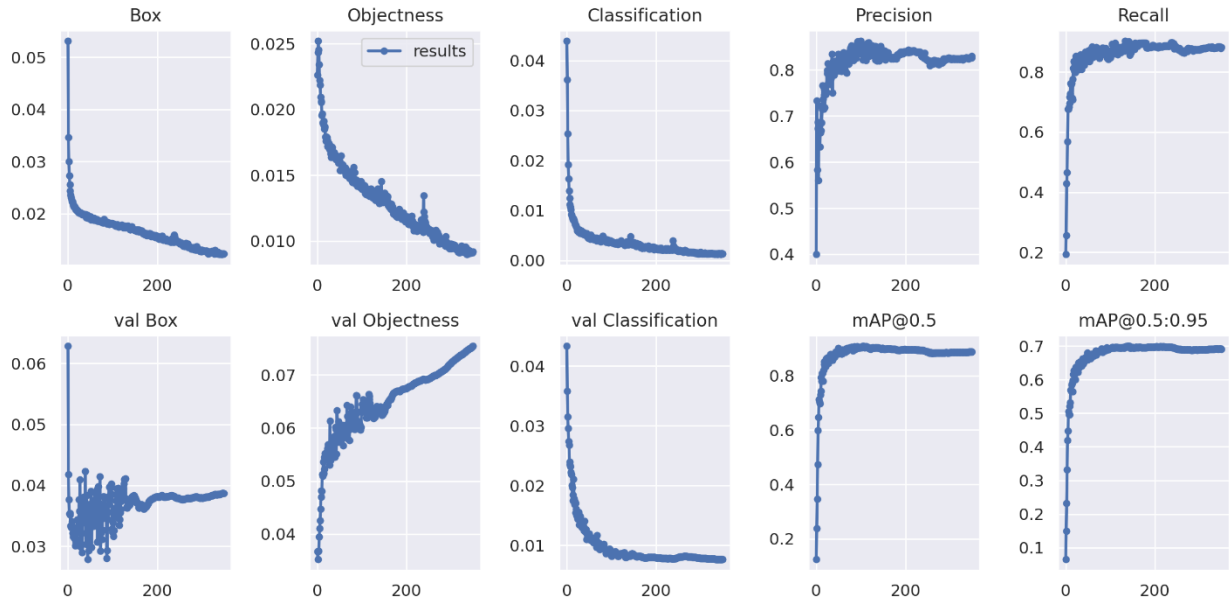


Figure A31. Results of box loss, objectness loss, classification loss, precision, recall and mean average precision (mAP) over the training epochs for the training and validation set of model exp 58.

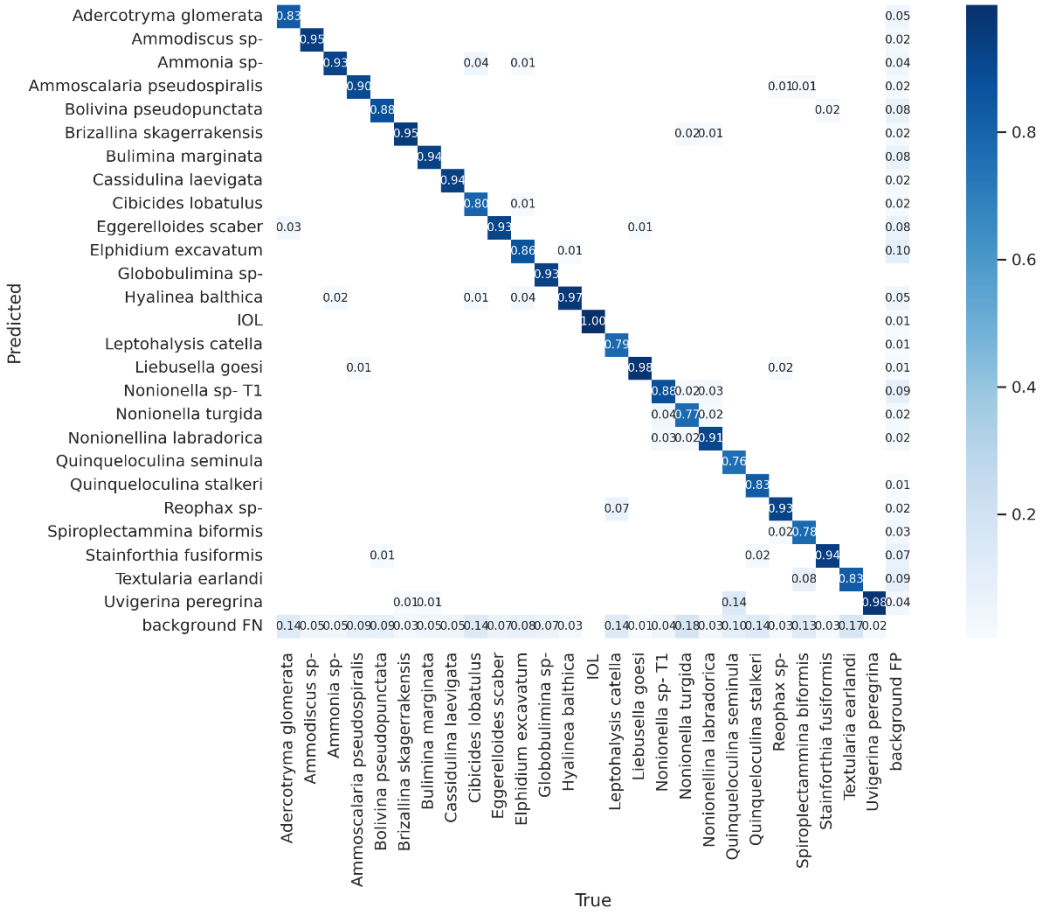


Figure A32. Confusion matrix of exp 58 model. The shade of the blue indicates the probability of the model to correctly identify the given species (only values>0 are shown).

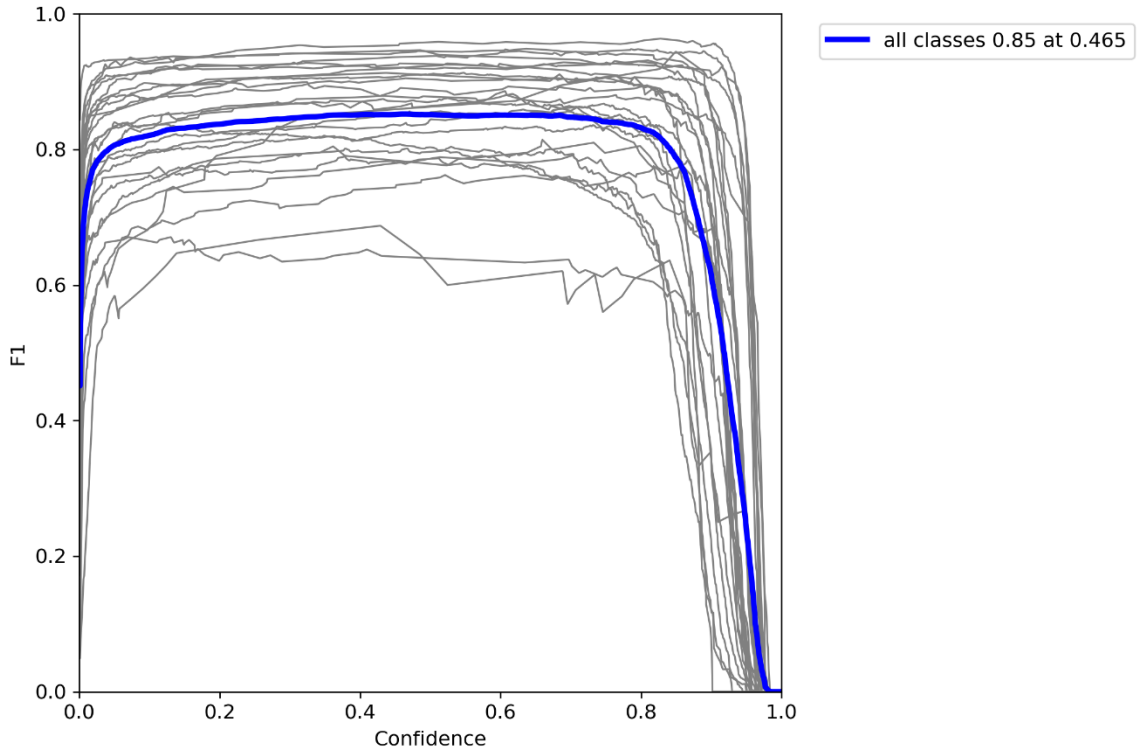


Figure A33. F1 score plot of exp 58 model.

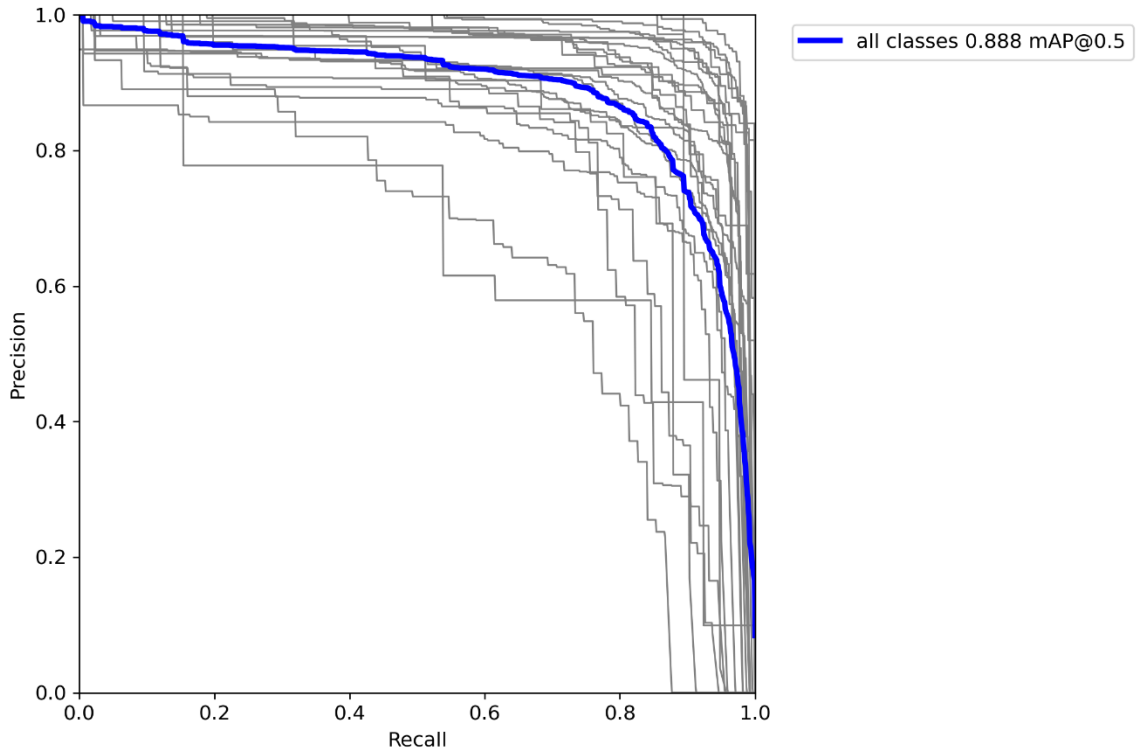


Figure A34. PR plot of exp 58 model.

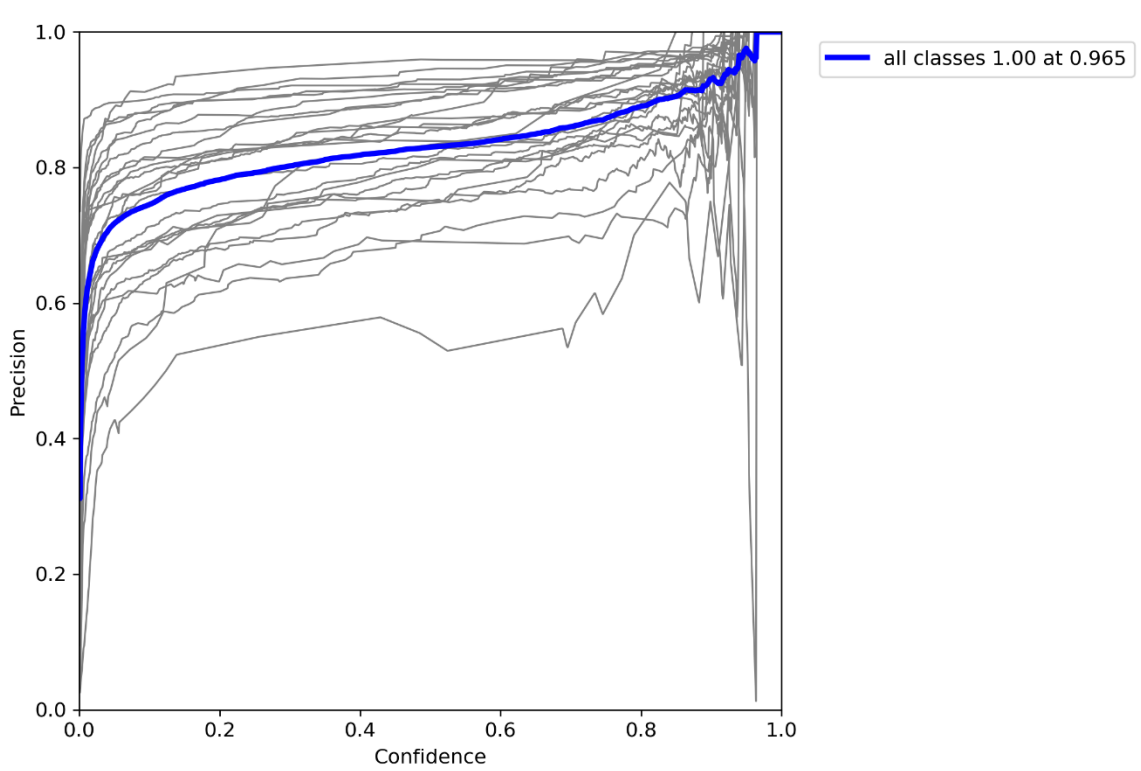


Figure A35. P plot of exp 58 model.

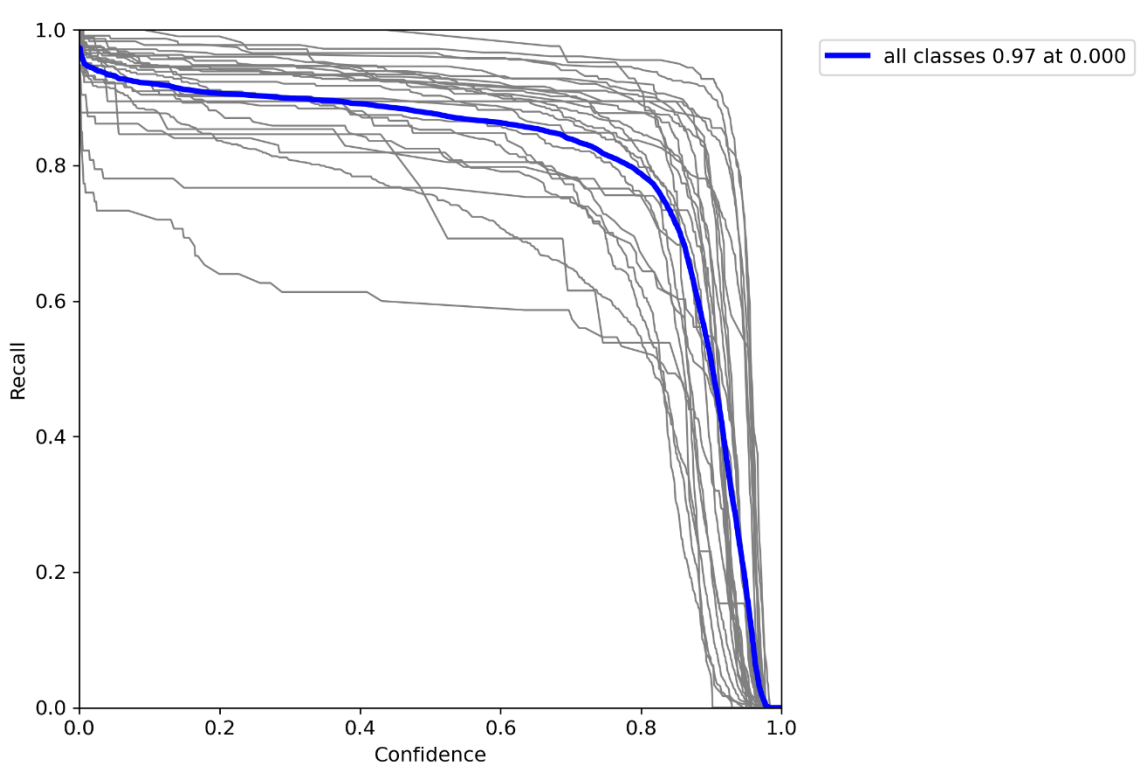


Figure A36. R plot of exp 58 model.