Detecting Floating Plastic Debris

Assessment of using few-shot meta-learning approach with active learning methods to detect floating plastic debris on satellite images

AE5810: MSc Thesis

Dilge Gül

TU Delft

EPFL

Detecting Floating Plastic Debris

Assessment of using few-shot meta-learning approach with active learning methods to detect floating plastic debris on satellite images

by

Dilge Gül

for partial fulfilment of the requirements for the degree of
Master of Science at the Delft University of Technology,
to be defended on 12 July 2023 at 14:00.

Version	**Date**	**Changes**	**Comments**
1.0	22/05/2023	N/A	First draft for the green light meeting
2.0	27/06/2023	Missing sections are completed, the literature study is restructured, the knowledge gaps are updated, the materials and method sections of the research article are edited for better clarity, the results section is extended with qualitative results, the future work section is improved	Final version

Student number:	4821165
Project Duration:	1 November 2022 - 12 July 2023
Faculty:	Faculty of Aerospace Engineering, Delft
Supervisors:	Dr. Jurgen Vanhamel
	Prof. Dr. Devis Tuia
	Dr. Marc Rußwurm

Cover: Rich Carey on Shutterstock.com (Modified)

TU Delft EPFL

Summary

Plastics, known for their lightweight and durability, can pose significant environmental problems when they become marine debris. Marine animals get entangled in this debris, their habitats are destroyed, and invasive species are transported to non-invaded regions. With the rapid increase in plastic debris in the oceans, the impact on human health is not yet fully understood.

A promising approach to monitor marine plastic debris is the combination of remote sensing and machine learning techniques. Early attempts involved hand-engineered spectral features and simple machine learning classifiers on satellite images. Subsequently, deep learning methods were employed, utilizing the full multi-spectral data from satellites. However, these methods require large amounts of training data, which are not readily available for floating marine debris. Therefore, new models that can be trained with limited data are needed. One promising approach is the combination of few-shot meta-learning and active learning.

Active learning involves selecting informative samples for annotation to enhance learning capacity, while few-shot meta-learning focuses on quickly adapting models to limited labeled examples. By actively querying and selecting informative samples, active few-shot meta-learning improves generalization and adaptation capabilities, resulting in better performance and faster adaptation in few-shot learning scenarios.

In this study, the main goal was to assess the effectiveness of an active learning approach combined with a few-shot meta-learning model in detecting floating marine debris. Different active learning strategies were tested alongside the METEOR model, which is specifically designed for various Earth observation challenges. METEOR utilizes a deep ResNet-12 architecture and a special meta-learning algorithm to extract important information from extensive land cover classification data. The extracted knowledge, known as the "meta-model," can then be transferred from previous tasks to the current task using model-based transfer learning. The study also employed ResNet-18, a widely-used deep neural network architecture in computer vision tasks, as a comparison model.

The study compared various sampling strategies for creating the support set of the METEOR model, evaluating their performance compared to random sampling and the ResNet-18 model. Two groups of sampling strategies were tested: uncertainty-based active learning methods and diversity-based active learning methods. Uncertainty-based methods measure the model's uncertainty in predicting sample labels and select the most uncertain samples, while diversity-based sampling strategies aim to maximize the diversity of samples in the support set by considering their representation in the chosen feature space.

To evaluate the sampling strategies, the study focused on recall and average precision as metrics, as accuracy can be misleading in the presence of class imbalance. The results consistently showed that active learning methods incorporating uncertainty-based sampling, such as entropy and query by committee, outperform other strategies in terms of recall and average precision. Diversity-based methods employed in this study struggled to select informative and representative samples, emphasizing the importance of considering feature space representativeness. The performance of the detection model was also influenced by regional characteristics, with the ResNet-18 model showing enhanced performance in the Accra region compared to Durban region. Additionally, class imbalance affected the performance of the active learning framework, with random sampling showing greater improvements in the Accra region compared to Durban. Overall, the study contributed to the field of floating marine debris detection by highlighting the value of utilizing sampling techniques for a few-shot meta-learning model.

The study suggested several directions for future work. While few-shot learning models show promise for marine debris detection with limited labeled samples, effective sampling and training strategies need to be developed. Specifically, for cluster-based methods, testing various feature spaces and exploring different cluster selection methods could yield better results. Additionally, a custom atmospheric correction for each region of interest could further improve performance.

Contents

Abbreviations

Abbreviation	Definition
CNN	Convolutional Neural Network
EM	Electromagnetic
EPFL	Swiss Federal Institute of Technology Lausanne
FDI	Floating Debris Index
GEE	Google Earth Engine
GESAMP	Group of Experts on the Scientific Aspects of Marine Environmental Protection
MAML	Model Agnostic Meta-Learning
METEOR	Meta-Learning Framework for Earth Observation Problems Across Different Resolutions
MSI	Multi-Spectral Instrument
NB	Naïve Bayes
NDVI	Normalized Difference Vegetation Index
PI	Plastic Index
QBC	Query by Committee
RF	Random Forest
RGB	Red Green Blue
SSO	Single Shot Oracle
SVM	Support Vector Machine

List of Figures

List of Tables

1

Background & Report Outline

Plastics are used often due to their lightweight and durability. However, these useful traits can become adversities once plastic objects turn into marine debris and cause significant problems for the environment [1]. This debris ends up entangling marine animals or destroying their habitats, and transporting invasive species to non-invaded regions [2]. The amount of plastic debris ending up in the oceans is increasing rapidly as can be seen in Figure 1.1 [3]. It is not yet understood how this impacts human health, but it is accepted to be a potential hazard [2, 4]. Given the results of the assessment made by the Group of Experts on the Scientific Aspects of Marine Environmental Protection (GESAMP) and the efforts of the United Nations to remediate the marine plastic pollution problem, investments into research on understanding this problem are warranted [5].

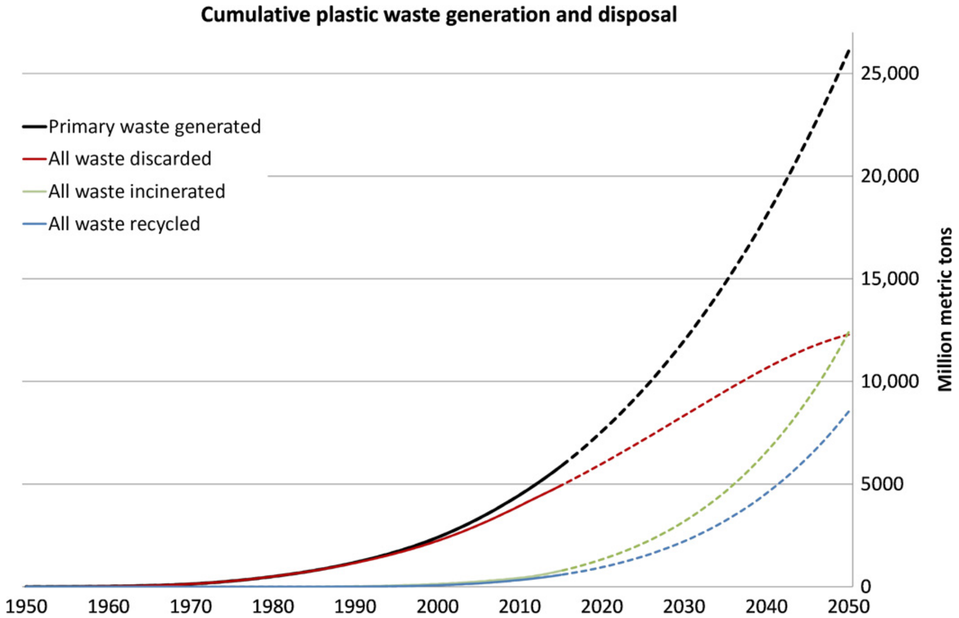


Figure 1.1: Generation and disposal rates of plastic waste between 1950-2015. Dashed lines after 2015 represent projections of the previous trends [3].

In the frame of discovering the threats of plastic pollution in the marine environment, significant clean-up efforts have been made along with implementing measures on limiting the amount of plastic entering any water body in the first place. Some early methods of plastic monitoring include conducting impact assessments by analyzing the stomach contents of deceased marine animals, conducting beach surveys, and implementing large-scale debris collection efforts in coastal regions [6]. These methods can indicate how much plastic is disposed on these coastal areas, but are not very effective in detecting the changes in the total amount of offshore debris [1]. Observations from ships, made both visually and by net trawls, were also used in more recent studies [7]. The data gathered using these methods can be useful to a certain extent, but becomes hard to interpret when the complex aquatic dynamics are considered [1]. Furthermore, the efficacy of these efforts can only be measured if the plastic pollution can be monitored in target areas. This is a challenging task due to the complexity of temporal and spatial distributions of debris as much as the marine dynamics [1].

Recent developments in remote sensing point to the promising approach of using satellite data to monitor marine plastic debris. New Earth observation satellites such as Sentinel-2 started to raise the bar in how much satellite data can be utilized thanks to improvements in their temporal and spatial resolutions [8]. Today, remote sensing is one of the most promising ways forward to monitor marine plastic pollution around the globe and make use of this data for clean-up efforts or mitigation policies. However, considering the vast amount of data this approach brings with it, automated methods are necessary to detect plastics using spectral and spatial data on target locations. Deep learning methods have shown potential for such complex tasks, especially when combined with transfer learning techniques [9]. The progress on the topic so far suggests that a combination of remote sensing and machine learning techniques can provide a general method that can be used on a global scale [10, 11]. This is why a literature study was carried out on this topic and led to the thesis research presented in this report.

This thesis report has the following structure. First, the most relevant parts of the literature study performed before the research phase will be presented in chapter 2. The goal of including this chapter in this thesis report is to demonstrate the identified knowledge gap and how this thesis research helps close it. The research description will be presented in chapter 3 which will explain the main objective of the research and state the research questions investigated in this thesis. These two chapters will serve as introductory information before the research paper written for this thesis study is presented in chapter 4. This chapter will follow the format of a conventional journal article and will be independent of the rest of the report. Finally, chapter 5 will present how the results of this study answer the research questions, and finalize this thesis report.

2
Literature Study

This chapter offers an in-depth analysis of the most recent research on important areas of marine plastic detection. The first section looks at the earliest studies that used multi-spectral indices and simple machine learning classifiers for debris detection. The adoption of deep learning models then comes into focus, which has significantly improved the abilities of plastic detection systems. The chapter then explores the newly emerging field of active few-shot meta-learning as a promising direction for improving these systems. The chapter ends by outlining a collection of discovered knowledge gaps in the field. By providing this overview, this chapter aims to provide useful insights into the development of marine plastic detection research over time, as well as the current state of the art and potential future directions.

2.1. Feature Engineering

In the early years of plastic detection research, pioneering studies focused on utilizing multi-spectral indices as features [12]. Spectral indices are developed using the fact that every matter reflects a different amount of energy at different wavelengths of the electromagnetic (EM) spectrum. This is visualized in Figure 2.1, where each line represents the unique spectral signature of a material. A spectral index represents the relationship between reflectance in various spectral bands, in other words: reflectance at different points of a material's signature.

Spectral indices used in plastic detection research included ones that have been utilized in similar tasks already for many years such as Normalized Difference Vegetation Index (NDVI), and also some new indices created specifically for the detection of plastics such as Floating Debris Index (FDI) and Plastic Index (PI) [8, 13]. NDVI leverages the difference between near-infrared and visible light reflectance to identify areas with vegetation or other objects [14, 15]. FDI and PI, specifically designed for detecting floating marine debris, exploit the unique spectral properties of plastic materials [8, 13]. Studies using these indices demonstrated the potential of using specific indices as indicators for plastic presence.

To leverage the extracted features effectively, researchers employed machine learning classifiers such as Naive Bayes (NB), Support Vector Machine (SVM) and Random Forest (RF) [12, 16]. These classifiers could utilize the selected features to distinguish between floating marine debris and other objects present in the satellite images. Naive Bayes algorithms make probabilistic assumptions based on feature independence, while SVMs act as an "hyperplane" that serves to distinguish observations belonging to one class from another based on patterns of features, and Random Forest utilize an ensemble of decision trees for classification [17–19].

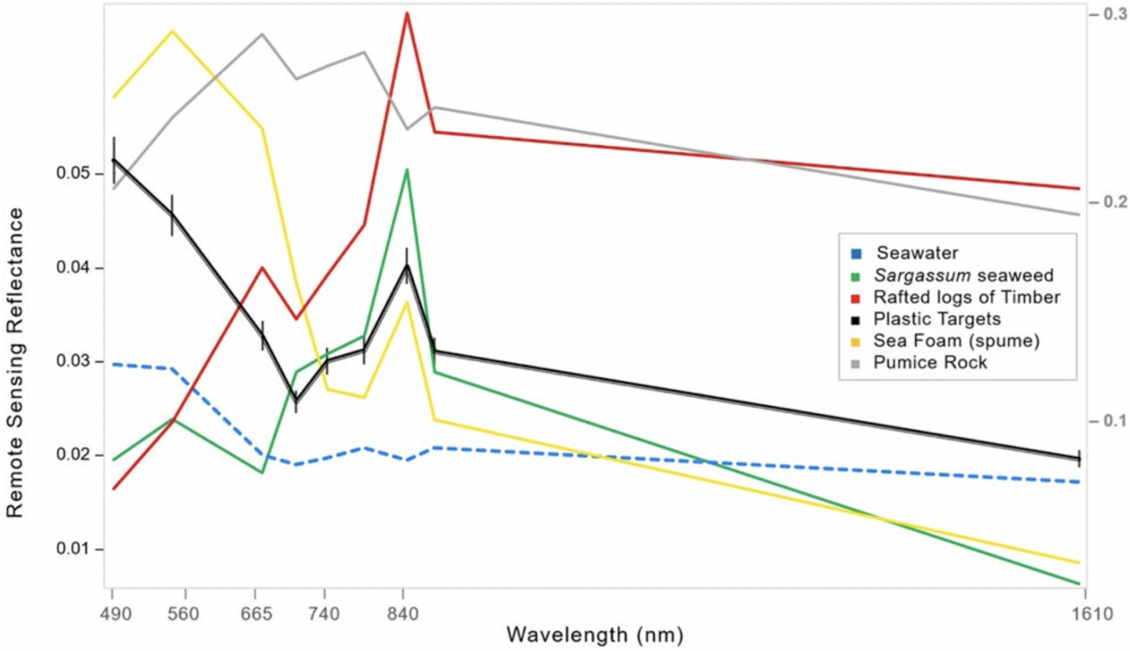


Figure 2.1: Spectral signatures for six different materials including plastics. The left-hand y-axis displays remote sensing reflectance, while the x-axis displays the range of Sentinel-2's Multi-Spectral Instrument bands. The corresponding reflectance of timber and pumice are presented on the right-hand y-axis in grey [8].

The studies using simple classifiers with spectral indices as input features employed similar procedures. The approach of the study from Biermann et al. is a good example for the general characteristics of these studies [8]. Their procedure starts with exploring the available data which consists of finding the right subset of data for the chosen study goals. Then, the next step is processing the data which consists of atmospheric correction followed by applying the spectral indices and extracting the features from the chosen data. The final step is classifying the data using these extracted features and feeding them into a simple classifier as input. This procedure is visualized in Figure 2.2.

The evaluation of early studies involved using accuracy as the evaluation metric to measure the effectiveness of plastic detection algorithms. One study by Mifdal et al. compared several simple classifiers' performance when they use NDVI and FDI indices as input [16]. The performance of these classifiers can be seen in Table 2.1. These results demonstrate significant limitations, possibly due to the reliance on manually engineered features which pose challenges in capturing the full complexity and variability of plastic materials [16]. The process of manually choosing suitable features frequently involved subjectivity and called for domain knowledge. Furthermore, the fixed nature of engineered features constrained their ability to adapt to shifting environmental factors or different kinds of plastic debris. These issues demonstrated the need for more sophisticated methods that could automatically discover discriminative features and address the shortcomings of manual feature engineering.

Table 2.1: Accuracy of different classifiers applied on the same data to detect floating plastics, where the input of these classifiers were NDVI and FDI features [16].

Method	Input	Accuracy (%)
SVM	NDVI + FDI	58.82
RF	NDVI + FDI	58.83
NB	NDVI + FDI	60.81

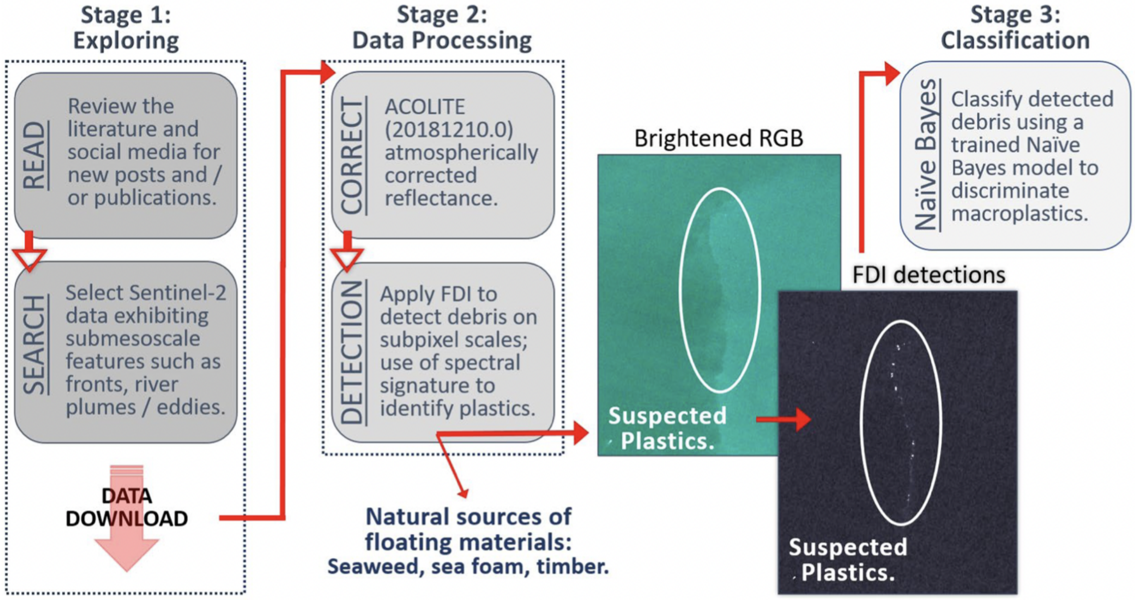

Stage 1:
Exploring
READ
Review the literature and social media for new posts and / or publications.
SEARCH
Select Sentinel-2 data exhibiting submesoscale features such as fronts, river plumes / eddies.
DATA DOWNLOAD
Stage 2:
Data Processing
CORRECT
ACOLITE (20181210.0) atmospherically corrected reflectance.
DETECTION
Apply FDI to detect debris on subpixel scales; use of spectral signature to identify plastics.
Natural sources of floating materials:
Seaweed, sea foam, timber.
Brightened RGB
Suspected Plastics.
FDI detections
Suspected Plastics.
Stage 3:
Classification
Naïve Bayes
Classify detected debris using a trained Naïve Bayes model to discriminate macroplastics.

Figure 2.2: Flowchart depicting and summarizing the actions required for detecting, identifying, and categorizing floating debris in Sentinel-2 imagery using FDI and a Naive Bayes classifier [8].

2.2. Feature Learning

Feature learning represents a significant advancement in the detection of floating marine debris by leveraging deep learning models to automatically extract features from satellite images. Unlike traditional feature engineering, which relies on manually selecting relevant features, feature learning algorithms can learn and adapt to extract discriminative features directly from the data [20]. Figure 2.3 shows an example of how a deep learning algorithm learns features, and how these features get more complex with more layers. The figure demonstrates that each layer recognizes a feature of the image such as edges or parts. Then, using these features, the algorithm classifies the image between three classes: car, person, and animal.

Convolutional Neural Networks (CNNs) have revolutionized image analysis and feature extraction [22]. CNNs are particularly well-suited for detecting patterns and extracting relevant features from images. They consist of multiple layers, including convolutional layers that learn spatial hierarchies of features [23]. CNNs have been adopted in plastic detection research as well due to their ability to automatically learn and represent complex patterns present in satellite images.

In the same study by Mifdal et al. which was previously mentioned, a U-Net CNN model which takes the whole multi-spectral data as input was also implemented [16]. The accuracy comparison of this U-Net with the simple classifiers can be seen in Table 2.2. These results suggest that the CNN model performs better than the other classifiers in terms of accuracy. Similar results were achieved by Kikaki et al. in their study which compared a Random Forest classifier to the U-Net model [24]. On the other hand, Mifdal et al. also found out that the U-Net model overfits the data more than the simple classifiers, resulting in a poor generalization performance. The conclusion of the study was that this method alone was not sufficient for marine plastic detection with the limited amount of training data that is currently available. A similar conclusion was made by another study conducted by Sóle Gómez et al. where two different CNN models were compared [25].

Carmo et al. [26] presented that a possible solution to poor generalization, especially in the case of domain shifts, is self-supervised learning which could potentially perform better with less training data. Self-supervised learning methods can learn visual features without using any labelled training data [27]. This approach was developed due to the expensive and time-consuming nature of the data labelling process. However, after testing self-supervised learning on various Earth observation tasks, Rußwurm

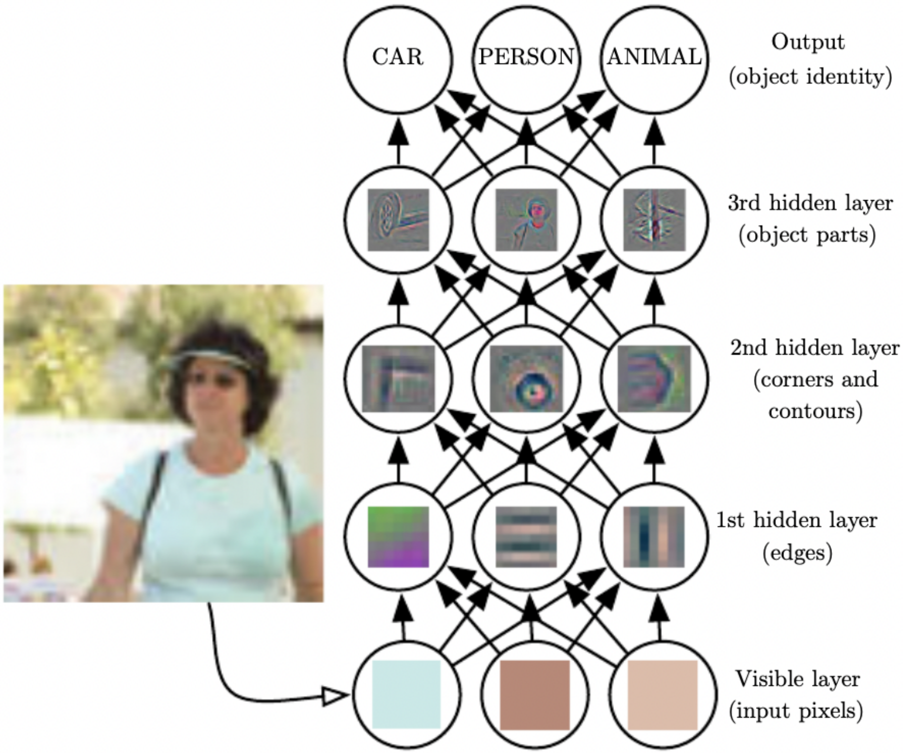


Figure 2.3: How an example neural network for classification learns different features at each layer [21].

Table 2.2: Accuracy of different classifiers applied on the same data to detect floating plastics with their specific inputs specified [16].

Method	Input	Accuracy (%)
SVM	NDVI + FDI	58.82
RF	NDVI + FDI	58.83
NB	NDVI + FDI	60.81
CNN	multi-spectral data	84.28

et al. [9] concluded that it is still not a sufficient solution for poor generalization. This study suggested that a meta-learning framework could solve this problem better than self-supervised learning. The framework they developed called "Meta-learning to address diverse Earth observation problems across resolutions" (METEOR) outperformed self-supervised algorithms and supported their claim.

The ability of few shot learning models such as METEOR to be able to classify large data sets after being trained on only a few samples is their main advantage. However, as the number of training samples decreases, the significance of each training sample increases [28]. In the floating marine detection context, this sginficiance becomes even more important due to the uneven class distirbution. According to Rußwurm et al., only about 0.05% of the pixels contain marine debris on an average Sentinel-2 image used in their study for detecting floating plastics [29]. This severe class imbalance is a challenge for any object detection model, and for a few-shot learning model, makes it crucial for the support set to have enough representation from each class to perform well. Previous research has demonstrated that including expertise of some users in the support set selection by developing a human-in-the-loop sampling strategy increases the performance of few-shot models by optimizing the support set [30].

In conclusion, the state-of-the-art feature learning approaches have significantly advanced the detection of floating marine debris. By employing deep learning models, researchers have achieved promising results in automatically extracting relevant features from satellite images. However, the limited availability of training data poses the biggest challenge since deep learning models depend on a rigorous training procedure in order to perform well. One study suggested few-shot meta-learning as the solution to this challenge, but the problem of the support-set selection remains unresolved. The subsequent section will explore few-shot meta-learning and active learning as potential approaches to further understand the current state of research in these fields. This will help evaluate if they could be used for the floating marine debris detection task.

2.3. Active Few-Shot Meta-Learning

This section delves into the domain of active few-shot meta-learning which combines the concepts of few-shot learning, meta-learning, and active learning. This new domain represents a promising new approach to potentially develop efficient and adaptable plastic detection systems. The section will be presented in two distinct phases: Few-Shot Meta-Learning and Active Learning.

2.3.1. Few-Shot Meta-Learning

Few-shot learning is a rapidly evolving field in machine learning that addresses the challenge of training models with limited labeled data [31]. Few-shot learning aims to overcome this limitation by enabling models to generalize and classify new instances based on only a few labeled examples. Meta-learning, on the other hand, focuses on acquiring knowledge from multiple tasks or datasets to facilitate rapid adaptation and generalization to new tasks or datasets [32]. In the context of few-shot learning, meta-learning algorithms aim to train models that can quickly adapt to new instances with only a small number of labeled examples. By effectively capturing and generalizing from the underlying patterns and characteristics of data, few-shot meta-learning models exhibit impressive adaptability and can classify new instances accurately.

In the few-shot meta-learning context, a task refers to a specific learning problem or classification problem. Each task consists of a small set of labelled samples called a support set, and a query set which contains unlabeled examples that need to be classified [9, 32]. The support set provides the model with the necessary information to learn and generalize to the query set.

Model-Agnostic Meta-Learning (MAML) is a popular few-shot meta-learning algorithm [31, 32]. MAML works by training a model's initial parameters on a variety of tasks in such a way that it can quickly adapt to new tasks with minimal data. The model is trained to find an initialization that can be fine-tuned efficiently using a small support set, leading to improved performance on the query set. During the training process, MAML optimizes the model's initial parameters to minimize the task-specific loss after a few gradient updates on the support set. This enables the model to learn generic representations and adapt them quickly to new tasks with limited data. By leveraging the shared knowledge from multiple tasks, MAML enhances the model's ability to generalize and make accurate predictions on unseen examples.

Only one few-shot meta-learning model was ever tested for floating marine debris detection: METEOR. As previously introduced, this model was developed to be used for various Earth observation problems across various resolutions [9]. This model showed promising results in terms of floating debris detection capability. Results of this study showed that METEOR outperformed competitor and baseline methods. However, this study did not analyse the impact of the support set selection on the model's performance. Using a few-shot meta-learning approach such as METEOR for floating plastic detection seem promising considering their ability to adapt to new tasks only with a few training samples. Yet, it should still be tested how the selection of the support set impacts the plastic detection performance of such a model.

2.3.2. Active Learning

Active learning is a technique that aims to optimize the annotation process by selecting the most informative samples for labeling [33]. In the context of detecting floating marine debris, where acquiring labeled data can be time-consuming and expensive, active learning offers a valuable approach. By actively selecting samples that are likely to improve the model's performance, active learning reduces the annotation effort and enhances the efficiency of plastic detection systems.

Various strategies and techniques have been employed in the field of active learning to enhance the annotation process and improve model performance. These strategies aim to intelligently select samples for annotation that would provide the most valuable information to the learning algorithm. Some commonly used strategies will be presented below.

Uncertainty-Based Sampling:

Uncertainty-based sampling is a widely adopted strategy in active learning [34]. It involves selecting samples for annotation that are associated with high uncertainty in model predictions. This can be achieved by measuring the entropy of the model's predicted probability distribution or using other uncertainty estimation methods. By focusing on samples that the model finds most challenging or uncertain, uncertainty-based sampling aims to refine the model's understanding of the decision boundary and improve overall accuracy.

Query-by-Committee (QBC) is another popular active learning strategy that falls under uncertainty-based sampling [34]. It involves training an ensemble of models with different initializations or variations in the training data. These models form a committee that represents different hypotheses about the unlabeled data. The committee members then "vote" on which samples to query for annotation. Disagreement among the committee members indicates uncertainty in the predictions and prompts the selection of samples for annotation. QBC aims to capture diverse perspectives and address model uncertainty by selecting samples where the committee shows the most disagreement.

Diversity-Based Sampling:

Diversity-based active learning methods play a crucial role in selecting informative samples that cover a wide range of instances and improve the generalization capabilities of machine learning models [35]. These methods aim to ensure that the selected samples are diverse enough that they can represent the whole dataset. One example of a diversity-based active learning method is cluster-based sampling.

Cluster-based sampling involves identifying clusters or groups of similar instances in the dataset and selecting representative samples from each cluster. By considering the diversity among clusters, cluster-based sampling ensures that the selected samples span a wide range of variations present in the data [36]. This approach enhances the model's ability to generalize effectively by capturing the variability and distribution of instances [34]. Cluster-based sampling techniques often employ clustering algorithms, such as k-means, to group similar instances together [36]. The representative samples selected from each cluster contribute to a more comprehensive understanding of the dataset and can help address the bias that may arise from focusing on specific regions or instances. By incorporating cluster-based sampling into the active learning process, models can benefit from diverse perspectives and achieve better coverage of the data space, ultimately leading to improved performance in tasks such as plastic detection.

These strategies represent a subset of the wide range of techniques used in active learning. Researchers continue to explore and develop new approaches that suit specific domains and tasks. The choice of strategy depends on the characteristics of the data, the learning algorithm employed, and the available resources.

The combination of few-shot meta-learning and active learning offers a promising approach for floating marine debris detection. By leveraging the strengths of these methodologies, this innovative approach addresses the challenges of limited labeled data. Few-shot meta-learning enables quick adaptation and classification with minimal labeled examples, while active learning enhances the annotation process by selecting informative samples. This synergy enhances efficiency, and generalization capabilities, contributing to effective detection and monitoring of floating marine debris, as well as supporting environmental preservation.

2.4. Knowledge Gap

As mentioned in the previous section, the marine debris detection is a new field of research and this also means there is a large knowledge gap to be covered by new studies. This also means that there are various possible directions for research, and any future study should carefully select focus points to ensure valuable results. This section will identify the most important parts of the knowledge gap within the field and clarify the choice of research objective for this thesis presented in the next chapter.

G1. **Limited availability of training data:** The one challenge mentioned by all studies is the limited availability of training data. Especially since floating debris detection is a niche topic, there is only a limited amount of available data. Performance of machine learning algorithms heavily depends on the size and the quality of the training data set. As it is very expensive and time-consuming to collect more data on marine floating debris, the best approach would be using detection methods which do not require a large amount of training data to perform well. One such approach is few-shot meta-learning which makes it possible to train a model on a large set of data from another application and then fine-tune and apply it to floating debris detection with only a few labelled samples. This approach has only been tested once for plastic detection so far. Experimenting with the use of few-shot meta-learning on floating debris detection would therefore produce valuable knowledge for the scientific community.

G2. **Active few-shot meta-learning being unexplored:** As mentioned in the previous point, few-shot meta-learning is a promising solution to limited availability of training data. However, few-shot meta-learning models are expected to suffer from a low-performance support set if the samples are not selected to optimize for diversity and representativeness. The literature study suggested that active learning could ensure that informative samples are selected for the support set. However, there is no research on using active few-shot meta-learning for plastic detection. Even in other fields, this is a very novel topic and therefore any new study is expected to benefit the scientific community.

G3. **Informative sampling selection challenge:** Active learning seeks to identify the most informative samples from a vast dataset to optimize model learning with minimal data. However, the precise definition of "informative" and the selection of appropriate active learning methods for specific applications remain uncertain. Consequently, when it comes to detecting floating debris using active few-shot learning, the suitability of various active learning strategies remains unclear. Therefore, it is crucial to explore different methods extensively before drawing any conclusions regarding the effectiveness of this approach.

G4. **Unexplored performance on realistic class imbalance:** The proportion of pixels in satellite images of coastal areas with plastic debris is noticeably low compared to the vast majority of pixels that represent land or water. Due to their tendency to favor the dominant classes, machine learning models may find it difficult to recognize and classify the minority class of plastic debris because of this inherent class imbalance. As a result, little is known about how well these models perform under a realistic class distribution. The fact that this aspect has not been the subject of in-depth research suggests that there is a significant knowledge gap. Closing this knowledge gap will aid in the creation of more effective plastic detection systems.

G5. **Poor generalization ability of existing models:** Another challenge often mentioned by previous studies is the poor generalization ability of models. This mainly has to do with the limited availability of training data with large spatial coverage. An ideal model should be able to detect plastics at any location and also at different times at the same location. Future research should develop advanced methods to improve generalization abilities of detection models and test their generalization performance.

Analysis of the identified knowledge gaps has provided insightful information about current constraints and areas that need more research. The research questions and an explanation of how they aim to address and close the aforementioned gaps will be presented in the following chapter.

3 Research Description

Considering the knowledge gaps which were identified in the previous chapter, a new research has been proposed. The primary objective of this research is *to evaluate the efficacy of an active learning approach coupled with a few-shot meta-learning model for the detection of floating marine debris*. By incorporating active learning, the most informative samples for annotation will be selected, maximizing the learning gain of the model within the limitations of available resources.

The main research question is:

> **How can informative training images be selected for a few-shot meta-learning model to detect floating marine debris patches on publicly available satellite data?**

The main research question can be further divided into the following sub-questions. The answers to these sub-questions will combine to provide an answer to the main research question.

SQ-1. How does a few-shot meta-learning model perform compared to a conventional deep neural network for marine debris detection?

SQ-2. How do uncertainty based and diversity based active learning methods compare for this application?

SQ-3. How do the sampling strategies perform when there is a more realistic class imbalance?

Answering these research questions will attempt to close each knowledge gap identified in the previous chapter in the following manner:

G1. **Limited availability of training data:** Using few-shot meta-learning and active learning together reduces the number of labelled samples needed for training as much as possible. By selecting the best samples for the model to learn how to distinguish plastics on satellite images, it is ensured that the model will reach an acceptable performance only with a small amount of training data. This would address the problem of needing an extensive data collection and labelling campaign for floating debris detection. With a well-developed few-shot meta-learning model, only a few labelled samples can be sufficient. All these mean that answering the main research question directly addresses this knowledge gap.

G2. **Active few-shot meta-learning being unexplored:** This study will close this knowledge gap by answering the first sub-question (SQ-1). The comparison between the active few-shot meta-learning model and a conventional deep learning model will provide insights on how such a model performs compared to a well-known model. The outcomes of this study will provide new information on how active learning and few-shot learning can be combined.

G3. **Informative sampling selection challenge:** Different active learning methods will be compared as a part of this study to answer the second sub-question (SQ-2). By comparing various active learning methods, this study will explore which strategies work better with the few-shot learning model and the data at hand. The performance of these different methods will provide information on how well these selection strategies can pick informative samples.

G4. **Unexplored performance on realistic class imbalance:** Answering the sub-question 3 (SQ-3) will provide information on how a more realistic class imbalance would impact the performance of an active few-shot meta-learning model. By comparing the performance results in a more balanced setting to an imbalanced setting will demonstrate the difference and reveal the possible impacts of the severe class imbalance on floating marine debris detection field.

G5. **Poor generalization ability of existing models:** While the primary focus of this study does not specifically evaluate generalization performance in-depth, it indirectly addresses this knowledge gap. The ability to effectively fine-tune a classification model using only a limited number of labeled samples and subsequently apply it to a new region would effectively mitigate the generalization challenges faced by debris detection systems. If the model can be easily fine-tuned for any region with only a few labeled samples, the issue of generalization would no longer present a significant challenge.

The study and its results that answer these research questions will be presented in the next chapter.

4

Research Article

This page is intentionally left blank.

Effective support set selection for few-shot detection of floating marine debris

Dilge Gül, Devis Tuia, Jurgen Vanhamel, Marc Rußwurm

(Delft University of Technology, Delft, Netherlands
E.D.Gul@student.tudelft.nl)

Abstract: Marine litter, particularly plastic debris, poses a significant environmental challenge globally. Detecting floating debris in the marine environment using satellite remote sensing remains a complex task due to the limited availability of high-resolution data and the coarseness of existing datasets. This study explores the potential of active few-shot meta-learning for improving the detection of marine debris. The results demonstrate that active learning methods incorporating uncertainty-based sampling, such as entropy and query by committee, outperform other strategies in terms of recall and average precision. Yet, diversity-based methods are found to be limited by the poor representativeness of the feature space used for clustering samples. Additionally, the study highlights the influence of regional characteristics on detection performance and the impact of class imbalance on active learning strategies. To further enhance marine debris detection, future research directions are identified, including training meta-models specifically on marine debris data and tuning decision thresholds. The suggested methodology shows promise for enhancing the efficiency of remote sensing-based monitoring of marine debris, thereby assisting environmental management and conservation efforts.

Keywords: *marine plastic pollution, floating debris, remote sensing, multi-spectral detection, deep learning, meta-learning, few-shot learning, active learning*

4.1. Introduction

Marine litter is a growing problem caused mainly by human-created trash, and significant amounts of plastic can be found in the oceans due to the unfiltered discharge of waste into rivers, poor waste management, or lost fishing nets. According to the United Nations Environment Program, about 70% of marine litter sinks to the ocean floor, while the remaining, which mainly consists of plastic, floats on the surface of water bodies and can be aggregated by processes such as river plumes, windrows, oceanic fronts, or currents [16].

Research on macro-debris detection is recent, but studies on plastic detection using airborne data, models, and theoretical studies have demonstrated the potential to detect macro-plastics in optical data [12]. Satellite remote sensing is the leading technique for collecting high-quality, standardized optical imagery on global scales, but few studies have succeeded in detecting floating macro-plastics in the marine environment due to temporal, spatial, and spectral coarseness of available data [8]. The currently available datasets are relatively limited in number and do not usually use open-access high-resolution satellite data over geographically extended areas. These facts limit the utilization of satellite data to detect marine debris by machine learning frameworks [24].

Hand-engineered spectral characteristics and basic classifiers on satellite photos were used in early attempts to combine remote sensing and machine learning to detect floating marine debris [12]. Then, the next step was to use the full multi-spectral data from satellites together with deep learning techniques. However, these techniques require a significant amount of training data which is difficult to get for marine debris. Few-shot meta-learning has emerged as a promising approach for addressing this issue as it enables the development of models that can effectively learn from a limited number of labeled examples [9]. Few-shot meta-learning algorithms can quickly adapt to new scenarios by selecting an effective support set, making them an ideal choice for marine debris detection applications. In this article, we investigate the use of effective support set selection for few-shot detection of floating marine debris, demonstrating the efficacy of this approach on a real-world dataset.

Marine debris detection can benefit from the human-in-the-loop concept, such as active learning, in addition to few-shot meta-learning. By incorporating human-in-the-loop approaches, the detection process can leverage the expertise and knowledge of human annotators. With active learning, the model can perform well with fewer labeled samples by iteratively choosing the most informative samples for fine-tuning [37]. In the case of detecting marine debris, where there is a lack of labeled data, active learning can significantly lessen the labeling effort while enhancing the model's performance. The few-shot meta-learning framework can be furthered by incorporating active learning to increase the effectiveness of the detection procedure [30]. In order to improve the effectiveness of machine learning algorithms in the detection of marine debris, this research analyzes the possibility of human-in-the-loop approaches in addition to addressing the issue of limited labeled data.

4.2. Related Works

There are two main fields that are relevant to this paper's research topic: few-shot meta-learning and active learning. These two fields are relevant since the goal of this research is to analyse how the support set of a few-shot meta-learning model could be selected effectively using active learning methods. Below, some background information on these two topics is provided.

Few-shot meta-learning: While machine learning has been successfully used for many applications, it often fails when the amount of training data is limited. Few-shot meta-learning has recently been developed to tackle this problem [31]. The goal of few-shot meta-learning is to mimic the ability of humans to adapt to new concepts using their prior knowledge [38]. In the context of few-shot detection of floating marine debris, few-shot meta-learning can be used to quickly adapt to new detection tasks with limited labeled data. To apply few-shot meta-learning, a meta-model is trained on a large set of tasks with few labeled examples or one task with large amount of training data first [9]. Then, a task-model is created by fine-tuning the meta-model for the desired task using a few labeled examples as a support set. The model is able to generalize to the new detection task by learning the underlying patterns or "meta-knowledge" from the initial training step. The quality of the few labeled samples in the support set has a significant impact on the performance of the model [28].

Active learning: Another approach to machine learning that seeks to improve the efficiency of the learning process is selecting the informative samples for the training of models. This approach is called active learning, and it introduces the human-in-the-loop concept as a human annotator is actively labelling new samples and feeding them to the model as the training progresses [30]. Two active learning categories are uncertainty-based and diversity-based learning. Uncertainty-based methods involve selecting samples with high prediction uncertainty, while diversity-based methods focus on maximizing diversity in the selected samples [28]. In the context of few-shot detection of floating marine debris, active learning can be used to further reduce the number of support set samples required for fine-tuning of the meta-model. At any point during fine-tuning, the support set can be enriched by samples which are actively selected and this process can be repeated iteratively to improve the model's accuracy. Ideally, the active selection will ensure the selection of most effective support set samples earlier in the process and will always outperform a randomly selected support set with the same size. Active learning has shown promising results in various computer vision tasks, and can be combined with few-shot meta-learning to further improve the detection accuracy while minimizing the labeling cost [39].

Despite the potential of active few-shot learning in floating marine debris detection, its application in this field has been largely unexplored. Both active learning and few-shot learning techniques have yet to be utilized for the specific task of detecting marine debris. Initial studies in the field relied on hand-engineered spectral features and simple classifiers, which were constrained by the limited informativeness of these manually crafted features. Subsequently, researchers introduced deep learning models to extract more complex features for marine debris detection. However, this approach heavily relied on the availability of extensive training data, which is often lacking in the context of floating marine debris. The scarcity of labeled training data poses a significant challenge and calls for innovative approaches, such as active few-shot learning, to address the limitations and enhance the effectiveness of floating marine debris detection.

4.3. Materials

This section will present the materials used in this study, namely the data and the models. This study uses the *RefinedFloatingObjects* data archive as floating marine debris data [29], which will be presented in more detail in subsection 4.3.1. The chosen few-shot deep learning methods for this research is the METEOR ("a **MET**a-learning framework for **E**arth **O**bservation problems across different **R**esolutions") [9]. A ResNet-18 trained for this application in a fully supervised way is going to be used as the comparison model since it provides a high accuracy example. More information on these models will be presented in subsection 4.3.2.

4.3.1. Data

There are six datasets from six distinct regions used as part of this research. These regions are Lagos (Nigeria), Marmara (Turkey), Venice (Italy), New Orleans (United States), Accra (Ghana), and Durban (South Africa). These datasets are from the *RefinedFloatingObjects* archive created by Rußwurm et al. [29]. Part of this archive is created by re-annotating a subset of the *FloatingObjects* archive created by Mifdal et al. [16] in Google Earth Engine (GEE) to reduce label noise. This part consists of the regions Lagos, Venice, New Orleans and Accra. The other two regions, Marmara and Durban are added for the validation of the model developed by the study of Rußwurm et al. [29]. The selection of all six regions was done by exploring news items or social media posts that pointed to existing floating marine debris on the sea surface. The same applies for the dates and time the data was retrieved since the existence of debris depends on the dynamic motions of ocean currents, and the location or the shape of the debris can change over time.

This study will use the same approach as to how the scenes in the *RefinedFloatingObjects* archive are sampled for training or validation. Using this approach, 128 px x 128 px patches centered on each annotated point will be extracted, each labelled either as marine debris (class 1) or other/non-debris (class 0). To provide their model with diverse set of non-debris examples, which are anything but marine debris, the class 0 contains various objects or materials such as water, land, coastline, and ships [29]. It is important to note that the nature of debris is different in each dataset. Marine debris refers to any floating object on the surface of sea water, but the floating objects can still be anything from natural debris to plastic litter. This brings significant diversity to the characteristics of pixels labelled as marine debris.

The Sentinel-2 imagery retrieved and labeled for the creation of the *RefinedFloatingObjects* archive includes two different formats: L1C (top-of-atmosphere) and L2A (bottom-of-atmosphere). L2A data undergoes atmospheric correction using the Sen2Cor processor, reducing noise and improving quality [40]. Whenever available, L2A data was used due to its higher quality. However, for the Marmara and Accra regions where L2A data was not published, L1C data was used. Preliminary tests demonstrated that combining L1C and L2A data did not significantly impact model performance. Hence, the *RefinedFloatingObjects* data is used as it is. However, it is recommended to employ specialized processing with atmospherically corrected data for optimal model performance if resources and time allow.

The Sentinel-2 Multi Spectral Instrument (MSI) consists of 13 spectral bands, ranging from visible to near infrared and short-wave infrared, at various spatial resolutions, with the highest being 10 m [41]. Therefore, the *RefinedFloatingObjects* used in this study includes these 13 spectral bands as well. However, band 10, which does not provide bottom-of-atmosphere information, is excluded from L2A data, making it possible to use L1C and L2A data together. The information from the remaining 12 spectral bands is utilized, as multi-spectral data enhances detection performance by providing detailed scene information and enabling the detection of subtle differences between debris and non-debris pixels

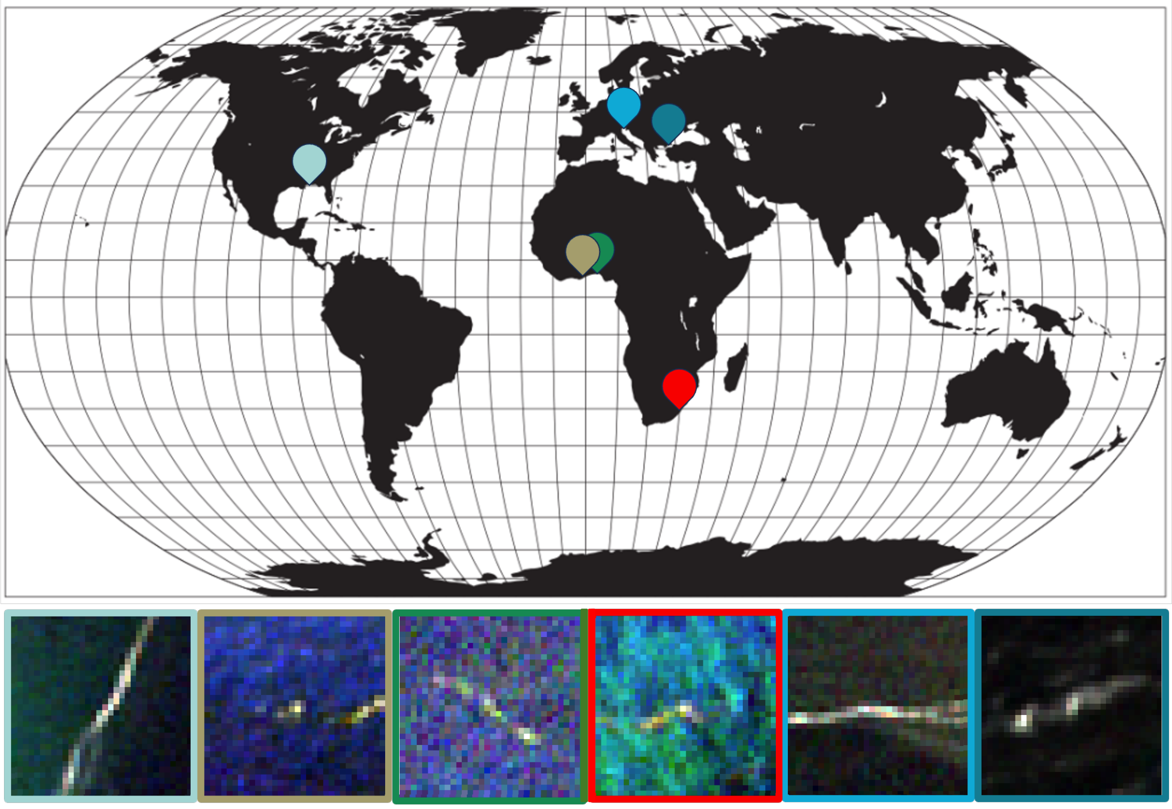


Figure 4.1: Red Green Blue (RGB) visuals of debris samples from each region. The samples are visualized using multi-spectral data, highlighting the differences in appearance attributed to the unique spectral characteristics of each location. [World map from `https://www.maptorian.com`]

[16].

The ResNet-18 model used in this study for comparison follows the same training procedure as performed by Rußwurm et al. for their study's model [29]. The four regions Lagos, Marmara, Venice and New Orleans are used to train the model while regions Accra and Durban are used for testing. The summary of all information on the six datasets presented in this section can be found in Table 4.1.

Table 4.1: Information about the datasets used in this study including the type of data, size of datasets, nature of debris at each region and whether the dataset was used for training or testing of the ResNet-18 model.

ID	Region	Type	Size	Nature of debris	ResNet-18 split
0	Lagos	L2A (corrected)	678	Plastic, pumice, and other debris	Training
1	Marmara	L1C (raw)	197	Floating algae (sea snot)	Training
2	Venice	L2A (corrected)	569	Sea foam with plastics and other debris	Training
3	New Orleans	L2A (corrected)	1029	Mostly other debris	Training
4	Accra	L1C (raw)	1506	Sea foam with pumice, plastics and other debris	Test
5	Durban	L2A (corrected)	701	Sargassum patches with entangled plastics	Test

The experiments will include testing the performance of all the methods on a more realistic class distribution as well. According to Rußwurm et al., on a Sentinel-2 image that is used in the *FloatingObjects* data archive, only about 0.05% of the pixels contain marine debris [29]. Looking at Table 4.2, it is evident that the data utilized in this study deviates significantly from a realistic class distribution. The training data for ResNet-18 contains a higher proportion of debris pixels to non-debris pixels compared to a realistic scenario, and the test regions also show an unrealistic distribution. Consequently, it is important to assess the performance of the approach developed in this study under a more realistic class distribution. While the severe class imbalance poses a challenge for any object detection model, evaluating their performance under such conditions is crucial as it can reveal specific areas of difficulty and inform future improvements.

Table 4.2: Class distribution of all training and test regions. "%Debris" represents the percentage of debris out of the total size.

ID	Region	Total Size	Debris	Non-Debris	%Debris
0	Lagos	678	336	342	49.6%
1	Marmara	197	62	135	31.5%
2	Venice	569	197	372	34.6%
3	New Orleans	1029	275	754	26.7%
	Total training set	2473	870	1603	**35.2%**
4	Accra	1506	740	766	**49.1%**
5	Durban	701	163	538	**23.3%**

4.3.2. Models

This section will present the two models that are used as part of this study: METEOR which is the few-shot meta-learning model, and ResNet-18 which is the comparison model. These models will be presented in three subsections: the first one introducing their architectures, the second one describing their pre-training procedures, and the third one explaining their fine-tuning processes.

Model Architectures

The METEOR model is based on a deep ResNet-12 neural network architecture, which has been designed to address various Earth observation problems across different sensors and geographies, including marine plastic debris detection. METEOR utilizes a model-agnostic meta-learning algorithm, which enables the extraction of meta-data from extensive land cover classification data. This extracted meta-data forms the "meta-model" of METEOR. This meta-model can effectively encode knowledge from source task and transfer it to a target task, a process referred to as model-based transfer learning. The meta-model is then fine-tuned into a task-specific model for a particular target task, such as the detection of deforestation or urban scene classification. This fine-tuning process, known as few-shot learning, requires only a few labeled images for adaptation. The overall process of METEOR is summarized in Figure 4.2.

ResNet-18 is a widely used deep neural network architecture that has shown remarkable performance in various computer vision tasks, including remote sensing and Earth observation applications [42]. However, training such deep models can be challenging because the information flowing through the layers can become very weak and difficult to learn from. To overcome this problem, ResNet-18 uses a technique called residual learning [43]. Instead of trying to directly learn the complete transformation from input to output, the model focuses on learning the difference between the input and the desired output. This difference is called the residual.

ResNet-18 includes special connections called shortcut connections to make learning easier [43]. These connections allow information to skip one or more layers and directly reach further layers. These shortcuts help the model by providing alternative paths for information to flow through the network. They ensure that important information from earlier layers is preserved and that the model can effectively learn from it. The main advantage of these shortcuts is that they prevent the information from getting too weak or disappearing altogether as it passes through many layers. By keeping the information flow strong, ResNet-18 can learn intricate patterns and achieve high accuracy even with a large number of layers. So, these shortcut connections are like shortcuts that help the model learn better and handle deep networks successfully.

The architecture of the ResNet-18 model used in this study is depicted in Figure 4.3.

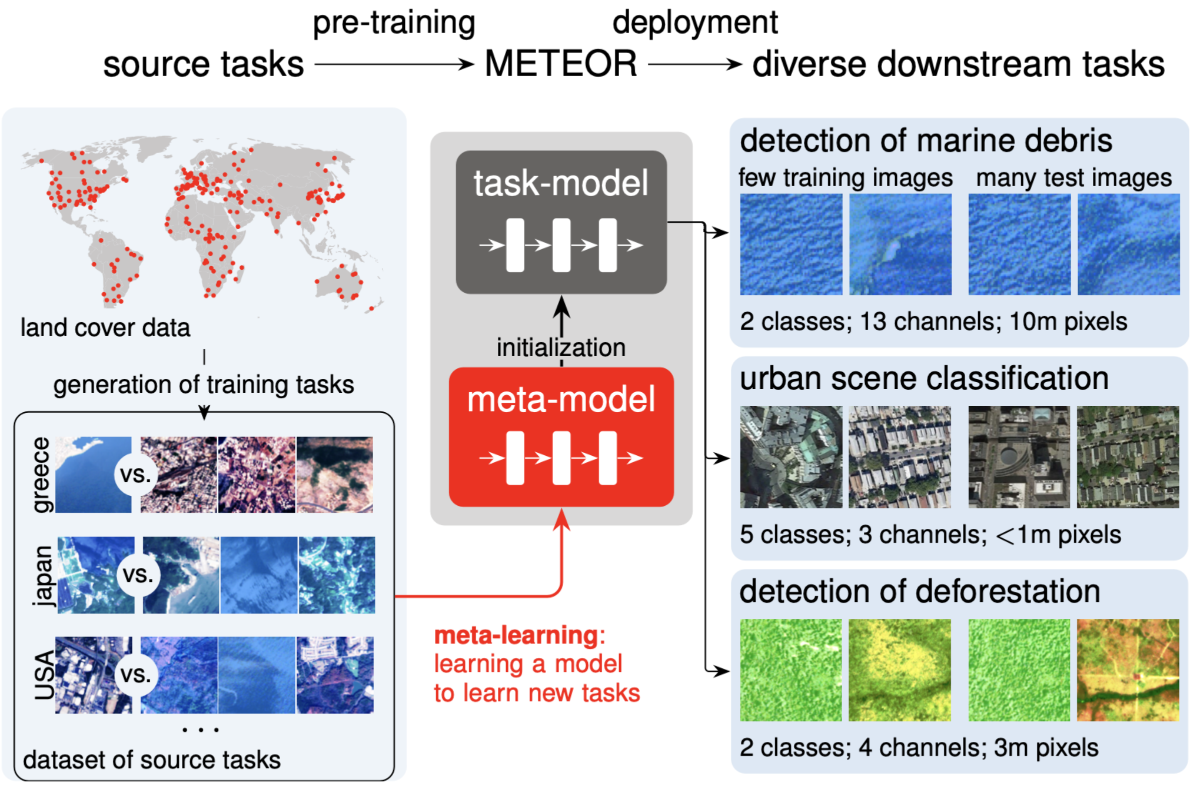

pre-training
deployment
source tasks
METEOR
diverse downstream tasks
land cover data
generation of training tasks
greece
vs.
japan
vs.
USA
vs.
...
dataset of source tasks
task-model
initialization
meta-model
meta-learning: learning a model to learn new tasks
detection of marine debris
few training images
many test images
2 classes; 13 channels; 10m pixels
urban scene classification
5 classes; 3 channels; <1m pixels
detection of deforestation
2 classes; 4 channels; 3m pixels

Figure 4.2: Concept of METEOR summarized. The left side shows the model's pre-training on land cover classification, and the right side shows the implementation of the model on other tasks [9].

Pre-Training Procedures

For the METEOR model, the pre-training process involves training the meta-learning algorithm on extensive land cover classification data. A deep ResNet-12 neural network was trained using 16 photos from the Sen12MS dataset, which contains imagery with 15 input channels representing two radar bands from Sentinel-1 and 13 spectral channels from Sentinel-2 [44]. The meta-model was pre-trained on four randomly selected land use and land cover classes from a specific geographic area [9]. The resulting meta-model captures the knowledge from the land cover classification task and serves as the basis for subsequent fine-tuning.

Unlike METEOR, the ResNet-18 model follows a conventional supervised pre-training procedure. It requires a large amount of labeled training data to perform well on a classification task. In remote sensing applications, ResNet-18 has been utilized for various tasks, including land cover classification, object detection, and semantic segmentation [14]. While ResNet-18 has demonstrated high accuracy and robustness in similar applications, it has not been directly applied to floating marine debris detection. Nevertheless, it is expected to perform well in this domain, given its track record and ability to handle changes in image quality and environmental factors [14].

Fine-Tuning Processes

Fine-tuning involves updating the model parameters using a limited number of labeled samples from the target task. For METEOR, the fine-tuning process focuses on transforming the meta-model into a task-specific model using few-shot learning techniques. The fine-tuning is facilitated by the knowledge encoded in the meta-model, allowing for effective adaptation to the debris detection task. A few labelled samples from the test set is used to fine-tune the meta-model, and then the model is used to classify the rest of the test set. On the other hand, ResNet-18 comparison model does not have a fine-tuning process

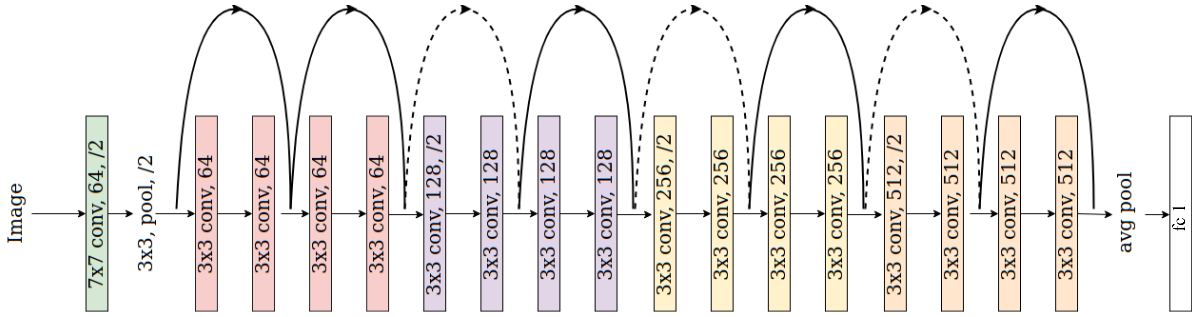


Figure 4.3: Architecture of the ResNet-18 model where (both the solid and dashed) arrows between layers represent the shortcut connections [43].

in this study. The pre-trained model is directly tested on the test samples.

Overall, in spite of their shared architecture, the METEOR and ResNet-18 models have unique qualities that should be emphasized. The different training goals of these models have a big effect on how they are supposed to be used and deployed. Additionally, each model's fine-tuning procedure differs due to their various pre-training regimens. As a result, there are significant functional and optimizational differences between the METEOR and ResNet-18 models.

4.4. Methods

In this study, the main goal is to test and compare various sampling strategies to sample the support set of METEOR. The strategies will be compared to random sampling as well as to the performance of the comparison model ResNet-18. To design a fair comparison experiment, all strategies will be tested using the same procedure. The basis of this procedure is explained by Algorithm 1

Algorithm 1 Basis of sampling procedure

```
Initialize METEOR meta-model
Select 1 random sample from each class
while support set size ≤ desired value do
    Pick 2 samples using the sampling method
    Add those 2 samples to the support set
    Fine tune the model with the extended support set
    Test on remaining samples and update evaluation metric
```

The selection of a sample represent the user labelling the samples picked by the sampling strategy to feed them into the model in the real life use case. Similarly, for the first step, the user is supposed to select one sample from each class (debris and other) and label them to create the initial one-shot support set. This selection can be made deliberately if the user has experience in selecting representative samples. However, for the experiments in this study, these two samples will be picked randomly.

There are two groups of sampling strategies that will be tested and compared in this study: uncertainty-based active learning methods and diversity-based active learning methods. The uncertainty-based methods refer to selecting samples that the model is most uncertain about their class prediction. Two common uncertainty-based sampling strategies that will be used in this report are *entropy* and *query-by-committee* methods. The diversity-based sampling strategies select samples to optimize their diversity. This study will use different feature space mappings as the representation of samples and select the ones that maximize diversity in the support set using a *clustering based* method. This section will explain the principles of these sampling strategies.

4.4.1. Uncertainty-Based Sampling

For uncertainty-based sampling methods, the first sampling step consists of measuring the uncertainty of each sample's prediction in terms of the chosen metric. Then, the most uncertainly classified samples are added to the support set and the prediction uncertainties are updated once the model is fine-tuned by the extended set and evaluated on the test set. This procedure is visualized in Figure 4.4.

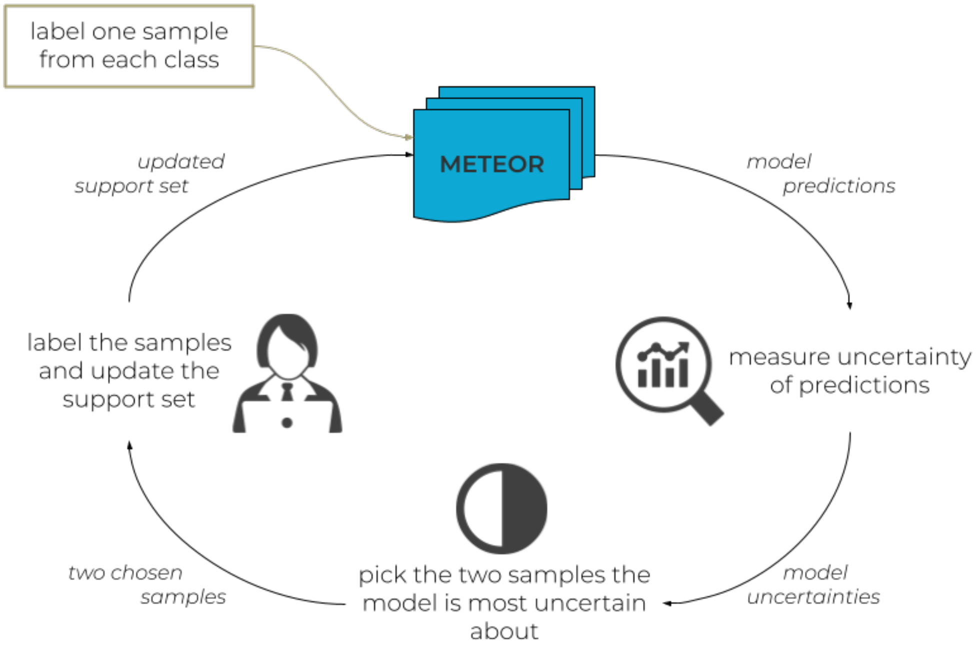

label one sample from each class
METEOR
updated support set
model predictions
label the samples and update the support set
measure uncertainty of predictions
two chosen samples
pick the two samples the model is most uncertain about
model uncertainties

Figure 4.4: The diagram showing the procedure used for uncertainty-based sampling methods.

In this study, two distinct uncertainty-based sampling methods will be tested, each employing a different metric to quantify the model's uncertainty regarding a sample's prediction: entropy and disagreement among a committee of classifiers.

Entropy Based Sampling

Entropy is a measure of uncertainty, and is the most common uncertainty metric used in uncertainty-based sampling [34]. This study defines entropy of the k^{th} sample as follows:

$$\mathbb{H}(p_k = -(p_{i,k}\log(p_{i,k}) + (1 - p_{i,k})\log(1 - p_{i,k})) \tag{4.1}$$

where $p_{i,k}$ is the confidence score predicted for class 1 and hence making $1 - p_{i,k}$ the confidence score for class 0 in the binary classification setting [45].

The entropy value will be interpreted as follows: If the entropy is close to 50%, it means that the model is very uncertain. However, if the entropy is close to 1% or 99%, it means that the model is fairly certain about its prediction.

To actively sample using the entropy metric, the value for the entropy will be calculated for each sample at every step, and the samples with the highest uncertainty will have priority when the support set is being extended.

Query-By-Committee Method

Query-by-committee method involves the use of a committee of models which use different subsets of the training data and therefore each represent a different hypothesis about the underlying data distribution [34]. For active selection of the samples, the disagreement between the committee members are used as the uncertainty metric. In this study, the disagreement is defined as the highest standard deviation of predictions made by all committee member models, where there is three of them. This metric can be formulated as follows:

$$\sigma = \sqrt{\sum_{i=1}^{N} \frac{(p_{i,k} - \overline{p_{i,k}})^2}{N}} \tag{4.2}$$

where N is the number of committee members and $p_{i,k}$ is the confidence score predicted for class 1.

The samples with the highest standard deviation between class probabilities will be selected and labelled by the user to extend the support set.

4.4.2. Diversity-Based Sampling

This study is going to employ a clustering method as the diversity-based sampling strategy. The main difference between the uncertainty-based and clustering based methods is the additional aspect of considering the relationship between different samples for diversity-based models. These methods use the location of samples in the chosen feature space to make sure the support set represents all "kinds" of samples in the training set. This adds another layer to the sampling procedure. At each step, not only the specific samples but also the clusters the samples belong to are selected. The clustering is done before the sampling starts, so each sample belongs to a cluster. The sampling strategy is then to pick samples looking at both their uncertainty in prediction and cluster label. The diversity-based sampling strategy is visualized in Figure 4.5.

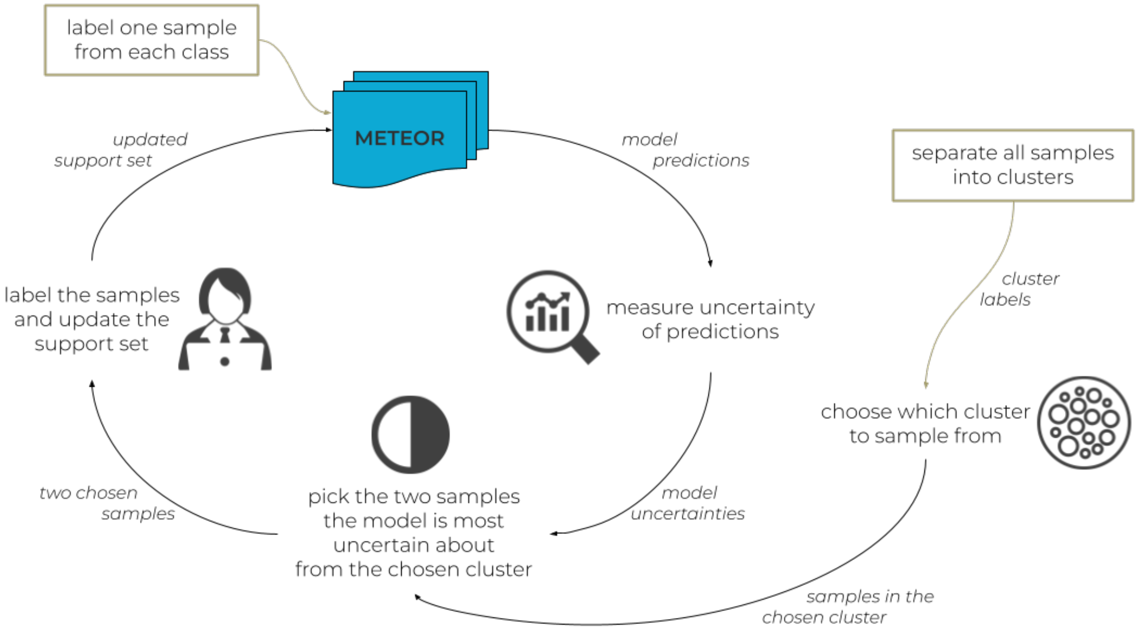

label one sample from each class
METEOR
updated support set
model predictions
separate all samples into clusters
label the samples and update the support set
measure uncertainty of predictions
cluster labels
choose which cluster to sample from
pick the two samples the model is most uncertain about from the chosen cluster
two chosen samples
model uncertainties
samples in the chosen cluster

Figure 4.5: The diagram showing the procedure used for the most advanced diversity-based sampling method which picks the most uncertain samples from the least represented cluster at each step.

Since a few-shot learning model only sees a handful of samples before making predictions on the whole dataset, it is crucial that the support set represents the whole dataset. It is expected that the model will predict more accurately if the support set is diverse and therefore can represent every part of the data distribution [35]. Cluster based sampling is a method that aims to maximize diversity among the samples in the support set. Increasing diversity increases the representativeness of samples as well [46]. Unlike uncertainty-based sampling methods, cluster based methods explore the parts of the feature space that is not in close proximity to the classification boundary. The difference can be visualized as such in Figure 4.6. In this figure, the gray area is where uncertainty-based methods are likely to explore as it is the region the model is most uncertain about [46]. diversity-based methods are able to explore the green and blue regions at the same time as both regions would increase the diversity in terms of different locations on the feature space. Furthermore, the blue region might never be explored unless diversity is considered since it does not contain any existing samples and is also far from the decision boundary. This is why diversity-based methods often outperform uncertainty-based methods [46].

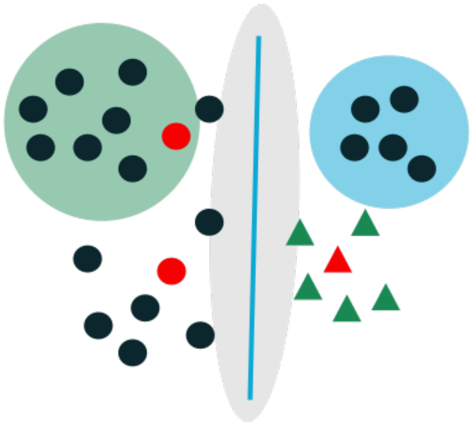
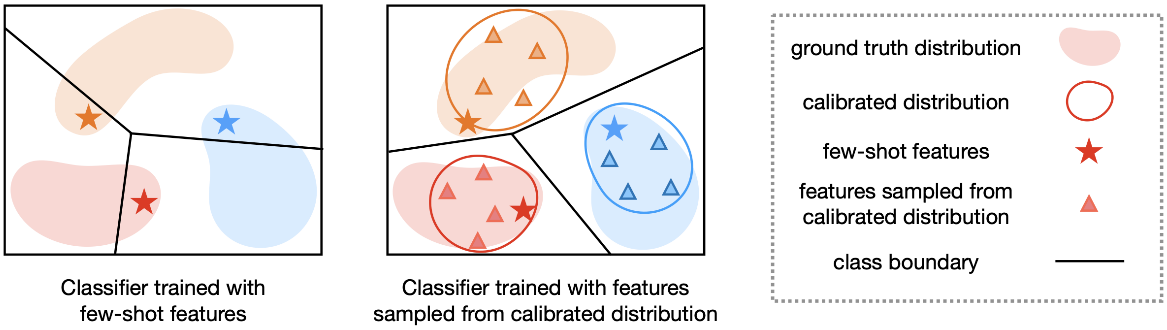


Figure 4.6: Example feature space where blue line is the classification boundary and the red samples are the labelled ones. The gray region can be explored by uncertainty-based sampling, while the green and blue regions can be explored by diversity-based sampling. The visual is inspired by [46].

Exploring a larger area of the feature space is useful for better generalization of the model. However, it also increases performance by preventing the model from overfitting [47]. Figure 4.7 shows how using clustering methods for actively sampling training data can improve classification performance. This improves the model's generalization performance as well.



Classifier trained with few-shot features

Classifier trained with features sampled from calibrated distribution

ground truth distribution
calibrated distribution
few-shot features
features sampled from calibrated distribution
class boundary

Figure 4.7: The left side shows how the model overfits when only a few samples are provided while the right side shows how the performance improves when more training points are sampled using cluster based selection [47].

In this study, the clustering based sampling will be performed as follows. First, the samples need to be clustered in terms of their representation. This representation is chosen as a randomly initialized ResNet-18 model's features. Once the features are extracted for each sample, the feature space will be divided into clusters using k-means clustering. K-means algorithm is a simple and popular clustering algorithm in machine learning [48]. These clusters are used to ensure that samples from different clusters are selected while sampling, hence increasing diversity in the support set.

This study will explore three different cluster based sampling strategies. The first one is a random selection strategy and is performed mostly as a comparison case. In this cluster sampling, first a random cluster is picked, and then a random sample within that cluster is added to the support set. This strategy is different than selecting random samples from the complete dataset since now the probability of picking any cluster is uniformly distributed.

The second cluster based method still samples a random sample from the picked cluster, but this time the clusters are picked according to how well they are represented in the support set. This strategy prioritizes selecting samples from the least seen clusters. This way, it is ensured that each cluster is represented int he support set, maximizing diversity of clusters.

The final cluster based sampling strategy again picks the most unseen cluster at each step, but to build upon that, it also considers the uncertainty of samples in each cluster. Once the least seen cluster is picked, the entropy of the samples in that cluster is used to pick the ones that the model is most uncertain about. Therefore, it is ensured not only that all clusters are represented in the support set, but also that the model learns the labels of the samples it is least confident about.

4.5. Experimental Setup

This section provides an overview of the evaluation metrics and hyper-parameter tuning employed in this study. It delves into the evaluation metrics used to measure the effectiveness of the model and the process of hyper-parameter tuning to enhance its performance. By systematically examining these aspects, the study aims to ensure robust and reliable experimentation, leading to meaningful insights and advancements in the field of floating marine debris detection.

4.5.1. Evaluation Metrics

This study will focus on comparing the models and the sampling strategies in terms of recall and average precision. The selection of the evaluation metric was done after a short study of various potential metrics. Accuracy is the most common evaluation metric to compare machine learning models. However, especially when there is a significant class imbalance in the data, accuracy can be meaningless. If a model is predicting only one class in a binary classification setting, it would still have high accuracy but the model would not actually be performing well in terms of detecting both classes. This is why it is often necessary to look into recall and precision as well, which are calculated as shown in Equation 4.3 and Equation 4.4.

$$\text{Recall} = \frac{\text{True Positives}}{\text{True Positives} + \text{False Negatives}} \tag{4.3}$$

$$\text{Precision} = \frac{\text{True Positives}}{\text{True Positives} + \text{False Positives}} \tag{4.4}$$

For this study, recall represents how much of the total debris was detected by the model while precision represents how much of what was labelled as debris by the model was actually debris. Since the goal of this study is to build towards a model which can detect floating plastics for cleanup purposes, it is prioritized that all plastics are detected. Only then, the marine environment can be plastic free. This would mean that recall is a prioritized evaluation metric. However, only focusing on recall would mean that the precision is not considered. This could lead towards a model that is predicting too many false positives. Therefore, it is important to have another metric that can be used to check the balance between recall and precision. This study will use average precision as this metric.

Average precision is equal to the area under the precision-recall curve which is constructed by recall and precision values at different thresholds [49]. By using the weighted mean of precision values achieved at different thresholds, average precision becomes independent of the specific threshold used for predictions [50]. This is a very useful characteristic especially when the distribution of classes in the dataset is very imbalanced. Marine debris detection is such an application as explained previously in this section. As a metric which summarizes the trade-off between recall and precision in one value, average precision is useful as an additional metric to recall considering the goal of this study to prioritize recall without risking too low of a precision.

Experimenting on the performance of ResNet-18 and METEOR with various sampling strategies make it possible to compare them with each other. However, none of these numbers represent what the highest achievable performance is. Therefore, a Single Shot Oracle (SSO) method, inspired by the Single Instance Oracle [37], will be included in the results. The SSO will test various additions to the support set at each step and pick the ones that increase the chosen metric, in this case the average precision, the most. The performance of SSO will represent how much performance improvement can be expected from an "ideal" sampling strategy. This will both ensure that the expectations are kept realistic, and also how well each sampling strategy is actually performing compared to what is achievable.

4.5.2. Hyper-Parameter Tuning

In order to optimize the performance of the active few-shot learning model, the following hyper-parameters were considered and tuned.

- **Number of shots:** Up to how many shots the experiments will run and the step size have been picked considering the real use case the method is being developed for. The main goal for using a few-shot meta-learning model with active sample selection methods is to keep the labeling effort to a minimum. Therefore, the upper limit to how many shots have been tested has been selected as 20. This represents a number that is low enough that labeling effort is reduced but is also large enough of a support set size to compare with lower number of shots. The step size has been chosen to be 2 since few-shot approach often increments one sample from each class while extending the support set.
- **Number of random initializations:** As introduced in section 4.5, the randomness due to different initializations were also accounted for in the experiments. To reduce the bias, each method has been run 20 times and the mean results of all the runs have been used as the output. The number 20 was selected as a result of the trade off between variance and bias.
- **Number of clusters for diversity-based methods:** For the selection of this hyper-parameter, 2 to 5 number of clusters have been tested. The final value was determined based on this comparison study based on manual experimentation and the monitoring of performance on a validation set. Even though there were no large differences in the maximum accuracy reached by models using different numbers of clusters, some of them showed faster convergence. Using 4 clusters yielded the fastest convergence speed and therefore has been chosen as the number of clusters to be used throughout the rest of this study.

4.6. Results and Discussion

In this section the results of the study will be presented and discussed. Table 4.3 showcases the performance of different sampling strategies, comparison sampling methods, and the ResNet-18 model in terms of recall and average precision on the Durban dataset.

Table 4.3: Recall and average precision of various methods on Durban region with different sizes of support sets. These results are for the original distribution of the two classes. The values represent the mean value of 20 different random initializations.

	3-shots	5-shots	10-shots	15-shots
Recall				
ResNet-18	*0.22*	*0.22*	*0.22*	*0.22*
Single Shot Oracle	0.40	0.41	0.43	0.45
Random	0.34	0.24	0.22	0.26
Entropy	**0.30**	0.22	**0.25**	0.26
Query by committee	0.25	0.23	**0.25**	0.**29**
Random clusters & random samples	0.29	**0.29**	0.21	0.19
Unseen clusters & random samples	0.23	0.18	0.23	0.21
Unseen clusters & uncertain samples	0.20	0.15	0.14	0.12
Average Precision				
ResNet-18	*0.63*	*0.63*	*0.62*	*0.63*
Single Shot Oracle	0.35	0.40	0.48	0.50
Random	0.29	0.31	0.36	0.41
Entropy	0.31	0.34	0.39	0.42
Query by committee	**0.32**	**0.35**	**0.40**	**0.43**
Random clusters & random samples	0.29	0.33	0.38	0.40
Unseen clusters & random samples	0.27	0.28	0.33	0.37
Unseen clusters & uncertain samples	0.29	0.29	0.29	0.28

These results suggest that ResNet-18 does not perform very well in terms of recall, even though it reaches to 82% accuracy in the original class distribution. Recall represents the amount of samples that were successfully labelled as debris out of all debris instances. Therefore, the lower recall of ResNet-18 suggests that the model is not necessarily good at detecting every debris in the data. This could mean the model is not suited for this task, but it could also be that the threshold is not selected well considering the class imbalance. The classification threshold refers to the probability or prediction score above which an instance is assigned to one class, and below which it is assigned to another class. In this particular setting, a threshold of 50% is used, where a prediction score above 50% indicates debris (class 1) with greater certainty than class 0. The precision of ResNet-18 can be up to 97% in the original class distribution. This suggest that almost all the samples labelled as debris by the model actually contain debris. Reducing the threshold would mean reducing the precision while increasing the recall. This would happen because the model would start labelling samples as debris even when it is not at least 50% sure that it actually is debris. On the other hand, since recall is priority for this application, this can be a way to improve ResNet-18's detection performance.

In this study, the Single Shot Oracle method has been used for the average precision metric. This method's performance does not necessarily represent the upper bound in terms of recall since it is optimized for average precision, but it is still a valuable comparison method for the recall results. Optimizing for recall is intentionally avoided since it could mean the model is only predicting the positive class, debris. Therefore, SSO that uses average precision is more reliable. The results show that the SSO method outperforms ResNet-18 model in terms of recall, but cannot compete with it in terms of average precision. This is due to the fact that ResNet-18 outperforms SSO method in terms of precision. Just the way it would improve the ResNet-18's performance, tuning the threshold could potentially improve the average precision performance of the SSO method.

Comparing the different active learning methods, it seems that the best performing one is the query by committee method. Both recall and average precision results show that uncertainty-based methods (entropy and query by committee sampling) outperform diversity-based methods (cluster sampling). Between the two uncertainty-based methods, query by committee sampling performs slightly better than the entropy sampling method. However, neither of these methods perform significantly better than the random sampling method which is the lower bound of the selection strategies. This suggests that using random sampling instead of a more advanced selection strategy is not necessarily a disadvantage. On the other hand, the SSO method shows promise as the upper bound. This concludes that more advanced sampling strategies are needed to reach that upper bound of performance with active learning. Figure 4.8 shows the behavior of the query by committee method compared to all the comparison methods over the increasing number of shots.

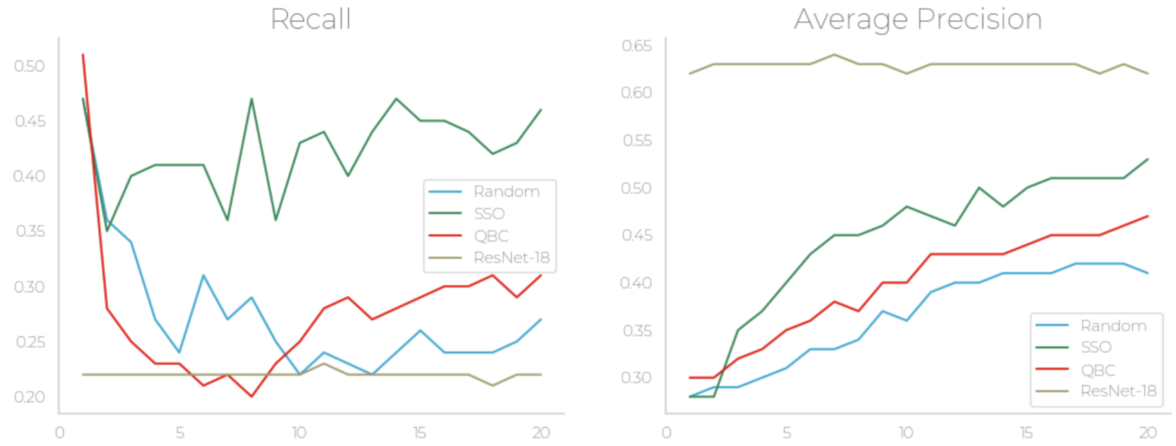


Figure 4.8: The graphs showing the performance progression over increasing number of shots of query by committee (QBC) method and the three comparison methods in terms of recall and average precision.

This figure suggests that the QBC method is placed right in the middle of random sampling and SSO methods' performance, which is expected. ResNet-18 stays above any of these methods for any number of shots. For recall, the situation changes. ResNet-18 is the worst of them in terms of recall. QBC is still placed in between random sampling and SSO, but only after the 10-shots mark. Before this point, it shows very poor and diminishing performance.

On the other side of the performance spectrum, the diversity-based methods are doing very poorly. This suggests that either the feature space used for the clustering step is not representative or the selection made from the clustered samples are not done well. To understand the feature space that is being used for these methods, a 2D t-SNE plot of the features were made. This plot can be seen in Figure 4.9. The distribution in this plot shows that the samples are indeed not separable by clustering them. The debris and non-debris samples are scattered around without following any specific pattern. This suggests that the randomly initialized ResNet-18 model used to obtain the features for samples cannot capture the characteristics of different classes. These results explain why the cluster based methods perform poorly.

The recall and average precision results further suggest that some sampling methods do not show a consistent behavior as the number of shots increase. For example, the recall of random sampling goes down from 5 to 10-shots. This could mean that addition of low-quality samples to the support set can actually hinder the model's ability to detect debris. The same negative trend is visible for all diversity-based methods as well. One possible explanation for this can be that the sampling strategies sample too many non-debris samples, resulting in the model not being able to learn how to distinguish debris sufficiently. In order to confirm this, additional tests are performed.

First of all, to understand the selection behavior of the SSO method as the high-performance upper bound, a random run of this selection strategy was evaluated. It was observed that the steepest increase in average precision happened in the first 4 steps. The samples that were added to the support set as a part of the SSO method in the first 4 steps are visualized in Figure 4.10. These results show that the SSO method tends to pick one debris and one non-debris sample to be added to the support set. Furthermore, it is also seen that the average precision increases rapidly when this happens. This further supports the hypothesis that a few-shot model such as METEOR depends on a good selection of support set samples to reach a high performance.

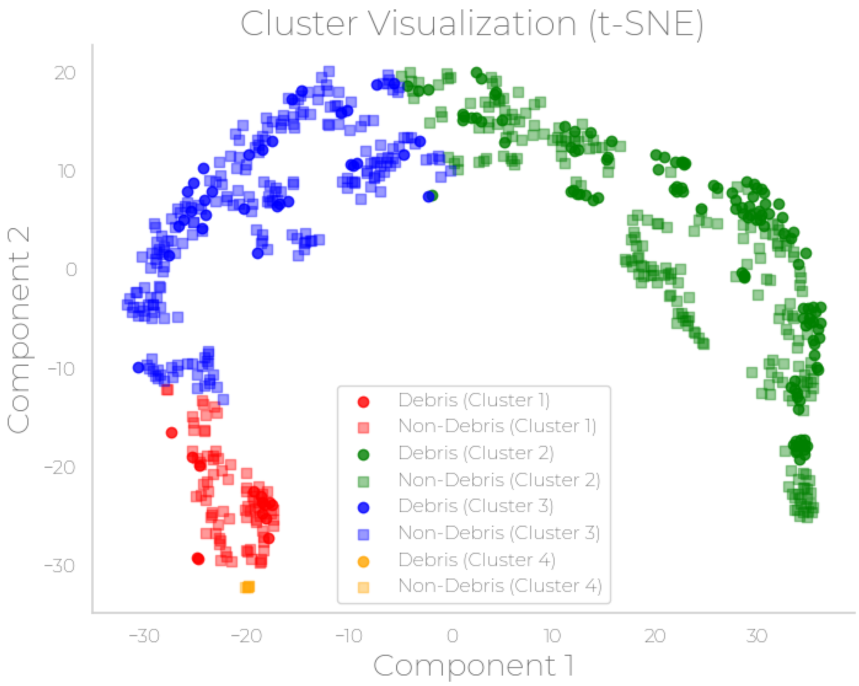


Figure 4.9: 2D t-SNE plot created using the feature vectors of the randomly initialized ResNet-18 which are used for the clustering based sampling methods.

To further analyze these outcomes, several runs of the query by committee method were manually compared. It was observed that the number of debris samples added to the support set was visible as an increase in the recall performance. Table 4.4 shows three example runs and how the performance of the model was affected by the selection of samples. Each run corresponds to the query by committee sampling strategy which is initialized with a different random seed. This means that the first two samples that are fed to the model were different, while the selection strategy was exactly the same in each case.

Table 4.4: Comparison of model performance across different random seed initializations in the query by committee method, highlighting the impact of sample selection on recall and average precision (AP) values. Debris samples are represented by their class value of 1, and non-debris samples are similarly represented by a 0.

#1 (Seed = 18)			**#2 (Seed = 19)**			**#3 (Seed = 20)**		
Additions	**Recall**	**AP**	**Additions**	**Recall**	**AP**	**Additions**	**Recall**	**AP**
1, 0	0.38	0.24	1, 0	0.62	0.26	1, 0	0.17	0.25
0, 0	0.02	0.23	0, 0	0.48	0.34	1, 0	0.65	0.26
0, 0	0.02	0.24	1, 0	0.54	0.32	1, 0	0.82	0.28
0, 0	0.02	0.24	0, 0	0.39	0.32	1, 0	0.96	0.25
0, 0	0.03	0.24	0, 0	0.06	0.25	0, 0	0.77	0.26

The results displayed in Table 4.4 highlight notable differences in recall performance as a consequence of the varying number of additions to the support set. For instance, in the second run (#2), it is seen that the recall drops when a pair of non-debris samples are added to the support set, but goes up again when a non-debris sample is added in the next step. The third run (#3) demonstrated a steady recall increase as long as the addition of samples consisted of one debris and one non-debris sample. These findings indicate a clear relationship between the number of debris samples included in the support set and the resulting model performance, specifically in terms of recall. The performance change is less visible in terms of average precision. For example, the third run (#3) shows that the AP values did not change significantly while the recall values did as more samples were added. This indicates that the

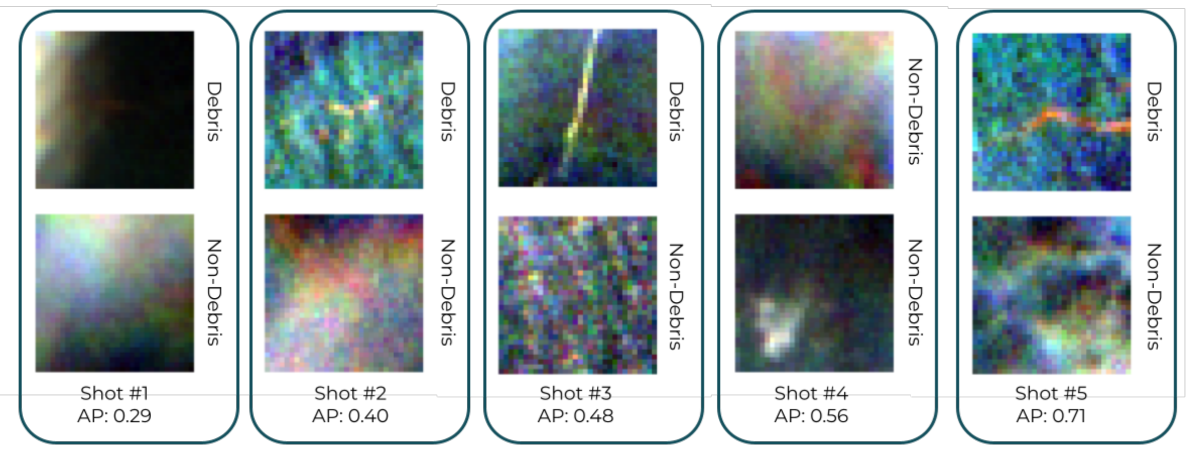


Figure 4.10: The first 4 set of samples which are picked by the SSO method in a single run. The average precision of the model at each step is indicated under the visuals of the added samples.

precision of the model went down as the recall went up. Even though the priority metric was chosen as recall for this application, it is important to evaluate the cost of low precision for the future marine debris monitoring missions using such models. On another note, it is seen that the starting recall values are very different for these three example runs despite all of them starting with one debris and one non-debris sample. Such observations emphasize the significance of sample selection for a few-shot meta-learning model like METEOR and its impact on overall model effectiveness.

As explained in subsection 4.3.1, the number of debris pixels on a satellite image compared to the non-debris pixels would be very little in reality. To evaluate how the models perform in a more realistic class distribution setting, the experiments were also performed on a modified test set. This test set has reduced number of debris samples so that they only amount to 5% of the total number of samples in the set. The results for the Durban region on this more realistic setting are presented in Table 4.5.

Table 4.5: Recall and average precision of various methods on Durban region with different sizes of support sets. These results are for the more imbalanced distribution of the two classes. The values represent the mean value of 20 different random initializations.

	3-shots	5-shots	10-shots	15-shots
Recall				
ResNet-18	*0.24*	*0.18*	*0.25*	*0.13*
Single Shot Oracle	0.14	0.11	0.12	0.11
Random	0.13	0.08	0.04	0.03
Entropy	0.33	0.43	0.35	0.45
Query by committee	0.51	**0.59**	**0.41**	**0.41**
Random clusters & random samples	0.15	0.08	0.03	0.02
Unseen clusters & random samples	0.18	0.08	0.06	0.04
Unseen clusters & uncertain samples	0.15	0.10	0.04	0.02
Average Precision				
ResNet-18	*0.52*	*0.38*	*0.45*	*0.29*
Single Shot Oracle	0.08	0.11	0.14	0.16
Random	0.07	0.07	0.08	0.08
Entropy	0.07	**0.09**	0.10	0.12
Query by committee	**0.09**	**0.09**	**0.13**	**0.18**
Random clusters & random samples	0.07	0.07	0.07	0.08
Unseen clusters & random samples	0.07	0.07	0.07	0.07
Unseen clusters & uncertain samples	0.07	0.07	0.06	0.06

The comparison of the random sampling, QBC sampling and SSO methods in the original and realistic cases are visualized in Figure 4.11.

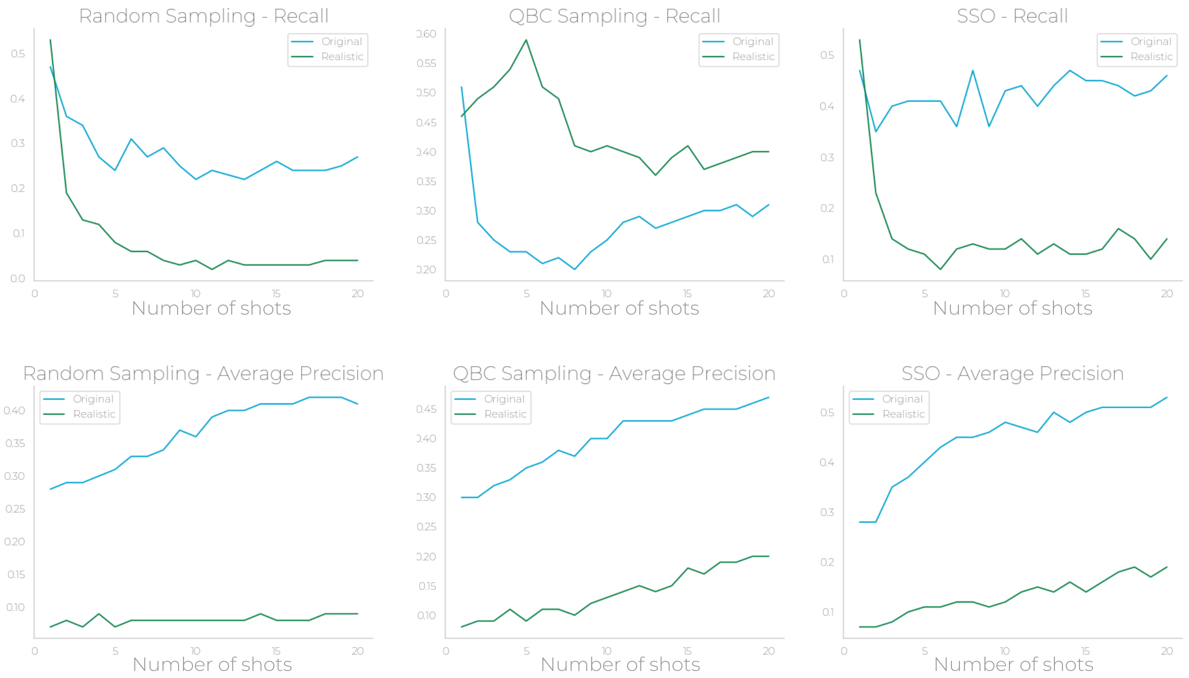


Figure 4.11: Comparison of sampling strategies in the original and realistic class distribution cases. The top row is in terms of recall while the second row is in terms of average precision.

According to these findings, ResNet-18 performed worse than it did in the initial experiments. This decline primarily reflects the model's capacity to identify debris samples because the same model was applied in both instances. Recall and average precision are directly impacted by changes in the test set, demonstrating the ResNet-18 model's limited ability to identify debris samples in this particular dataset. Additionally, in contrast to its relatively stable performance in the initial experiments, ResNet-18's performance changes with an increase in the number of shots in this realistic scenario. This finding suggests that the evaluation of a model's detection performance is influenced by the number of debris samples included in the entire test set.

These results also show that the SSO method performs more poorly both in terms of recall and average precision in the more realistic setting. Furthermore, the random sampling method performs significantly worse in this realistic setting compared to the original experiments, while the QBC method shows a better performance than the random sampling in the realistic setting. This suggests that a non-random sampling strategy is needed more in the realistic setting, while just randomly sampling was sufficient when there is a more balanced class distribution in the dataset. Additionally, the QBC method performed better in terms of recall and worse in terms of average precision compared to the original experiments, meaning that the precision in the realistic case must have been significantly below than the precision in the original case for this method.

The final experiment performed in this study was to apply the same procedures on a different test region, Accra, to see how the performance of various methods changed when the characteristics of the data changed. The results for the original class distribution for the Accra region can be seen in Table 4.6.

Table 4.6: Recall and average precision of various methods on Accra region with different sizes of support sets. These results are for the original distribution of the two classes. The values represent the mean value of 20 different random initializations.

	3-shots	5-shots	10-shots	15-shots
Recall				
ResNet-18	*0.96*	*0.96*	*0.96*	*0.96*
Single Shot Oracle	0.80	0.86	0.93	0.93

Continued on next page

Table 4.6: Continued

	3-shots	5-shots	10-shots	15-shots
Random	0.73	0.85	0.90	0.92
Entropy	0.76	**0.83**	0.86	0.88
Query by committee	0.51	0.67	0.78	0.77
Random clusters & random samples	**0.78**	**0.83**	**0.90**	**0.91**
Unseen clusters & random samples	0.47	0.42	0.57	0.73
Unseen clusters & uncertain samples	0.51	0.67	0.67	0.71
Average Precision				
ResNet-18	*0.96*	*0.96*	*0.96*	*0.96*
Single Shot Oracle	0.88	0.95	0.97	0.98
Random	0.82	0.93	0.95	0.96
Entropy	0.80	0.88	**0.94**	0.95
Query by committee	**0.86**	**0.92**	**0.94**	0.95
Random clusters & random samples	0.85	0.89	0.96	**0.97**
Unseen clusters & random samples	0.73	0.75	0.81	0.88
Unseen clusters & uncertain samples	0.73	0.82	0.89	0.92

These findings demonstrate an overall improvement in performance for all methods in Accra compared to Durban, as indicated by both recall and average precision metrics. The ResNet-18 model exhibits enhanced performance in Accra, suggesting that its ability to detect debris is influenced by the specific characteristics of the region. Notably, the SSO method showcases considerable advancement in terms of both recall and average precision in Accra, reinforcing the notion that it serves as a reliable upper bound for performance evaluation.

Furthermore, it is worth noting that the random sampling method demonstrates a more substantial improvement in the Accra region, compared to its performance in Durban. This observation suggests that the non-random sampling strategies, such as query by committee and entropy-based sampling, may play a more critical role in enhancing performance in scenarios with imbalanced class distributions, as encountered in Durban.

The results presented in Table 4.7 provide further insights into the performance of various methods on the Accra region with respect to the more imbalanced class distribution.

Table 4.7: Recall and average precision of various methods on Accra region with different sizes of support sets. These results are for the more imbalanced distribution of the two classes. The values represent the mean value of 20 different random initializations.

	3-shots	5-shots	10-shots	15-shots
Recall				
ResNet-18	*0.89*	*0.89*	*0.90*	*0.93*
Single Shot Oracle	0.30	0.46	0.64	0.60
Random	0.19	0.06	0.10	0.10
Entropy	0.98	**1.00**	**1.00**	**1.00**
Query by committee	**0.99**	**1.00**	**1.00**	**1.00**
Random clusters & random samples	0.16	0.17	0.11	0.15
Unseen clusters & random samples	0.32	0.15	0.07	0.05
Unseen clusters & uncertain samples	0.38	0.24	0.20	0.17
Average Precision				
ResNet-18	*0.54*	*0.56*	*0.48*	*0.57*
Single Shot Oracle	0.34	0.49	0.61	0.65
Random	0.18	0.18	0.21	0.22

Continued on next page

Table 4.7: Continued

	3-shots	5-shots	10-shots	15-shots
Entropy	0.20	0.24	0.28	0.26
Query by committee	**0.27**	0.19	0.12	0.08
Random clusters & random samples	0.15	0.22	0.18	0.22
Unseen clusters & random samples	0.19	0.20	0.24	0.28
Unseen clusters & uncertain samples	0.18	**0.26**	**0.32**	**0.33**

The recall values demonstrate the effectiveness of the uncertainty-based methods, with both entropy and query by committee sampling achieving high recall scores across different support set sizes. Notably, these methods consistently outperform other strategies, including the ResNet-18 model and the random sampling approach. The entropy sampling method achieves near-perfect recall, indicating its ability to identify debris instances accurately. Similarly, the query by committee method exhibits exceptional recall performance, reinforcing its effectiveness in selecting informative samples. On the other hand, the diversity-based methods, such as random clusters and random samples, as well as unseen clusters and random samples, show relatively lower recall values, suggesting their limitations in identifying debris instances accurately. Furthermore, the average precision values highlight the trade-off between precision and recall. The uncertainty-based methods, particularly unseen clusters and uncertain samples, demonstrate higher average precision scores, indicating their ability to make more precise predictions. Conversely, the query by committee method exhibits lower average precision compared to other strategies, emphasizing the challenge of balancing precision and recall in the context of marine debris detection.

These findings further emphasize the need for customized active learning approaches that account for the specific characteristics of each region, including the class imbalance and environmental conditions. By tailoring the sampling strategies and considering the unique context, it becomes possible to optimize the performance of marine debris detection models. Such insights will be invaluable in future marine debris monitoring missions, where accurate identification and localization of debris instances are essential for effective environmental management.

4.7. Conclusion

The objective of this study was to evaluate various sample selection techniques to achieve high performance in floating marine debris detection using a few-shot meta-learning model. Since being able to detect all debris samples in a new region is the most important aspect, recall was the primary evaluation metric while average precision was the secondary evaluation metric. The comparison study using these two metrics concluded that the active learning methods incorporating uncertainty-based sampling, such as entropy and query by committee, consistently outperformed other strategies in terms of recall and average precision. These findings highlight the effectiveness of leveraging sample uncertainty to select informative data points, resulting in improved debris detection performance.

One reason for the limited effectiveness of diversity-based methods was the poor representativeness of the feature space used to cluster the samples. The study revealed that the clustering of samples based on certain features did not adequately capture the underlying variability and distribution of debris in the target regions. As a result, the diversity-based methods struggled to select informative and representative samples, leading to sub-optimal performance in debris detection. These findings emphasize the importance of considering the representativeness of the feature space when employing diversity-based approaches and highlight the advantages of uncertainty-based methods in identifying informative samples for improved debris detection performance.

Additionally, the study revealed the influence of regional characteristics on the performance of the detection model. Specifically, the results indicated that the ResNet-18 model exhibited enhanced performance in the Accra region, suggesting that the specific environmental conditions and debris characteristics of the region influenced its ability to detect debris accurately. This observation emphasizes the importance of considering regional variability and tailoring the detection framework accordingly to achieve optimal performance in different areas.

Furthermore, the investigation of sampling strategies highlighted the role of class imbalance in the performance of the active learning framework. The random sampling method showed more substantial improvements in the Accra region compared to Durban, indicating its effectiveness in scenarios with imbalanced class distributions. On the other hand, the non-random sampling strategies, such as query by committee and entropy-based sampling, demonstrated superior performance in the presence of imbalanced classes encountered in Durban. These findings suggest that the choice of sampling strategy should be carefully considered based on the class distribution characteristics of the target region.

Ultimately, by offering important insights into the efficacy of various sample selection techniques, this study makes a contribution to the field of floating marine debris detection. The findings emphasize the value of utilizing uncertainty-based sampling techniques and the necessity of adjusting the detection framework to take into account regional variation and class imbalances. These factors can be taken into account to improve the few-shot meta-learning models' ability to detect floating marine debris, aiding in efficient environmental management and conservation efforts. To continue enhancing the precision and effectiveness of marine debris monitoring, various research directions can be identified for future studies, which will be presented in the next section.

4.8. Future Work

Several potential directions for future studies can be identified in the light of results and conclusions of this study. Especially considering the novelty of the topic of using few-shot active learning in marine debris detection, there are many possible areas to be further investigated. This section will list the most relevant research directions which were identified as a result of this study.

1. **Atmospheric correction:** One potential avenue for improving the performance of the proposed method is to incorporate atmospheric correction techniques. This study made use of L1C and L2A data together, and did not perform any additional atmospheric correction techniques. By accounting for atmospheric effects in the remote sensing data, it may be possible to mitigate their influence on the detection performance of the classification models. Especially if the atmospheric correction is performed using local in-situ measurements for each region, the quality of data can be increased substantially, improving reliability of the detection models as well.
2. **Alternative active learning methods:** In this study, the SSO method showed the upper limit of how a sampling method could perform, and even if this might be too unrealistic, it is still possible to find or devise a sampling strategy that performs better than the ones already tested. This could be other uncertainty or diversity-based methods. Further comparative studies involving various other techniques can help identify the most suitable methods for the specific task of marine debris detection.
3. **Feature space exploration:** The only feature space that was used for the clustering of the samples for diversity-based sampling methods was of a randomly initialized ResNet-18 model. However, a more representative feature space could improve the performance of diversity-based methods. For example, METEOR's own feature space could be used. At each step, this feature space would evolve as the task-model is fine-tuned further for a given region. This dynamic use of feature space shall be further investigated in a future study. Another option is to add spectral indices designed to detect debris such as FDI to the features and evaluate if they help with improving detection performance. A more advanced diversity-based sampling strategy has the potential of outperforming the uncertainty-based methods since it can explore a larger part of the sample space, as explained in subsection 4.4.2.
4. **Cluster selection strategies:** Investigating alternative cluster selection strategies can be beneficial for the diversity-based sampling methods. For instance, considering different selection criteria such as prioritizing the largest cluster first, can provide valuable insights into improving the

diversity-based sampling methods' performance. Evaluating and comparing the effectiveness of various cluster selection approaches can lead to more informed decisions in the development of effective sampling strategies.

5. **Training the METEOR meta-model using marine debris data :** In this study, METEOR meta-model trained on land cover classification data was compared to a ResNet-18 model trained on floating marine debris data. This difference in the training process of the two models raise question marks on how fair this comparison can be. A comparative study between METEOR and ResNet-18 both trained on marine debris data would be insightful. This analysis can shed light on the most suitable architecture for marine debris detection task.
6. **Threshold tuning:** This study used average precision as an evaluation metric which measures recall and precision at different thresholds. This metric sometimes is used to hand pick the threshold when there is a desired level of recall or precision to be reached. Experimenting with different thresholds can provide valuable information, especially considering the significant class imbalance in marine debris detection task. Therefore, an important future step would involve tuning the threshold on a test region and subsequently evaluating the model's performance on a separate validation region. This approach will allow for a more rigorous assessment of fine-tuning the decision threshold for optimal performance.

By addressing these points in future studies, it is anticipated that the proposed methodology can be further enhanced, leading to improved effectiveness in marine debris detection using remote sensing data.

5 Synthesis

The aim of this research was to evaluate the efficacy of an active learning approach coupled with a few-shot meta-learning model for the detection of floating marine debris. The main research question together with sub-questions were formulated in and presented in chapter 3. This chapter will demonstrate how the research presented in the journal article answered these research questions.

SQ-1. How does a few-shot meta-learning model perform compared to a conventional deep neural network for marine debris detection?

In terms of recall, METEOR model outperformed the ResNet-18 model in the original class distribution. Just random sampling method was enough surpass ResNet-18's performance. However, no method was able to compete with ResNet-18 in terms of average precision. This indicates that Resnet-18's precision was better than the METEOR model regardless of which sampling strategy was employed.

In a more realistic setting where there is a stark class imbalance, the METEOR model outperformed ResNet-18 in terms of recall but not in terms of average precision. On the other hand, only the SSO method came close to ResNet-18's average precision. While this means that METEOR is a promising model to compete with ResNet-18 with the right selection of support set samples, the right sampling strategy for this has not been discovered yet.

Another important consideration is the labelling effort. It should not be forgotten that ResNet-18 model has seen thousands of marine debris data before being tested on the Durban and Accra datasets, while the METEOR model has only seen a handful of samples that make up the support set while fine-tuning. Therefore, even if the METEOR model does not reach ResNet-18's performance, it still has a significant advantage in terms of reducing the labelling effort and the ability to be fine-tuned for a new region with only a few samples.

SQ-2. How do uncertainty based and diversity based active learning methods compare for this application?

The results of this study demonstrated that uncertainty-based active learning methods, such as entropy and query by committee, consistently outperformed diversity-based methods in terms of recall and average precision. This finding suggests that uncertainty-based sampling strategies are more effective in selecting informative samples for the detection of floating marine debris.

On the other hand, as explained in subsection 4.4.2, diversity based methods has a promising advantage that they can explore regions of the data space that is farther from the classification boundary. This is why this study suggested further studies on different diversity based methods, especially using various feature spaces. The exact methods used in this study performed very poorly but it was discovered that the selected feature space for the clustering step was not very representative. Developing better diversity based methods still holds a promise of being able to compete with or outperform uncertainty-based methods.

SQ-3. How do the sampling strategies perform when there is a more realistic class imbalance?

Every sampling method tested in this study performed worse in a more realistic class distribution setting compared to the original class distribution. This is due to the limited number of debris samples in the dataset once it is reduced to make it more realistic. When the classes are so imbalanced, the sampling strategies struggle selecting samples from both classes. This results in a weakened ability to detect marine debris samples in the test set compared to the original class distribution case.

Developing representative feature spaces and using more advanced cluster based methods are therefore an interesting topic for future studies. If a sampling strategy can keep the number of debris and non-debris samples in the support set balanced, then the METEOR model will not be impacted by the class imbalance as much.

By answering these sub-questions, the results of this study were able to answer the main research question:

How can informative training images be selected for a few-shot meta-learning model to detect floating marine debris patches on publicly available satellite data?

This study highlighted the effectiveness of the active learning approach coupled with the METEOR model for selecting informative training images for the detection of floating marine debris. The findings provided insights into the comparative performance of different models and sampling strategies, pointing towards the potential of uncertainty-based methods and the need for further exploration of diversity-based methods. Furthermore, the study emphasized the impact of class imbalance and the importance of developing strategies to address this challenge for improved detection performance.

Acknowledgments

We all feel it. Too many things are going wrong in this world. It looks like as we ”advance” as the human race, we ruin more and more. Species going extinct, arctic ice melting, plastics ending up in the ocean and killing marine life... It was around the end of my bachelor’s that I was just done with going on with my life, ignoring everything that is going wrong. I knew I could not do much, but doing nothing seemed worse than trying and failing. I believe my first step was educating myself via elective courses during my MSc studies. Then, I did my internship at The Ocean Cleanup and finally felt like I was contributing. This thesis will be the next step on my path to acknowledging all the things that go wrong but still hanging onto some hope that we can maybe fix them.

I would like to thank my supervisors at EPFL, Devis Tuia and Marc Rußwurm, and my supervisor at TU Delft, Jurgen Vanhamel for making it possible in the first place for me to work on this topic. They all taught me so much and I am extremely happy I got to spend the year working along (and with the help of) these brilliant people

As you know, dear reader, a thesis is never only about its content. It is a part of the person who wrote it, and this person has a life of their own and hopefully an acceptable amount of mental health. I am grateful for my family, especially my mom, who valued my education above everything and always supported me no matter what. I would also like to thank my housemates for all the times they were ready to hug me when most needed, all my friends who listened to my complaints for hours on one end, and all my fellow Forcies who helped me stay physically and mentally healthy even at my most stressful times. Finally, I would like to thank two very important people for two very special reasons. First one is my best friend Çağlar: Thank you for being my best friend, always listening to me even if you are across the ocean and of course also for being the reason why I ended up Delft in the first place. Second one is Dave: Thank you for always picking me up if I am down, celebrating with me if something somehow went according to plan, and most importantly for loving me no matter how annoying I get when I am stressed. It is not only me but all of you who made this thesis project possible.

Dilge Gül
Delft, June 2023

References

1. Ryan, P. G., Moore, C. J., Van Franeker, J. A. & Moloney, C. L. Monitoring the abundance of plastic debris in the marine environment. *Philosophical Transactions of the Royal Society B: Biological Sciences* **364,** 1999–2012. doi:10.1098/rstb.2008.0207 (July 2009).
2. Schmaltz, E., Melvin, E. C., Diana, Z., Gunady, E. F., Rittschof, D., Somarelli, J. A., Virdin, J. & Dunphy-Daly, M. M. Plastic pollution solutions: emerging technologies to prevent and collect marine plastic pollution. *Environment International* **144.** doi:10.1016/j.envint.2020.106067 (Nov. 2020).
3. Geyer, R., Jambeck, J. R. & Law, K. L. Production, use, and fate of all plastics ever made. *Science Advances* **3.** doi:10.1126/sciadv.1700782 (July 2017).
4. Lim, X. Microplastics are everywhere — but are they harmful? *Nature* **593,** 22–25. doi:10.1038/d41586-021-01143-3 (May 2021).
5. Issifu, I. & Sumaila, U. R. A review of the production, recycling and management of marine plastic pollution. *Journal of Marine Science and Engineering* **8,** 1–16. doi:10.3390/jmse8110945 (Nov. 2020).
6. Iñiguez, M., Conesa, J. & Fullana, A. Marine debris occurrence and treatment: A review. *Renewable and Sustainable Energy Reviews* **64,** 394–402. doi:10.1016/j.rser.2016.06.031 (Oct. 2016).
7. Ruiz, I., Basurko, O. C., Rubio, A., Delpey, M., Granado, I., Declerck, A., Mader, J. & Cózar, A. Litter Windrows in the South-East Coast of the Bay of Biscay: An Ocean Process Enabling Effective Active Fishing for Litter. *Frontiers in Marine Science* **7.** doi:10.3389/fmars.2020.00308 (May 2020).
8. Biermann, L., Clewley, D., Martinez-Vicente, V. & Topouzelis, K. Finding Plastic Patches in Coastal Waters using Optical Satellite Data. *Scientific Reports* **10.** doi:10.1038/s41598-020-62298-z (Dec. 2020).
9. Rußwurm, M., Wang, S., Kellenberger, B., Roscher, R. & Tuia, D. Meta-learning to address diverse Earth observation problems across resolutions [Manuscript submitted for publication].
10. Maxwell, A. E., Warner, T. A. & Fang, F. Implementation of machine-learning classification in remote sensing: an applied review. *International Journal of Remote Sensing* **39,** 2784–2817. doi:10.1080/01431161.2018.1433343 (May 2018).
11. Zhang, Y. Ten Years of Technology Advancement in Remote Sensing and the Research in the CRC-AGIP Lab in GCE. *Geomatica* **64,** 173. https://link.gale.com/apps/doc/A674565583/AONE (2010).
12. Topouzelis, K., Papageorgiou, D., Suaria, G. & Aliani, S. Floating marine litter detection algorithms and techniques using optical remote sensing data: A review. *Marine Pollution Bulletin* **170,** 112675. doi:10.1016/j.marpolbul.2021.112675 (Sept. 2021).
13. Themistocleous, K., Papoutsa, C., Michaelides, S. & Hadjimitsis, D. Investigating detection of floating plastic litter from space using sentinel-2 imagery. *Remote Sensing* **12.** doi:10.3390/RS12162648 (Aug. 2020).
14. Huang, S., Tang, L., Hupy, J. P., Wang, Y. & Shao, G. A commentary review on the use of normalized difference vegetation index (NDVI) in the era of popular remote sensing. *Journal of Forestry Research* **32.** doi:10.1007/s11676-020-01155-1 (Feb. 2021).
15. Xue, J. & Su, B. Significant remote sensing vegetation indices: A review of developments and applications. *Journal of Sensors* **2017.** doi:10.1155/2017/1353691 (2017).
16. Mifdal, J., Longepe, N. & Rußwurm, M. *Towards detecting floating objects on a global scale with learned spatial features using sentinel 2* in *ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences* **5** (Copernicus GmbH, June 2021), 285–293. doi:10.5194/isprs-annals-V-3-2021-285-2021.

17. Berrar, D. in *Encyclopedia of Bioinformatics and Computational Biology: ABC of Bioinformatics* (eds Ranganathan, S., Nakai, K. & Schonbach, C.) 403–412 (Elsevier, 2019).

18. Pisner, D. A. & Schnyer, D. M. in *Machine Learning* 101–121 (Elsevier, 2020). doi:10.1016/B978-0-12-815739-8.00006-7.

19. Dietterich, T. G. in *Multiple Classifier Systems* 1–15 (Springer, 2000). doi:10.1007/3-540-45014-9-1. http://link.springer.com/10.1007/3-540-45014-9_1.

20. Politikos, D. V., Adamopoulou, A., Petasis, G. & Galgani, F. Using artificial intelligence to support marine macrolitter research: A content analysis and an online database. *Ocean & Coastal Management* **233,** 106466. doi:10.1016/j.ocecoaman.2022.106466 (Feb. 2023).

21. Goodfellow, I., Bengio, Y. & Courville, A. *Deep Learning* (MIT Press, 2016).

22. Ajit, A., Acharya, K. & Samanta, A. *A Review of Convolutional Neural Networks* in *2020 International Conference on Emerging Trends in Information Technology and Engineering (ic-ETITE)* (IEEE, Feb. 2020), 1–5. doi:10.1109/ic-ETITE47903.2020.049.

23. Géron, A. *Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow, 2nd Edition* 2nd ed. (O'Reilly Media, Inc., Sept. 2019).

24. Kikaki, K., Kakogeorgiou, I., Mikeli, P., Raitsos, D. E. & Karantzalos, K. MARIDA: A benchmark for Marine Debris detection from Sentinel-2 remote sensing data. *PLoS ONE* **17.** doi:10.1371/journal.pone.0262247 (Jan. 2022).

25. Solé Gómez, À., Scandolo, L. & Eisemann, E. A learning approach for river debris detection. *International Journal of Applied Earth Observation and Geoinformation* **107.** doi:10.1016/j.jag.2022.102682 (Mar. 2022).

26. Carmo, R., Mifdal, J. & Ruswurm, M. *Detecting Macro Floating Objects on Coastal Water Bodies using Sentinel-2 Data* in *OCEANS 2021: San Diego – Porto* (IEEE, Sept. 2021), 1–7. doi:10.23919/OCEANS44145.2021.9705668.

27. Jing, L. & Tian, Y. Self-Supervised Visual Feature Learning With Deep Neural Networks: A Survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **43,** 4037–4058. doi:10.1109/TPAMI.2020.2992393 (Nov. 2021).

28. Shin, J., Kang, Y., Jung, S. & Choi, J. Active Instance Selection for Few-Shot Classification. *IEEE Access* **10,** 133186–133195. doi:10.1109/ACCESS.2022.3231365 (2022).

29. Rußwurm, M., Venkatesa, S. J. & Tuia, D. *Large-scale Detection of Marine Debris in Coastal Areas with Sentinel-2* Sion, May 2023.

30. Jakubik, J., Blumenstiel, B., Vössing, M. & Hemmer, P. Instance Selection Mechanisms for Human-in-the-Loop Systems in Few-Shot Learning (July 2022).

31. Wang, Y., Yao, Q., Kwok, J. T. & Ni, L. M. Generalizing from a Few Examples. *ACM Computing Surveys* **53,** 1–34. doi:10.1145/3386252 (May 2021).

32. Vanschoren, J. in *Automated Machine Learning* (eds Hutter, F., Kotthoff, L. & Vanschoren, J.) 1st ed., 35–61 (Springer Cham, 2019). doi:10.1007/978-3-030-05318-5-2.

33. Ren, P., Xiao, Y., Chang, X., Huang, P.-Y., Li, Z., Gupta, B. B., Chen, X. & Wang, X. A Survey of Deep Active Learning. *ACM Computing Surveys* **54,** 1–40. doi:10.1145/3472291 (Dec. 2022).

34. Settles, B. Active Learning Literature Survey. *University of Wisconsin, Madison* **52** (July 2010).

35. Liang, Z., Xu, X., Deng, S., Cai, L., Jiang, T. & Jia, K. Exploring Diversity-based Active Learning for 3D Object Detection in Autonomous Driving (May 2022).

36. Yen, S.-J. & Lee, Y.-S. Cluster-based under-sampling approaches for imbalanced data distributions. *Expert Systems with Applications* **36,** 5718–5727. doi:10.1016/j.eswa.2008.06.108 (Apr. 2009).

37. Pezeshkpour, P., Zhao, Z. & Singh, S. *On the utility of active instance selection for few-shot learning* in *NeurIPS HAMLETS* (2020).

38. Li, X., Sun, Z., Xue, J.-H. & Ma, Z. A concise review of recent few-shot meta-learning methods. *Neurocomputing* **456,** 463–468. doi:10.1016/j.neucom.2020.05.114 (Oct. 2021).

39. Tao, R., Zhang, H., Zheng, Y. & Savvides, M. Powering Finetuning in Few-Shot Learning: Domain-Agnostic Bias Reduction with Selected Sampling (Apr. 2022).

40. Main-Knorn, M., Pflug, B., Louis, J., Debaecker, V., Müller-Wilm, U. & Gascon, F. *Sen2Cor for Sentinel-2* in *Image and Signal Processing for Remote Sensing XXIII* (eds Bruzzone, L., Bovolo, F. & Benediktsson, J. A.) (SPIE, Oct. 2017), 3. doi:10.1117/12.2278218.

41. Drusch, M., Del Bello, U., Carlier, S., Colin, O., Fernandez, V., Gascon, F., Hoersch, B., Isola, C., Laberinti, P., Martimort, P., Meygret, A., Spoto, F., Sy, O., Marchese, F. & Bargellini, P. Sentinel-2: ESA's Optical High-Resolution Mission for GMES Operational Services. *Remote Sensing of Environment* **120,** 25–36. doi:10.1016/j.rse.2011.11.026 (May 2012).

42. Zhao, Y., Zhang, X., Feng, W. & Xu, J. Deep Learning Classification by ResNet-18 Based on the Real Spectral Dataset from Multispectral Remote Sensing Images. *Remote Sensing* **14,** 4883. doi:10.3390/rs14194883 (Sept. 2022).

43. He, K., Zhang, X., Ren, S. & Sun, J. *Deep Residual Learning for Image Recognition* in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (IEEE, June 2016), 770–778. doi:10.1109/CVPR.2016.90.

44. Schmitt, M., Hughes, L. H., Qiu, C. & Zhu, X. X. SEN12MS – A Curated Dataset of Georeferenced Multi-Spectral Sentinel-1/2 Imagery for Deep Learning and Data Fusion (June 2019).

45. Wu, J., Chen, J. & Huang, D. Entropy-based Active Learning for Object Detection with Progressive Diversity Constraint (Apr. 2022).

46. Abraham, A. *Diverse Mini-Batch Active Learning: A Reproduction Exercise* Mar. 2020. https://medium.com/data-from-the-trenches/diverse-mini-batch-active-learning-a-reproduction-exercise-2396cfee61df.

47. Yang, S., Liu, L. & Xu, M. Free Lunch for Few-shot Learning: Distribution Calibration (Jan. 2021).

48. Na, S., Xumin, L. & Yong, G. *Research on k-means Clustering Algorithm: An Improved k-means Clustering Algorithm* in *2010 Third International Symposium on Intelligent Information Technology and Security Informatics* (IEEE, Apr. 2010), 63–67. doi:10.1109/IITSI.2010.74.

49. Zhang, P. & Su, W. *Statistical inference on recall, precision and average precision under random selection* in *2012 9th International Conference on Fuzzy Systems and Knowledge Discovery* (IEEE, May 2012), 1348–1352. doi:10.1109/FSKD.2012.6234049.

50. Misiorek, P. & Janowski, S. Hypergraph-based importance assessment for binary classification data. *Knowledge and Information Systems* **65,** 1657–1683. doi:10.1007/s10115-022-01786-2 (Apr. 2023).