Lead detection in the Arctic Ocean

Assessment of thresholding and machine learning classification methods on Sentinel-3 SRAL altimeter for lead detection in the Arctic Ocean

Delft University of Technology

Ericka Martin

TU Delft

Lead Detection in the Arctic Ocean

Assessment of thresholding and machine learning classification methods on Sentinel-3 SRAL altimeter for lead detection in the Arctic Ocean

by

Ericka Martin

A thesis submitted to the Faculty of Aerospace Engineering
of the Delft University of Technology
in partial fulfillment of the requirements for
the degree of Master of Science in Aerospace Engineering,
to be defended publicly on Tuesday June 29, 2021 at 1:30PM.

Version	Date	Description	Change log
1.0	17/05/2021	First Draft	-
2.0	15/06/2021	Final Version	Implementing feedback from supervisors

Student number:	4442687
Project duration:	June 15, 2020 – June 29, 2021
Supervisors:	I. Bij de Vaate
	Dr.Ir. D.C. Slobbe
	Ir. M.C. Naeije
Thesis Committee:	Dr. Ir. E.J.O. Schrama (Chair)
	Prof. dr. ir. M. Snellen (External examiner)
	I. Bij de Vaate (Supervisor)
	Ir. M.C. Naeije (Supervisor)

An electronic version of this thesis is available at http://repository.tudelft.nl/.

Cover image taken by Annie Spratt: https://unsplash.com/photos/U1mQ3wGcvtQ

TU Delft

Acknowledgements

With this thesis report, I am concluding another big chapter of my life, a very special one. Never have I thought I would live in a small city in the Netherlands called Delft for six years and to complete my Masters in Aerospace Engineering. A place now that I can call 'home' has taught me a lot. TU Delft is unarguably where the brightest minds gather, and though it can be intimidating sometimes, I always found myself being supported by my incredible peers and professors. My journey in Delft would have not been the same without my friends, who always made this place a happier place. And my roommates, who are basically now my sisters, who always supported me in many ways: laughing with me, crying with me, studying with me, and most importantly feeding me. And of course Xavier, who always have my back no matter what; on my good days and my bad days, and somehow always make me feel better! I really would like to thank these amazing people that I met through my years at TU Delft.

I would not have been able to experience the incredible time I spent in Delft without the support of my family. I cannot thank them enough for everything they did for me. Our very mobile family have travelled together to many places and has let me see the world and inspired me to work even harder. I will not forget the time they took me all the way to Tanegashima to see the world's most beautiful rocket launching site. Or, our holidays in the Alps where we walked up and down, 25km everyday. These trips were not just for momentarily pleasure, but have fueled me to work for better. Witnessing the rapid glacier melting in the Alps every year definitely has gave me the driving force to purse this thesis topic. Seeing the ice melting left me with a sense of helplessness, which motivated me to contribute to this matter in my own way, using my specialization in aerospace engineering and remote sensing.

Finally, I am fortunate and incredibly grateful to have three wonderful supervisors who supported my thesis for the past year. I would first like to thank Inger, who worked closely with me throughout my thesis. I appreciated your useful input in our discussions, your time and energy to help improve my research, and moreover your constant encouragement and care. I would also like to thank Cornelis, who always provided me with insightful, precise and constructed feedback. Last but not least, I am grateful for the friendly supervision given by Marc, who has also provided me valuable input from the perspective of satellites and aerospace engineering. Strange enough, with the ongoing pandemic I have never physically met my supervisors, but I would like to express my gratitude to them for always willing to help me with my research.

Ericka Martin
Delft, June 2021

Abstract

Detection of the openings in the Arctic sea ice pack, or leads, allow to sample instantaneous sea surface height (SSH) and this information is crucial for quantifying the impact of sea ice melting. It is therefore important to correctly detect as many leads as possible to obtain more SSH references. This paper studies 12 different classification methods including supervised-, unsupervised machine learning methods and thresholding method, being applied to the Sentinel-3 Synthetic Aperture Radar (SAR) altimetry data collected in March/April of 2017-2020 and June/July of 2020 from areas all across the Arctic Ocean. These are compared and assessed with respect to images taken by Ocean and Land Colour Instrument (OLCI), also on board Sentinel-3, ensuring a perfect temporal alignment between the two measurements. The supervised Adaptive Boosting, Artificial Neural Network and Linear Discriminant classifiers showed excellent and robust results in March/April with overall accuracies up to 91.82%. The unsupervised K-medoid classifier produced excellent results achieving up to 91.51% accuracy and it is an attractive classifier as it does not require ground truth data. The classifiers perform poorly in the summer months, as sea ice returns show more ambiguous reflections due to melting. Therefore on summer data, classifications that are solely based on waveform data from SAR altimetry is unsuitable and auxiliary information is required. Furthermore, this paper attempted to identify off-nadir leads (ONL) by adding an extra class in supervised learning methods, intending to reduce the falsely detected leads. Most classifiers failed to detect leads and did not improve their false lead rate. However, as RUS Boost classifier was able to identify 61.6% of total ONLs, this can be used to initially reject these points for more conservative lead detection.

Contents

List of Acronyms

Acronym	Full Name
Ada Boost	Aaptive Boosting
ANN	Artificial Neural Network
ATLAS	Advanced Topographic Laser Altimeter System
AUC	Area Under Curve
Bagging	Bootstrap Aggregation
DBSCAN	Density-Based Spatial Clustering of Applications with Noise
DT	Decision Tree
CryoVEx	CryoSat-2 Validation Experiment
ESA	European Space Agency
FAST4NL	Forecast Arctic Surges and Tides for the Netherlands
FLR	False Lead Rate
HC	Agglomerative Hierarchical Clustering
KNN	K-Nearest Neighbors algorithm
L1-b	Level 1-b
LD	Linear Discriminant
MERIS	Medium Resolution Imaging Spectrometer
MODIS	Moderate-resolution Imaging Spectro-radiometer
NB	Naive Bayes Classifier
NCT	Non-Time Critical
OLCI	Ocean and Land Colour Instrument
ONL	Off Nadir Lead
OOP	Optimal Operating Points
RF	Random Forest
ROC	Receiver Operating Characteristics
RUS	Random Under Sampling
SAR	Synthetic Aperture Radar
SARIn	SAR Interferometric
SLSTR	Sea and Land Surface Temperature Radiometer
SIRAL	SAR/Interferometric Radar Altimeter
SOM	Self Organizing Maps
SRAL	Synthetic Aperture Radar Altimeter
SSD	Stack Standard Deviation
SSH	Sea Surface Height
SUP	Supervised machine learning classifier
SVM	Support Vector Machine
THR	Thresholding classifier
TLR	True Lead Rate
TWL	Total Water Level
UNSUP	Unsupervised machine learning classifier
wf	waveform

List of Symbols

Symbol	Description	Unit
ϵ	Radius of expanding clusters in DBSCAN	-
μ	Mean	-
σ	Standard deviation	-
C	Misclassification cost	-
E	Expectation	-
FI	False ice (number of falsely predicted ice)	-
FL	False lead (number of falsely predicted leads)	-
FLR	False Lead Rate	%
FLR_{ONL}	Total false lead over total waveforms, used in ONL analysis	-
I	Number of ground truth sea ice	-
K	Number of clusters	-
kurt	kurtosis	-
L	Number of ground truth leads	-
LeW	Leading-edge Width	-
MAX	Maximum power value	W
NrPeaks	Number of peaks	-
p_i	Power value at ith bin	W
p_{max}	Maximum power value	W
PP	Pulse Peakiness	-
PPL	Pulse Peakiness Left	-
PPloc	Pulse Peakiness local	-
PPR	Pulse Peakiness Right	-
S	Slope	-
sigma0	Backscatter coefficient	dB
skew	Skewness	-
TeW	Trailing-edge Width	-
TI	True ice (number of correctly predicted ice)	-
TL	True lead (number of correctly predicted leads)	-
TLR	True Lead Rate	%
Wn	Waveform noise	W
Ww	Waveform width	-
X	Random variable	-
X_{ij}	Number of counts in ijth cell of the confusion matrix	-

List of Figures

List of Tables

1 Introduction

The impact of climate change is significant and it is of utmost importance to monitor these drastic changes occurring in the world. In the past decades, satellite remote sensing has successfully provided quantitative information and spatio-temporal states of the climate system and its changes (Yang et al., 2013). Among many other observations, monitoring the Arctic sea ice is considered to be very important to understand the process of global climate change. For example, due to its high reflectivity, sea ice plays a crucial role in the Earth's radiation balance and thermal feedback processes (Screen and Simmonds, 2010). Measuring the sea surface height (SHH) may allow to better quantify the impact of sea ice melting, and this will be further discussed in this chapter.

This chapter introduces the relevance of this research and summarizes the main outcome of the literature study which was conducted prior to this thesis. First, Section 1.1 introduces the current state of the Arctic Ocean and the impact of its sea ice melting. Then, Section 1.2 focuses on the importance of detecting sea ice openings called leads. The state of the art lead detection methods are summarized in Section 1.3. This thesis aims to overcome the challenges and knowledge gaps found from these existing methods. The derived research objective and research questions are presented in Section 1.5. The relevance and novelty of the research are also presented in Section 1.6. Finally, Section 1.7 describes the outline of this report.

1.1. State of Arctic Sea Ice

The Arctic Ocean is partly covered with sea ice, which shows a dynamical behaviour over the seasons and it is a crucial element in various global climate models (Rose et al., 2013). The Arctic Ocean is also a very sensitive area to climate change and many studies have shown that the sea ice concentration is one of the most important influencing factors (C. Lüpkes, 2008). Satellite recordings have been showing a substantial decline in Arctic sea ice extent and sea ice volume during the last decades (Kwok and Rothrock, 2009; Morison et al., 2012; Poisson et al., 2018; Stroeve et al., 2007). Several studies even show that this decline is actually taking place faster than the simulated predictions (Stroeve et al., 2007; Wang and Overland, 2012). Figure 1.1 shows the substantial decline in sea ice extent monthly anomaly. This decrease in sea ice extent further enhances the sea ice melting as the increased open water areas allows more radiation from the sun to be absorbed and heat up the ocean (Rose et al., 2013). Thinner sea ice also has less mechanical strength, which allows it easily break and drift away.

The rapid melting of the Arctic sea ice hence has a profound impact to not only local but to global climate and environments. One of the largest concerns is the fact that the melting of sea ice influences the albedo feedback mechanism, which enhances the climate response, especially at high latitudes (Lindsay and Schweiger, 2015). With more areas of darker colored water being exposed, more radiation is absorbed rather than being reflected. This increases the temperature of Arctic waters and Arctic rivers, in turn warming the air above them and temperature rise can spread over land (Wadhams, 2016). Though recent studies show better understanding of the variation of tides in the Arctic Ocean (Bij de Vaate et al., 2021), is still not well known how the Arctic sea ice decline impact on global tides and surges (Slobbe, 2020). However, some studies also show that the sea ice melting can cause extreme water levels, which can lead to increased erosion (Barnhart et al., 2014), increased risk to the Arctic ecosystems (Kokelj et al., 2012), or impact the dynamics of the tidal

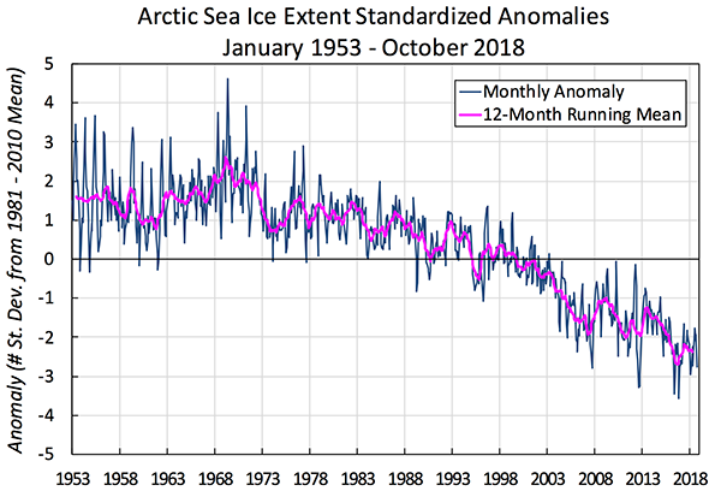


Figure 1.1: Mean sea ice extent anomalies in the years 1953-2018. From January 1953 to December 1979, the data was collected by the UK Hadley Centre and were based on operational ice charts and other sources. From January 1979 to October 2018, the data was collected by passive microwave satellite sensors. (Meier and Stroeve, 2020)

inlet systems in some coastal areas (De Swart and Zimmerman, 2009). For example, a project named Forecast Arctic Surges and Tides for the Netherlands (FAST4NL) recognizes this potential impact of extreme tides and surges in the Dutch coastal areas, and aims to quantify this impact by developing an accurate Arctic total water level (TWL) (Slobbe, 2020).

Measuring the sea surface height (SSH) in the Arctic Ocean could provide essential information in quantifying not only the aforementioned impact of sea ice melting, but also for computing the freeboard (i.e., height of ice surface that is above the surrounding water level (Quartly et al., 2019), see Figure 1.2), sea ice thickness, and ultimately the sea ice volume (Laxon et al., 2013; Wernecke and Kaleschke, 2015). Estimating SSH in the Arctic Ocean is challenging as large part of the ocean is covered by ice and current measurements have substantial uncertainties (Armitage et al., 2016). Tide gauges records are sparse both spatially and temporally, and most are located in along the coasts of Siberian and Scandinavian Arctic, hence many studies and models rely on space-borne measurements (Armitage et al., 2016). Measurements from satellite altimeters make use of sea ice openings called leads, and consider them as instantaneous sea surface height. It is therefore important to correctly find as many leads as possible and minimize the false detection, as falsely detected leads result in overestimation of the SSH (Wernecke and Kaleschke, 2015).

1.2. Lead Detection

Detection of leads can serve for various research purposes. In this thesis, lead detection for SSH estimation is the primary focus, however lead detection also plays a significant role for many other applications such as sea ice thickness computation, understanding of the Arctic climate system, and vessel navigation.

Knowledge of the locations of sea ice openings could aid improving the SSH estimation. These openings in the sea ice are essentially fractures in sea ice covers which occur in zones of divergence and shear motion, and are called polynyas or leads. Polynyas tend to remain at a given location for a longer period of time (Weeks, 2010). Leads on the other hand, are interesting as they have transient features that can form anywhere in an ice-covered ocean and can be refrozen in a short period of time. Leads and polynyas even appear in areas that are covered by thick ice, such as in the central Arctic (Wernecke and Kaleschke, 2015).

The SSH computed using the water levels of the leads can be assumed to be the instantaneous or local SSH (Dettmering et al., 2018). By interpolating these local SSH, the SSH over a larger area within the Arctic Ocean can be computed (Ricker et al., 2015). This is illustrated in Figure 1.2, showing the sea surface height tie points. Therefore, more correctly detected leads will allow for more data points for interpolation, and the SSH field can be better computed. However, is leads are falsely detected, this translate into an overestimation of the SSH and ultimately result in a negative bias in the estimated freeboard (Wernecke and Kaleschke, 2015). Therefore, it is of high interest to correctly find as many leads as possible to decrease the statistical error found in SSH estimations.

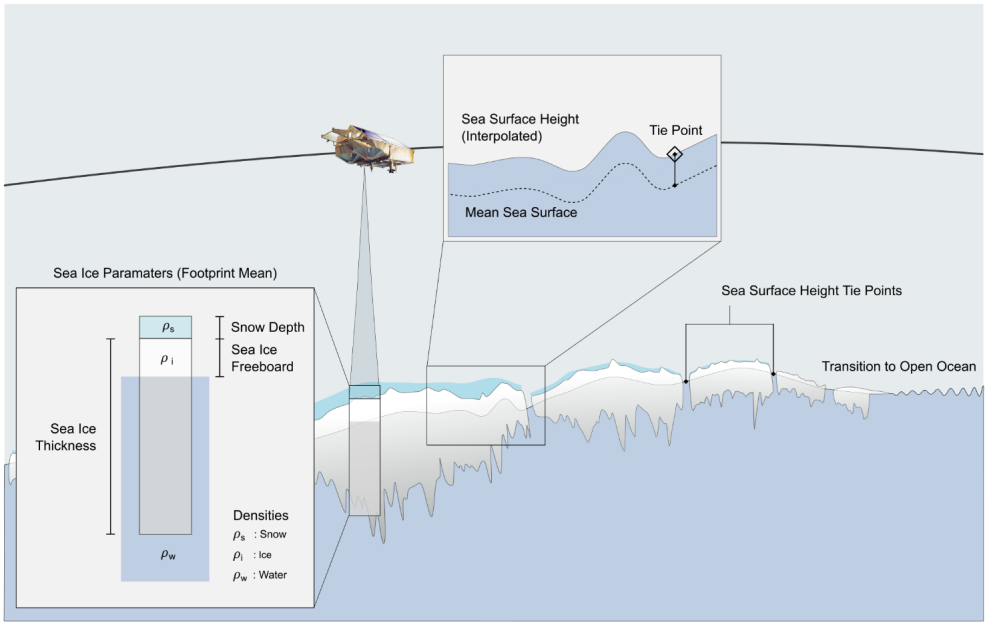


Figure 1.2: A schematic diagram showing a satellite with radar altimeter flying over the Arctic sea ice with several leads (Quartly et al., 2019)

This thesis focuses on SAR altimeter for many reasons. SAR altimeter exploits the Delay/Doppler effects and coherent processing of the group of transmitted pulses. Hence, the SAR altimeters can make more efficient use of the power reflected from the surface, compared to the conventional pulse limited altimeters (European Space Agency and CNES, 2020). Due to its simultaneous increase in the number of looks, the along-track resolution is also increased and the waveform returns are much sharper (Raney, 1998).

The waveform returns from SAR altimetry hold essential information in surface classifications. The discrimination relies on the different surface reflectivity and differences in the incidence angle of the radar pulse. This allows to distinguish whether the echo is dominated by specular reflections coming from smooth surfaces such as leads, showing peaky waveforms, or by diffuse reflections coming from rough surfaces such as the sea ice or ocean surface, which shows more noisy waveform returns (Laxon et al., 2013). Examples of these waveform returns are shown in Figure 1.3.

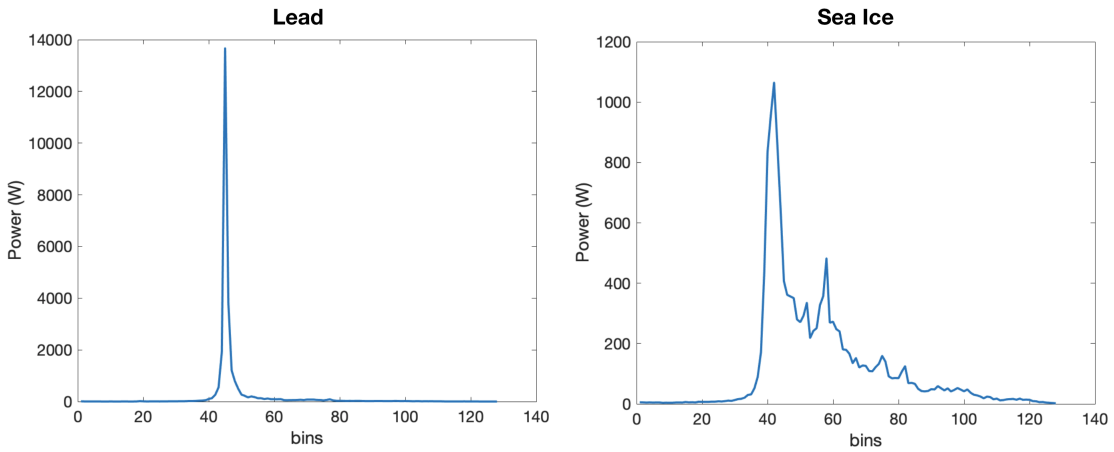


Figure 1.3: Example waveform of lead (left) and sea ice (right), acquired by the SRAL altimeter.

1.3. State-of-the-art Methods

Number of studies have demonstrated the use of SAR altimeter waveform returns for surface type classification in the Arctic Ocean. Historically, the empirical thresholding methods were used for waveform classifications, however with the rise of artificial intelligence in recent years, the adoption of machine learning algorithms in this context has been seen more frequently. The thresholding methods simply assign a range of values per waveform feature that should be met in order for them to be classified to a specific class (e.g. Laxon et al., 2013; Ricker et al., 2014; Schulz and Naeije, 2018). Machine learning classifiers on the other hand, use machine learning algorithms by training with labeled (supervised learning) or unlabeled (unsupervised learning) data set to distinguish different features of the waveform and assign them to their classes (e.g. Dettmering et al., 2018; Lee et al., 2016; Müller et al., 2017; Poisson et al., 2018)).

Empirical thresholding methods are based on setting thresholds to the waveform features to unique waveforms for different surface types. Earlier studies used a single waveform feature to classify various surface types in the Arctic Ocean. Peacock and Laxon (2004) for example have set the Pulse Peakiness (PP; waveform feature describing the "peakiness" of the waveform) value of 1.8 as the boundary. Waveforms with PP of less than 1.8 were processed as diffuse (classified as either open ocean or sea ice), and PP of greater than 1.8 were processed as specular (classified as leads). Later, other studies used more waveform features and added different thresholds to improve the results. For instance, the classification performances obtained by the studies from Laxon et al. (2013) and Rose et al. (2013) were assessed using the same ground truth and although both methods achieved high overall accuracies, the two methods tend to classify ice as leads, implying their tendency in overestimating leads (Lee et al., 2016). This tendency in overestimation of leads can lead to a strong bias in SSH and freeboard estimation.

Recently, more studies have adopted machine learning algorithms for waveform classification. Machine learning based methods can overcome several shortcomings associated with the simple thresholding methods. For example, simple thresholding methods may suffer because waveform features typically contain aliasing between leads and sea ice (Lee et al., 2016). The machine learning classifiers can be largely divided into two types; supervised learning and unsupervised learning.

Application of supervised machine learning was used for example by Lee et al. (2016), where Decision Trees and Random Forest classifiers were used for lead detection using CryoSat-2 waveform data. These machine learning classifiers were also compared to the thresholding classifiers using the same input data and ground truth. The accuracy obtained by machine learning algorithms clearly outperformed the thresholding classifiers. The derived sea ice thickness estimation based on these methods also showed that the supervised machine learning classifiers resulted in lower error values when comparing to ice thickness data obtained by CryoVEx (a dedicated aircraft campaign for CryoSat validation) (Lee et al., 2016). However, the remarkable accuracies obtained by DT and RF were computed based on 239 testing points on a very limited study area. For the purpose of estimating the SSH of larger part of the Arctic Ocean, the classifiers shall be applied to a larger testing set to fully comprehend their performances.

Müller et al. (2017) and Dettmering et al. (2018) applied unsupervised machine learning methods, specifically K-medoid clustering. Unlike supervised learning, unsupervised learning does not train with labeled data, hence the model does not rely on the ground truth but rather on natural patterns and dissimilarities of the training data. Because the data is unlabeled, once the clusters are obtained, the operator must manually label the clusters to the corresponding classes. The classification accuracies obtained by K-medoid clustering were promising, as they mostly outperformed the thresholding classifiers. This study has computed the accuracy with 14,231 testing points, but again were limited to a smaller area in the Arctic Ocean.

1.4. Knowledge Gap

Despite the promising machine learning algorithms presented in the earlier studies, there are still some knowledge gaps that must be filled. This thesis focuses on four issues that have been identified.

Firstly, as most of the previous studies focused on different study areas, input data and used different validation data, the results cannot be directly compared. Although many studies have demonstrated the advantages of machine learning classification over thresholding classification methods, it is for example still unknown whether unsupervised machine learning can outperform supervised machine learning methods when tested on the same study area.

Secondly, the validation data used in the previous studies were limited to small areas; Dettmering et al.

(2018) used waveform data measured only in the Greenland Sea, whereas Lee et al. (2016) used the MODIS image to validate only 239 waveforms. This is due to the fact that many studies manually validated the altimeter classification results with the ground truth data through visual inspection (Lee et al., 2016; Bij de Vaate, 2019; Quartly et al., 2019).

Thirdly, seasonal influence on the classification performance is not well known. Bij de Vaate (2019) demonstrated a significant increase in ambiguous class and lead class during the summer months, when applying the thresholding classifier. Shu et al. (2020) applied different supervised machine learning algorithms for the months of November to May, and showed worsening performance in May compared to the other months. Hence, classification performances, especially for machine learning classifiers during the summer months are not well studied and requires further research.

Finally, many studies acknowledge that the presence of off-nadir leads (ONL) can worsen the classification performance as ONL can dominate the radar footprint and dominate the range retrieval (Ricker et al., 2015). The waveform returns of ONLs often resemble a superposition of specular and diffuse returns but the contribution from the ice can be totally over-shadowed by the off-nadir lead. Therefore, there is no concrete understanding of the waveform returns of off-nadir leads (Quartly et al., 2019).

1.5. Research Objective and Research Questions

Given the analysis of the knowledge gap in the state of the art methods, the research objective and research (sub) questions were formulated for this thesis. These are presented in Section 1.5.1 and Section 1.5.2, respectively.

1.5.1. Research Objective

The research objective of the research is defined as follows.

> *Assess different SAR altimetry waveform classification methods for lead detection in the Arctic Ocean and identify the most suited ones to be applied for sea surface height estimation. .*

1.5.2. Research Questions

The main research question was formulated by carefully assessing the points that required further research which closely relates to the research objective. The main research question is given as follows.

> *How do empirical thresholding methods, supervised and unsupervised machine learning based classifiers compare in their performances, and what advantages/disadvantages do they have when detecting leads in the Arctic Ocean using SAR altimetry data?*

With this research question, this research serves as a follow up study from state-of-the-art. It will also be able to contribute to the research objective directly by comparing three very different methods that are assessed with same input and validation methods.

In order to support answering the main research question in a detailed manner, the following sub-questions are defined.

- **SQ-1:** How can SAR altimetry waveform classifications be validated in an effective way?
- **SQ-2:** How do the overall accuracy, True Lead Rate, and False Lead Rate compare between the classifiers?
- **SQ-3:** What is the combination of waveform features that results in best accuracy?
- **SQ-4:** How do the hyperparameters in machine learning algorithms influence the classification performance?
- **SQ-5:** How well do classifiers perform when trained with data from another year?
- **SQ-6:** How well do classifiers perform when trained with data from another study area?

- **SQ-7:** What is the seasonal influence on the performance of classifiers?
- **SQ-8:** How does the presence of off-nadir leads influence the performance of the classifiers?

1.6. Novelty and Relevance

Given the defined research objective and research questions of this thesis, its relevance and novelty are introduced in this section. Firstly, the identified novelty of the research are listed in the following.

- **Comparison of three types of classifiers on the same ground truth:** The most important novelty of this research lies in comparing thresholding, supervised and unsupervised machine learning classifiers on the same ground truth. These three types of classifiers have never been assessed and compared using the same input data and the ground truth, therefore this research will provide fair and valuable comparison between the classifiers.
- **Use of Sentinel-3 SRAL altimeter and OLCI images for lead detection:** A key opportunity of using the Sentinel-3 satellite was identified, as this satellite is equipped with Synthetic Aperture Radar Altimeter (SRAL) and Ocean and Land Colour Instrument (OLCI). Classification conducted on Level 1-b (L1-b) waveform data acquired by SRAL can be validated using OLCI images. This unique combination of instruments will provide a perfect temporal alignment between the optical and altimetry measurements.
- **Attempt on detecting off-nadir leads:** This study also attempts to understand the effect of adding another class to the classification problem. If off-nadir leads can be correctly detected, number of falsely predicted leads may be reduced. Furthermore, the points which were classified as off-nadir leads points can be rejected from the analysis, if a conservative prediction of leads is desired.

Finally, the results obtained in this research will be relevant to the research community in the Arctic or cryospheric studies. By identifying the best performing classification method for lead detection, sea surface height estimation can be improved in the Arctic region. This can positively impact to research in many areas, such as computation of the ice thickness, ice volume loss, and ultimately its impact to the Arctic and the global climate.

1.7. Report Outline

The structure of this thesis report is given as follows. Firstly, Chapter 2 presents the main content of the research written in a form of a draft journal article. The article presents the methodology, definition of classification models, results, conclusions and recommendation for future work. Then, Chapter 3 relates the conclusions back to the research questions formulated as seen in Section 1.5. It also includes more elaboration on the recommendation and future work. Finally, this report includes appendices; Appendix A is an appendix directly contributing to the journal article presented in Chapter 2, and Appendix B presents extra supporting materials.

2

Journal Article

Assessment of thresholding and machine learning classification methods on Sentinel-3 SRAL altimeter for lead detection in the Arctic Ocean

Ericka Martin, Inger Bij de Vaate, Cornelis Slobbe, Marc Naeije

(Delft University of Technology, Delft, Netherlands
E.Martin-2@student.tudelft.nl)

Abstract: Detection of the openings in the Arctic sea ice pack, or leads, allow to sample instantaneous sea surface height (SSH) and this information is crucial for quantifying the impact of sea ice melting. It is therefore important to correctly detect as many leads as possible to obtain more SSH references. This paper studies 12 different classification methods including supervised-, unsupervised machine learning methods and thresholding method, being applied to the Sentinel-3 Synthetic Aperture Radar (SAR) altimetry data collected in March/April of 2017-2020 and June/July of 2020 from areas all across the Arctic Ocean. These are compared and assessed with respect to images taken by Ocean and Land Colour Instrument (OLCI), also on board Sentinel-3, ensuring a perfect temporal alignment between the two measurements. The supervised Adaptive Boosting, Artificial Neural Network and Linear Discriminant classifiers showed excellent and robust results in March/April with overall accuracies up to 91.82%. The unsupervised K-medoid classifier produced excellent results achieving up to 91.51% accuracy and it is an attractive classifier as it does not require ground truth data. The classifiers perform poorly in the summer months, as sea ice returns show more ambiguous reflections due to melting. Therefore on summer data, classifications that are solely based on waveform data from SAR altimetry is unsuitable and auxiliary information is required. Furthermore, this paper attempted to identify off-nadir leads (ONL) by adding an extra class in supervised learning methods, intending to reduce the falsely detected leads. Most classifiers failed to detect leads and did not improve their false lead rate. However, as RUS Boost classifier was able to identify 61.6% of total ONLs, this can be used to initially reject these points for more conservative lead detection.

2.1. Introduction

Declining sea ice cover in the Arctic is a strong indicator of climate change. In the past decades, satellite recordings have been continuously measuring a substantial decline in Arctic sea ice extent (Kwok and Rothrock, 2009; Morison et al., 2012; Stroeve et al., 2007). Recent studies have also shown that the Arctic will be practically ice free in summer by the year of 2050 (Notz and SIMIP Community, 2020). Such drastic decline of Arctic sea ice does not only have an impact locally but also has profound impact on global climate and environments (Wadhams, 2016). For example, the melting of sea ice influences the albedo and thermal feedback mechanism, which may further enhances the warming in the Arctic (Lindsay and Schweiger, 2015). Though recent studies show better understanding of the variation of tides in the Arctic Ocean (Bij de Vaate et al., 2021), is still not well known how the Arctic sea ice decline impact on global tides and surges (Slobbe, 2020). However, some studies also show that the sea ice melting can cause extreme water levels, which can lead to increased erosion (Barnhart et al., 2014), increased risk to the Arctic ecosystems (Kokelj et al., 2012), or impact the dynamics of the tidal inlet systems in some coastal areas (De Swart and Zimmerman, 2009).

Measuring the sea surface height (SSH) in the Arctic Ocean could provide essential information in quantifying not only the aforementioned impact of sea ice melting, but also aid to estimate the sea ice thickness and its volume loss (Laxon et al., 2013; Wernecke and Kaleschke, 2015). Estimating SSH in the Arctic Ocean is challenging as its large area is covered by ice and current measurements have substantial uncertainties (Armitage et al., 2016). It is therefore essential to detect and measure the water levels of the sea ice openings called leads. It is of great interest to correctly detect as many leads as possible and minimize the false detection, as falsely detected leads result in overestimation of the SSH (Wernecke and Kaleschke, 2015).

Synthetic Aperture Radar (SAR) altimeters on satellites are useful for lead detection, as the surface types can distinguished by the waveform shapes. Its coverage is also advantageous, especially compared to tide gauges which are located sparsely, mostly located on the coasts of Siberian and Scandinavian Arctic (Armitage et al., 2016). As SAR altimeters exploit the Delay/Doppler effects and coherent processing of the group of the

reflected pulses, they allow for better along-track resolution and sharper waveform returns as compared to the conventional pulse-limited altimeters (Dinardo and Benveniste, 2013; Raney, 1998).

A number of classification methods have been developed, specifically for distinguishing leads from the ocean and sea ice, using SAR altimetry waveform data. The state of the art waveform classification methods can be largely divided into two group; empirical thresholding methods and machine learning classification methods. Most of the methods make use of waveform features, which describe the unique features of the waveform returns. Empirical thresholding method relies on setting thresholds to these waveform features in order to distinguish the surface types, and historically this simple method has been widely made use of. However, a study from Lee et al. (2016) showed that thresholding methods tend to over-detect leads, which ultimately leads to a strong bias in SHH estimation. Recently, more machine learning algorithms were adopted for waveform classifications, including the supervised tree based classification from Lee et al. (2016), supervised artificial neural network approach from Poisson et al. (2018), and the unsupervised K-medoid classification method from Dettmering et al. (2018) and Müller et al. (2017). These machine learning methods produced higher accuracies as it can overcome shortcomings associated with the simple thresholding methods, such as dealing with waveforms that contain aliasing between leads and sea ice (Lee et al., 2016).

Despite the promising machine learning algorithms presented in the earlier studies, there are still some knowledge gaps that must be filled. This paper focuses on four issues that have been identified. Firstly, the aforementioned classifiers and their performances cannot be directly compared as these studies involve different study areas, sensors, and validation data. For instance, it is still unknown whether using unsupervised machine learning classifiers can outperform supervised learning classifiers. Secondly, the validation data used in the previous studies were limited to small areas; Dettmering et al. (2018) used waveform data measured only in the Greenland Sea, whereas Lee et al. (2016) used the MODIS image to validate only 239 waveforms. This is due to the fact that many studies manually validated the altimeter classification results with the ground truth data through visual inspection (Bij de Vaate, 2019; Lee et al., 2016; Quartly et al., 2019). Thirdly, seasonal influence on the classification performance is not well known. Bij de Vaate (2019) demonstrated a significant increase in ambiguous class and lead class during the summer months, when applying the thresholding classifier. Shu et al. (2020) applied different supervised machine learning algorithms for the months of November to May, and showed worsening performance in May compared to the other months. Hence, classification performances, especially for machine learning classifiers during the summer months are not well studied and requires further research. Finally, many studies acknowledge that the presence of off-nadir leads (ONL) can worsen the classification performance as ONL can dominate the radar footprint and dominate the range retrieval (Ricker et al., 2015). The waveform returns of ONLs often resemble a superposition of specular and diffuse returns but the contribution from the ice can be totally over-shadowed by the off-nadir lead. Therefore, there is no concrete understanding of the waveform returns of off-nadir leads (Quartly et al., 2019).

To tackle this, different algorithms from all supervised-, unsupervised machine learning, and thresholding methods were applied to wide range of study areas in the Arctic Ocean, in order to gain a complete understanding of the performance of different classifiers and ultimately identify the most suited ones to be applied for SSH estimation. Although an operator-controlled selection of cloud free regions is required, this paper aims to automate the majority of validation data generation process by means of image segmentation and analysis of radiance change on optical image pixel nadir to the satellite. This way, a much larger study area can be validated in a consistent manner. The classifiers will also be applied in the summer months to better understand the seasonal influences. Furthermore, this study attempts to detect ONLs to reduce the false lead detection by adding another class to the classification problem, for supervised machine learning classifiers.

A key opportunity of using Sentinel-3 satellites was identified, as these satellites are equipped with Synthetic Aperture Radar Altimeter (SRAL) and Ocean and Land Colour Instrument (OLCI). Classification conducted on Level 1-b (L1-b) waveform data acquired by SRAL can be validated using OLCI images. This unique combination of instruments will provide a perfect temporal alignment between the optical and altimetry measurements. This approach is similar to the work by Poisson et al. (2018), where images from Medium Resolution Imaging Spectrometer (MERIS) were used to validate Radar Altimeter-2 (RA-2), in which both sensors were on board of the Envisat. This combination of instruments is extremely beneficial to this research because ice drift models need to be employed if the two measurements have few hours of delay, as the current speed can reach up to 0.1 km/h (Quartly et al., 2019).

2.2. Data and Methodology

2.2.1. Data

2.2.1.1 SRAL altimeter

This study makes use of SAR altimetry Non-Time Critical (NTC) L1-b waveform data retrieved by the Synthetic Aperture Radar Altimeter (SRAL) instrument of Sentinel-3A and Sentinel-3B satellites. The SAR mode in SRAL ensures high along-track resolution of approximately 300 m (EUMETSAT, 2017). The acquired waveform is sampled in 128 bins and this usually depicts a sudden rise in power on the leading edge and declining power on the trailing edge (Shu et al., 2020). Figure 2.1 shows an example of this waveform return.

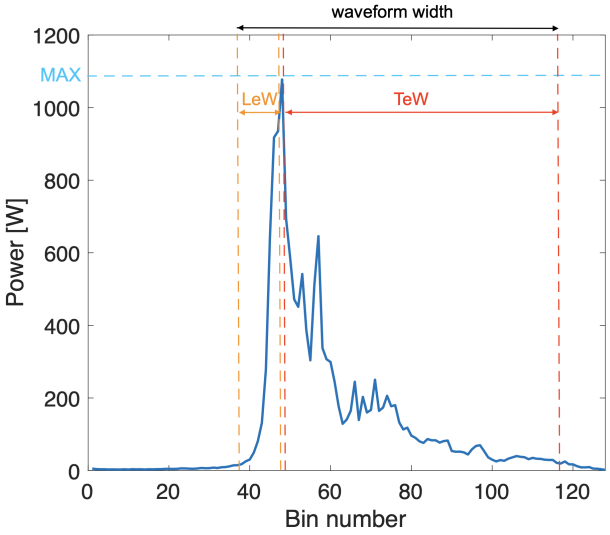


Figure 2.1: Example of SRAL L1b waveform return. Waveform features; maximum power, leading-edge width, trailing-edge width and the waveform width, are presented in the figure.

The shape of the altimetry waveform returns are dependent on the surface characteristics within the altimeter footprint (Müller et al., 2017). For example, specular reflections coming from smooth surfaces, such as leads, show narrow and peaky waveforms. In contrary, the diffuse reflections coming from rough surfaces, such as sea ice, show wider and more noisy waveform returns (Laxon et al., 2013; Poisson et al., 2018).

In order to distinguish these waveforms and identify their surfaces, number of waveform features are defined. These waveform features can describe the dissimilarities between the waveforms and are used as an input for the classification process. This study made use of 12 types of waveform features and these are summarized in Table 2.1.

2.2.1.2 OLCI images

This paper assesses the waveform classification results by directly comparing them to images taken by the Ocean and Land Colour Instrument (OLCI) on board of Sentinel-3. From the OLCI images, one can visually distinguish leads from ice by identifying the darker areas (lower radiance value) on the ice sheets (Ludwig et al., 2020). The surface classes determined by the OLCI images are considered to be the ground truth. With a spatial resolution of approximately 300m (Su et al., 2019), OLCI images cannot always depict the real ground truth as it cannot detect leads that are narrower than its resolution. Though it is an approximation of the reality, the use of OLCI images provides a common ground to compare different altimetry-based classifications.

OLCI is a push-broom imaging spectrometer which contains 21 spectral bands (Oa1 - Oa21) ranging from 400mm to 1020mm (Bourg et al., 2021). This study uses the L1-b product which consists of images of top of atmosphere radiance, calibrated to geophysical units (Wm^{-2} $sr^{-1}mm^{-1}$). The image is georeferenced onto the Earth's surface, and spatially resampled onto an evenly spaced grid (European Space Agency, 2016). In this study, pseudo-color images have been constructed using the three spectral bands of OLCI data; Oa3 (442.5nm), Oa5 (510nm) and Oa8(665nm), which accentuated the color differences in the images.

Table 2.1: Description and equations of waveform features considered in this study

Waveform Feature (Abbreviation)	Description	Equation
Maximum Power (MAX)	Maximum power value of the waveform in Watts	-
Kurtosis (kurt)	Kurtosis is a measure of peakiness of the power distribution (Lee et al., 2016). Kurtosis of a random variable is a 4th standardized moment, which can be computed by dividing the 4th central moment of the distribution by the 4th power of standard deviation.	$\text{Kurt}[X] = \text{E}\left[\left(\frac{X-\mu}{\sigma}\right)^4\right] = \frac{\text{E}[(X-\mu)^4]}{(\text{E}[(X-\mu)^2])^2}$ where X is the random distribution, μ is the mean and σ is the standard deviation.
Skewness (skew)	Skewness is a measure of how slanted the power distribution is. Skewness represents 3rd standardized moment, which can be computed by dividing the 3rd central moment of the distribution by the 3rd power of standard deviation.	$\text{Skew}[X] = \text{E}\left[\left(\frac{X-\mu}{\sigma}\right)^3\right] = \frac{\text{E}[(X-\mu)^3]}{(\text{E}[(X-\mu)^2])^{3/2}}$
Pulse Peakiness (PP)	PP is a measure of the peakiness of the waveform. It is found by the maximum power value divided by the total accumulated power ($\sum P_i$) from all the bins in the waveform (Wernecke and Kaleschke, 2015).	$\text{PP} = \frac{p_{max}}{\sum_{i=1}^{128} p_i}$
Backscatter Coefficient (sigma0)	Sigma0 is the radar backscatter coefficient, which describes the surface properties, radar frequency, polarization and incident angle (Wingham et al., 2006). Sigma0 values are taken from the SRAL L1b data.	-
Waveform Width (WW)	WW is defined as the number of range bins with their power greater than 1% of the maximum power (see (see Figure 2.1).	-
Leading-edge Width (LeW)	LeW is defined as the bin width between 1% and 99% of the maximum power value (see Figure 2.1).	-
Trailing-edge Width (TeW)	TeW is defined as the bin width between 99% and 1% of the maximum power value (see Figure 2.1).	-
Pulse Peakiness Left (PPL)	PPL is a modified form of PP, where it only considers only the three range bins on the left of the maximum bin. This show the peakiness of the left side of the waveform.	$\text{PP}_\text{l} = \frac{p_{max}}{\Sigma([p_{i_\text{max}-3}, p_{i_\text{max}-1}])}$
Pulse Peakiness Right (PPR)	PPR uses three bins on the right side of the maximum power to describe the peakiness of the right side of the waveform.	$\text{PP}_\text{r} = \frac{p_{max}}{\Sigma([p_{i_\text{max}+1}, p_{i_\text{max}+3}])}$
Pulse Peakiness local (PPloc)	PPloc uses three bins on the left and on the right of the maximum power to describe the 'local' peakiness of waveform surrounding the maximum power.	$\text{PP}_\text{loc} = \frac{p_{max}}{\Sigma([p_{i_\text{max}-3}, p_{i_\text{max}-1}], [p_{i_\text{max}+1}, p_{i_\text{max}+3}])}$
Number of Peaks (NrPeaks)	NrPeaks is found by counting the number of peaks which have the peak prominence larger than 0.01 in the normalized waveform.	

2.2.2. Study Areas and Dates

The selection of the specific tracks have been strongly dependent on the OLCI images, as only the cloud free images can be used. The corresponding altimetry track has been laid over these images to further specify the study areas. The resulting data stretches over a wide range of longitude and latitude (see Figure 2.2), which consists of 35 different OLCI images and 18,242 waveforms in total. Note that these study areas are limited up to 81.35° in latitude due to the orbital inclination of the Sentinel-3 satellite (98.65°). The study focuses on March and April for the winter seasons due to the polar nights and limited lighting conditions experienced in other months.

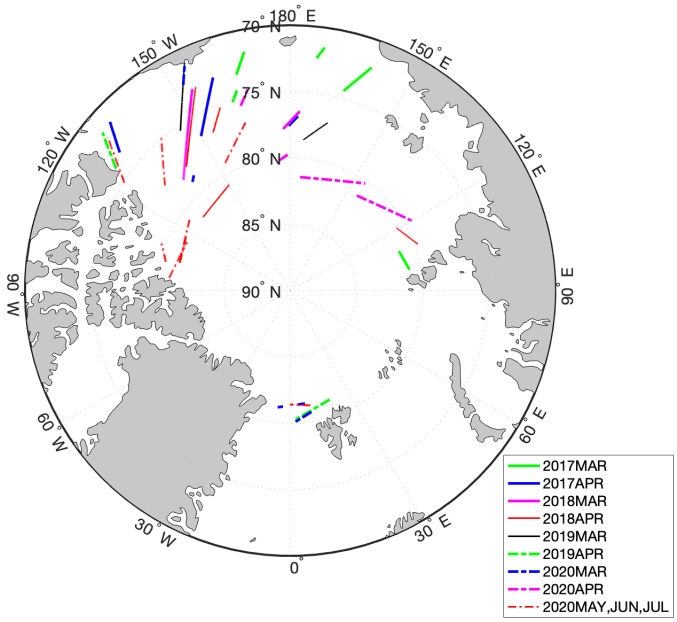


Figure 2.2: Areas to be studied in this paper; Sentinel-3A/3B altimetry tracks from March 2017 to July 2020.

2.2.3. Waveform Classifiers

This study has assessed a total of 12 classifiers, including nine different supervised machine learning classifiers, two unsupervised machine learning classifiers, and one thresholding classifier.

2.2.3.1 Supervised machine learning classifiers

Supervised machine learning classifiers make use of labeled training data, which consist of input features (waveform features) and their corresponding class (leads or sea ice). Supervised learning algorithms learn from the training data through an iterative optimization of an objective function to build a mathematical model (Mohri et al., 2012). The model is then also able to predict new input features, which are provided as the testing data. The nine supervised learning classifiers which are analysed in this report and their hyperparameters to be tuned are presented in the following.

- **Tree based classification**
  Tree-based classification methods have been widely used and showed promising results in many remote sensing applications (Shu et al., 2020; Xu et al., 2014), including lead detection (Lee et al., 2016). A decision tree model is the foundation of all tree-based models. In general, tree-based ensemble classifiers combine many classification trees for improved and robust prediction (Hastie et al., 2009). Ensemble tree classifiers, including bagging and boosting methods, are built with many decision trees in parallel (bagging) or sequential (boosting) manner. This paper explores these three types of tree based classifiers (decision tree, bagging and boosting), where two specific types of boosting methods are also analyzed.

  - **Decision tree (DT)** is expressed in a recursive partition of the input data in a tree-like structure (Breiman, 2001). The trees are constructed beginning with a root and this is split down to the leaves. The nodes represent 'tests' based on the input features, a branch for each possible outcome, and the leaves for the final class label. DT algorithms develop these 'tests' or split condition, at each node such that the error of class labels is minimized and a meaningful relationship between a class and the values of its features can be captured (Quinlan, 1986). Several splitting criteria are used to evaluate the effectiveness of the split, in which Gini index and Information gain are commonly used (Tangirala, 2020). The tuning parameters of this classifier include number of splits and split criterion.

  - **Bootstrap Aggregation (Bagging)** is a simple and powerful ensemble technique. The model is made by building many decision trees by random sampling with replacement, or bootstrapping, from the original data set. The class of the test data set is determined by majority voting among the outcome of these trees. This reduces the the sensitivity to a particular choice of the training set (Breiman, 1996), i.e., it reduces variance and over-fitting. The tuning parameters include number of splits, number of learners, and learning rate.
  - **Adaptive Boosting (Ada Boost)** is an ensemble classifier which makes use of boosting technique. Boosting technique attempts to iteratively improve the model and reduce the error of the combination of 'weak' learning algorithms (Freund and Schapire, 1999). AdaBoost in particular iteratively updates the weighted classification error depending on the mis-classified data points. In contrast to the bagging classifier, boosting method increases the complexity of the model to primarily reduce the bias (Breiman, 1996), that is, to reduce any under-fitting in the training data. The tuning parameters include number of splits and number of learners.
  - **RUS Boost** is a boosting method, often used specifically for when classes in training data is imbalanced. The method applies random under-sampling (RUS) which is a technique that randomly removes samples from the majority class to improve the classification performances especially for skewed (the classes in the data are unevenly represented) data sets (Seiffert et al., 2008). This method can be promising due to the smaller number of lead classes compared to sea ice class in the data set. The tuning parameters include number of splits, number of learners and learning rate.

- **Artificial Neural Network (ANN)**
  ANNs make use of a system of interconnected artificial neurons (processing units), which are inspired by the process of neurons in biological brains. The system in ANN is able to modify its internal structure with respect to the objective function (Grossi and Massimo, 2007). Each of the neurons has its own input and output, which allows them to communicate with other neurons and the environment. Each neuron also has a function which transforms the global input to output. These neurons can be organized into number of layers, depending on the data. Normally, the input layer has the amount of input variables, which transfers information to one or more hidden layers in the ANN, then to the output layer provides the results (Grossi and Massimo, 2007). ANN adaptively learns through the training samples and it is known to have a very high tolerance to noise and incomplete data (Jensen, 2005). The tuning parameters include number of layers, layer size, and the activation function.

- **Naive Bayes Classifier (NB)**
  Bayesian classifiers formulate the conditional probability of each feature given its class label. The samples are then classified by applying Bayes' theorem to compute the probability given a particular input of the features to predict the class with the highest posterior probability (Friedman et al., 1997). Naive Bayes (NB) classifier in particular is a constrained form of a general Bayesian network as it assumes a strong probabilistic independence; all input features are conditionally independent to each other for a given class label. This independence between the features is untrue in many real life problems, yet NB classifier performs excellently in many applications (Friedman et al., 1997; Zygmuntowska et al., 2013). The tuning parameters include the predictor distribution.

- **Linear Discriminant (LD)**
  In Linear Discriminant Analysis, the original feature space dimension is reduced by projecting it to a lower dimensional feature space. Then the algorithms aims to find an optimal set of discriminant project vectors such that classes separability is maximized (Qin et al., 2005). If the predictors have a singular covariance matrix, a diagonal covariance structure shall be used instead of using the covariance matrix (MATLAB, 2020).

- **Support Vector Machine (SVM)**
  Support Vector Machine (SVM) is the most widely used kernel learning algorithm, which the original covariate are non-linearly transformed into a higher dimensional feature space. Then, an optimal hyperplane which separates and distinctly classify the data points are found (Xu et al., 2014). Different kernel functions can be used in SVM implementation, namely, linear, polynomial, and Gaussian radial basis function (Savas and Dovis, 2019). The tuning parameters include kernel function, kernel scale

(scaling parameter for the input data), and box constrained level (penalty factor for misclassification) (MATLAB, 2020).

- **K-Nearest Neighbours (KNN)**
  K-Nearest Neighbors algorithm (KNN) finds the K number of closest data points that are closest to the test point. Then the class of the testing point is decided based on a majority vote by the classes of the K closest training data (Shen et al., 2017). In this study, KNN is also used for applying unsupervised algorithms to the test data. The tuning parameters include number of neighbors (K) and distance metrics which provides a measure of distance between a pair of points.

2.2.3.2 Unsupervised learning classifiers

Unsupervised machine learning algorithms do not require the input data to be labeled. They cluster the data set based on their natural similarities described by the waveform features. During the training phase of unsupervised algorithms, the waveforms are grouped into number of clusters and the user must label a class to each cluster. Then, KNN is applied to the testing data such that each test waveform finds its closest labeled clusters. This testing approach is adopted from the work of Dettmering et al. (2018) and Müller et al. (2017).

For this study, the following unsupervised learning classifiers are adopted.

- **K-medoid**

  K-medoid classification is a partitional cluster algorithm for clustering unlabeled data into K clusters, based on their different feature properties. In this algorithm, a representative element of a cluster is chosen by looking for an element which has the minimal dissimilarities to all the elements in the cluster; this is called the medoid of the cluster (Kaufman and Rousseeuw, 1987). The selection of cluster size is the only tuning parameter.

- **Agglomerative Hierarchical Clustering (HC)**

  Hierarchical clustering is another unsupervised clustering technique, in which every iteration the similar clusters are merged together. Initially, the individual samples are considered as one cluster, and will merge with the other closest cluster until one cluster is formed. The linkage function which describes the distance between any clusters be selected by considering for example the smallest, furthest or average distance between the objects in the two clusters (Nielsen, 2016). The tuning parameters include cluster size and the linkage function.

2.2.3.3 Thresholding classification

Thresholding method simply sets threshold values to the waveform features (see Table 2.1) to classify the samples. This empirical method was commonly used in the earlier studies for lead detection (Laxon et al., 2013; Peacock and Laxon, 2004; Rose et al., 2013; Schulz and Naeije, 2018). The thresholding values are typically selected based on theoretical values or can be found via solving an optimization problem to maximize the accuracy (Wernecke and Kaleschke, 2015). This study makes use of the latter, finding the optimal thresholding values by means of random search. Each of the waveform features distributions were inspected to select the initial range of values, then random search was used to find the optimal combination of the threshold values. See Appendix A.2 for more details.

2.2.4. Validation Data Generation

There have been a significant number of publications devoted to image segmentation of remotely sensed images (Shepherd et al., 2019). In the context of lead detection, Passaro et al. (2018) for example proposed a dedicated adaptive threshold algorithm to create binary images from Sentinel-1A SAR images, allowing for an automatic distinction between leads and sea ice. This paper makes use of K-means clustering, which is a very effective and simple method for detection analysis (Hamada et al., 2019).

The study areas are divided into a number of smaller sections, in which the image segmentation is individually applied to each of these sections using K-means clustering. These small sections are 0.6 ° in longitude and takes ± 0.05 ° in latitude from the satellite track within the specified longitude range. Image segmentation performs better for smaller sections within the study area as it will be more sensitive to local radiation

differences. Therefore, the results will not be affected by for example the different lighting conditions of the extreme ends of the image or the presence of large clouds. Then, K-means algorithm with cluster size of two (K= 2) is generated to produce a binary image, dividing the image to brighter and darker areas, with each pixel depicting either ice or lead. The algorithm considers the radiance values of each of the pixels and allocate them to the nearest cluster, while keeping the total distance to the centroids of the clusters minimal. Once the image segmentation is completed, each altimetry point is assigned with a class that is determined by a majority vote of the three closest pixels to the corresponding point. Three pixels are considered to determine the class label since altimetry returns may show surface properties from the surrounding pixels, especially when the altimetry measurement point is lying on the edge of a pixel. Furthermore, in order to make sure that the altimetry measurement point in the edge of the sections have enough reliable pixels to determine its class, there is an overlap of 0.1° in longitude between these sections.

Figure 2.3 shows an example of the segmented binary image (left) and its original image (right). This method allows for an efficient and flexible application to different study areas since the image segmentation relies on only the local radiance differences. This is advantageous as it can be applied to any study areas with different lighting conditions without any adjustments.

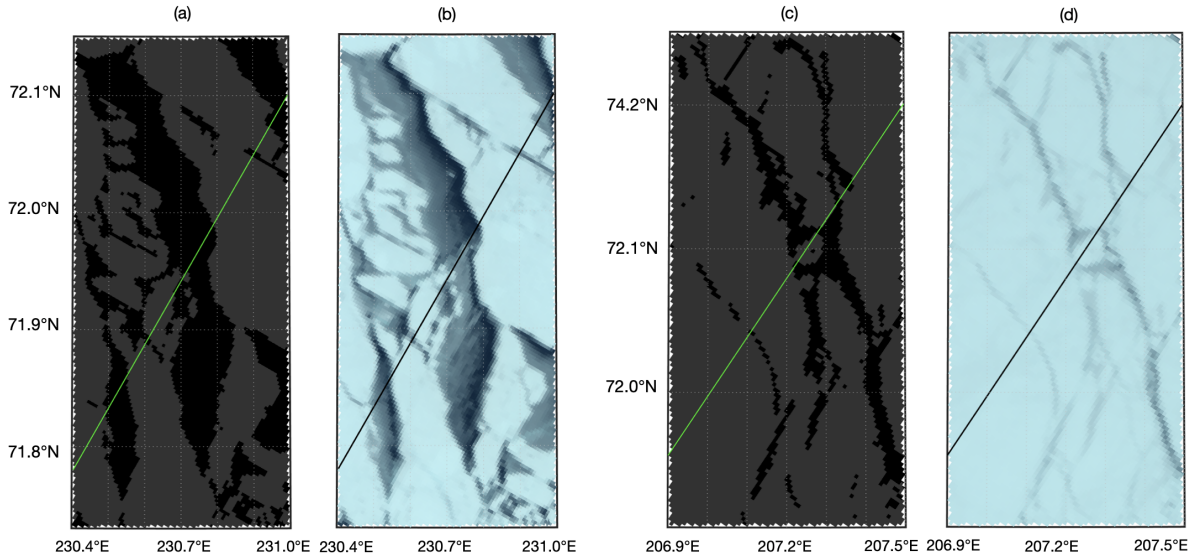


Figure 2.3: Examples of binary images (a,c) after the image segmentation scheme and their original OLCI images (b,d). The green/black line shows the Sentinel-3A ground track. The optical image (b) is taken on 15/04/2018 and (d) on 13/04/2019.

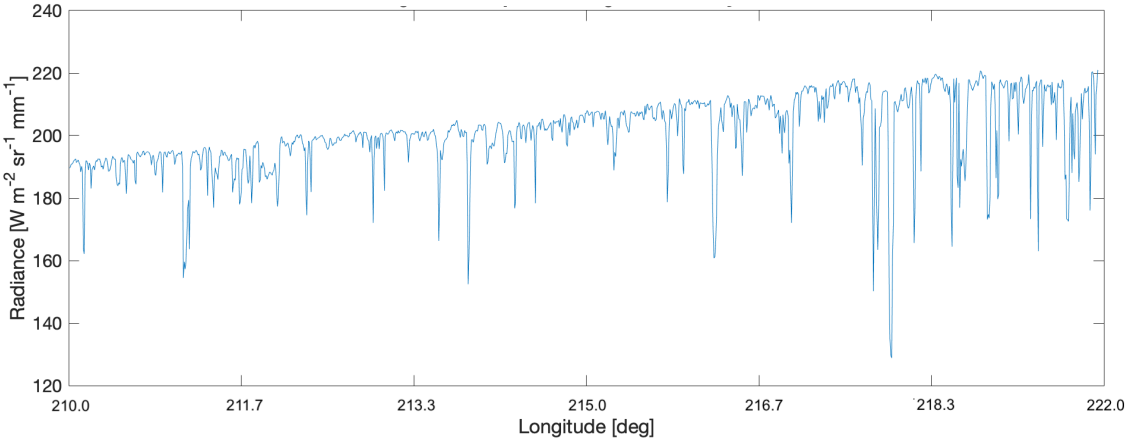


Figure 2.4: Example of radiance change of OLCI pixels along the altimetry track of Sentinel-3A. Data from 15/04/2018.

Though the image segmentation method generally picks up the shapes of the leads correctly, there could be mis-labeled pixels due to the irregular intensity of the radiance or the presence of small clouds. This can result in undetected leads or over/under estimation of their sizes. Therefore, another method was added to improve the reliability of the ground truth. The second method uses the changes of radiance along track of the satellite nadir, which shows a constant fluctuation (see Figure 2.4). The model first identifies all the maximum and minimum peaks in the radiance series. Then it relies on a certain percentage threshold that is set (in this paper 2% was used), and if the minimum peak is smaller than this percentage of the maximum peak, the points within the dip are considered to be leads. Addition of this second method is especially useful to properly label the points that are close to the edge of the leads. It also helps to discard points that image segmentation method was sensitive to (e.g. presence of small clouds). Consulting the results from both of the methods, the altimetry points were only labeled if both methods agreed to be either ice or leads.

2.3. Experimental Set Up

2.3.1. Division of Data

The total data set is divided into different train/test sets such that the classifiers can be assessed for different purposes and a complete understanding of the behaviour of the different classifiers can be gained. The specific description of these data sets are summarized in Table 2.2. Note that some data sets differ depending on the algorithm that are analyzed (depicted with SUP, UNSUP, and THR for supervised-, unsupervised machine learning and thresholding methods, respectively). General performance during the winter (March/April) seasons is our primary interest, and this is analyzed using the D-01 data set. Temporal and spatial biases in the winter seasons are also studied with D-02 and D-03 data sets, respectively. The summer performances are studied with D-04 data set, and the influence of the presence of off-nadir leads are studied with D-05 data set.

It must be noted that the purpose of analyzing the temporal and spatial biases is to understand whether the trained data could be applied to test data from another year or another study area. This has a practical relevance to the supervised machine learning algorithms because ground truth data might not be available for those regions/dates. Therefore, this specific data set division is only made for the supervised learning classifiers. In contrary, unsupervised learning and thresholding classifiers do not require any labeled data, therefore the algorithm is simply applied to the test data set, such that their performances can be compared to supervised learning classifiers.

Table 2.2: Table showing division of total data depending on their purposes.

Name	Description	Purpose
D-01	Randomly takes 80% data for training and 20% data for testing from the March/April data set from 2017 to 2020.	Evaluate the general performance of the classifiers.
D-02	**SUP:** uses 2017 March/April data set as training data to test the rest of March/April data set from 2018 to 2020. **UNSUP/THR:** Apply algorithm to the test data set.	Analyze possible bias in the yearly data.
D-03	**SUP:** Uses data from below 80°N and between 150°E - 240°E as training to test data lying above 80°N and between 120°E - 150°E. **UNSUP/THR:** Apply algorithm to the test data set.	Analyze possible bias in the study areas.
D-04	Randomly takes 80% data for training and 20% data for testing from the summer data set from 2020.	To evaluate classifier performances on summer seasons.
D-05	Manually selected study areas with off-nadir leads presence (data from 31/03/2017, 07/04/2017, 14/04/2020, 15/04/2020). Off-nadir leads are manually labeled as an additional class. All data are trained and 5-fold cross validation is applied for assessing the results.	To evaluate the influence of off-nadir leads on the classification performances.

2.3.2. Classification Assessment

The classification performances can be assessed in different ways. This paper focuses on using the overall accuracy, True Lead Rate (TLR), False Lead Rate (FLR) and Receiver Operating Characteristic (ROC) graphs to assess the different classifiers.

Firstly, the overall accuracy is simply defined as the number of total correct classification over the total number of data. The majority of this paper deals with binary classifications. The binary classification outcome can be one of the following. If the classifier correctly predicts a sample to be ice, it is called a True Ice. If ground truth suggests otherwise this is a False Ice. Similarly, if the classifier correctly predicts a lead, it is a True Lead and otherwise it is a False Lead. Therefore, the overall accuracy can be also described as (True Lead + True Ice) / total number of samples.

True Lead Rate (TLR) represents the number of correctly detected leads over the number of ground truth lead samples (Equation 2.1) and False Lead Rate (FLR) is the number of leads mis-classified as ice over the number of ground truth ice samples (Equation 2.2). Many studies use FLR and TLR to assess the performance (Wernecke and Kaleschke, 2015), as these parameter effectively describe what must be optimized in a classifier; TLR must be maximized while FLR being minimal.

$$\mathrm{TLR} = \frac{\mathrm{TrueLead}}{\mathrm{TrueLead} + \mathrm{FalseIce}} \tag{2.1}$$

$$\mathrm{FLR} = \frac{\mathrm{FalseLead}}{\mathrm{FalseLead} + \mathrm{TrueIce}} \tag{2.2}$$

The assessment parameters for ONL analysis (D-05) are slightly adjusted as it involves three classes. First, the True ONL Rate is introduced by computing the correctly predicted ONL over the total ONL from the ground truth. TLR still holds from the binary classification; it is the correctly predicted leads over the total ground truth leads. Finally, False Lead is computed by adding the total falsely predicted leads (predicted as leads but ground truth shows either ice or ONL). To be consistent when comparing the 3-class and the binary models the number of False Lead will be divided by the total number of waveforms such that a % fraction of mis-predicted leads can be computed. This value is called as $\mathrm{FLR}_{\mathrm{ONL}}$ in the context of ONL analysis. $\mathrm{FLR}_{\mathrm{ONL}}$ is the value that is most concerned in the ONLs analysis as it is aimed to reduce the falsely predicted leads by introducing an additional class.

Finally, Receiver Operating Characteristic (ROC) graph is used for visualizing and evaluating classifiers based on their TLR and FLR values. ROC graphs effectively show the trade-off between the TLR and the FLR. Also, its area under the curve (AUC) provides the overall performance measure. AUC value of 1.0 suggest a perfect classifier, whereas a diagonal line between (TLR, FLR) of (0%,0%) and (100%,100%) with an AUC value of 0.5 is equivalent to random-guessing. It must be noted that it is possible for a classifier with high AUC value to perform worse in specific regions of ROC curve compared to a low AUC classifier (Fawcett, 2006).

ROC graphs can be generated in different ways. Supervised learning classifiers generates a 'score' or posterior probability for each sample, which describes the probability of that sample belonging to one of the classes. A certain threshold value is then applied to the score to generate predictions. ROC graphs effectively show the classification results depending on the choice of this thresholding value (Fawcett, 2006). On the other hand, ROC graphs for unsupervised learning and thresholding classifiers can be generated by changing the dominant input parameters. This approach has been taken for example by Dettmering et al. (2018) and Müller et al. (2017), where the cluster size was changed to generate a ROC graph for K-medoid classifier.

As falsely detected leads result in a bias in SSH estimation, low FLR values are desired. Therefore in this paper, classifiers are evaluated with a focus in the low FLR region (<5%).

2.3.3. Experiment Overview

An overview of the experiment conducted in this paper is presented as a flowchart as see in Figure 2.5. Each item in the chart are discussed briefly in the following.

Firstly, the classification methods are defined by finding the optimal input settings including waveform features, (hyper)parameters and thresholds. Previous studies (e.g. Lee et al., 2018; Shen et al., 2017) have demonstrated the profound impact of using different combination of waveform features on the classification performances. Other input (hyper)parameters of the classifiers also have influence on their performances. An iterative procedure was taken to define these input parameters and settings. The (hyper)parameters were

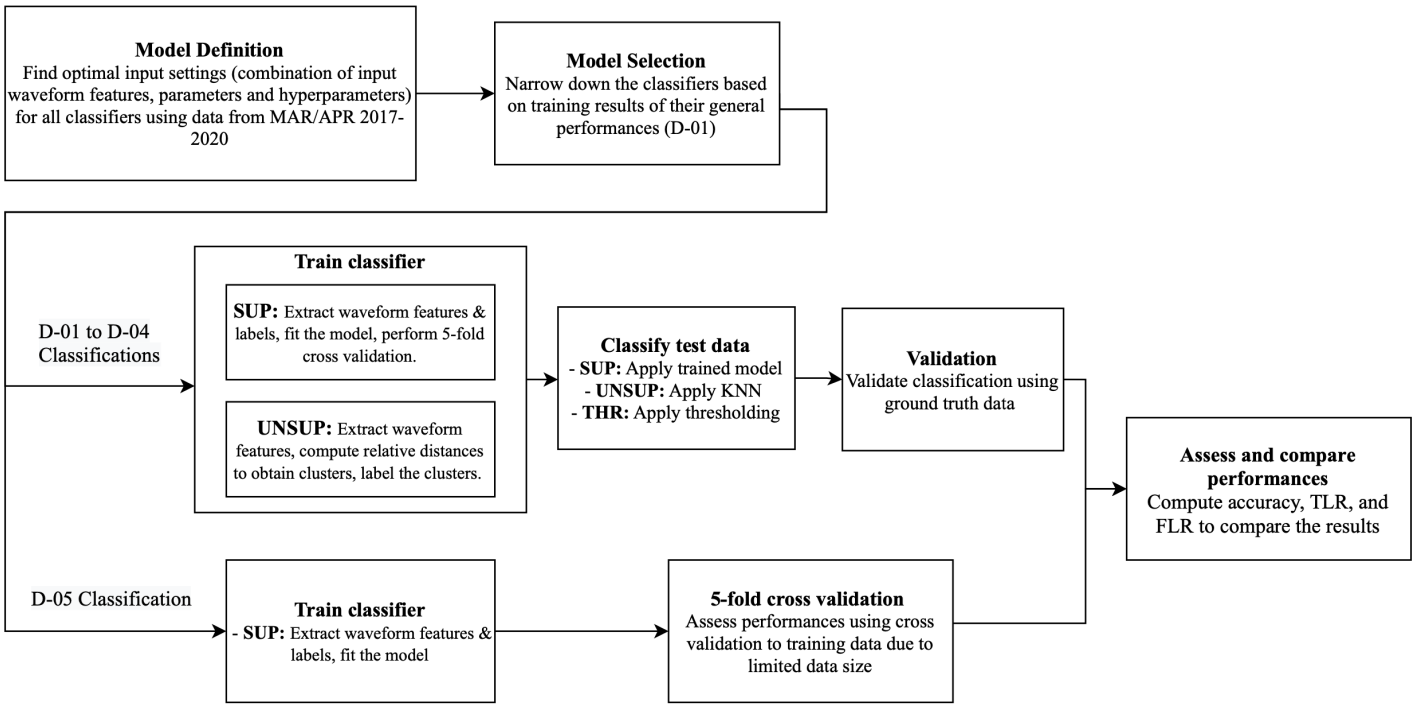


Figure 2.5: Flowchart showing the workflow from model definitions of the classifiers to assessing the classification performances for different study cases

first optimized by using all of the waveform features. Then, using these optimized (hyper)parameters, a combination of waveform features were selected by running the classifiers with one waveform feature at a time to understand the individual classification potential. For a fair comparison between the classifiers, this paper aims to select to a combination of waveform features which is beneficial to most of the classifiers and adhere to it when comparing the classifiers. Once the waveform features are selected, the classifiers are tuned again to maximize their performances. This iterative procedure is concluded as the relative performances of the classifiers stabilized. These input settings (combination of waveform features, parameters and hyperparameters) are kept constant throughout the study.

Then, a preliminary comparison of the classifiers is conducted based on their general performances, using the training results of D-01 data set. This stage aims to reduce and narrow down the number of classifiers to be compared, if some classifiers are found to clearly perform worse than the others.

The analysis of general classification performance in winter seasons, their spatial and temporal biases, and the seasonal analysis (D-01 to D-04) undergo similar training and testing process. During the training phase, the supervised machine learning algorithms (SUP) are trained with labeled data set provided by the OLCI ground truth. This study makes use of an internal 5-fold cross-validation technique to reduce overfitting. Cross-validation technique takes a full advantage of the available training data by repeatedly producing training and testing sets within the data (Hand, 1997). In a 5-fold cross validation, the data set is randomly divided into five subsets, of which four of them will act as the training set and the remaining one will act as a testing set to estimate the performances. This is repeated five times, and the final performance score is found by averaging of the five testing sets (Xu et al., 2014). To conclude the training phase of the supervised machine learning classifiers, optimal operating points (OOP) of their ROC graphs are computed. OOP of a ROC is an optimal point when considering the trade-off between the true positive (TLR) and false positive (FLR). OOP is computed by first finding the slope S,

$$S = \frac{C_{FL} - C_{TI}}{C_{FI} - C_{TL}} * \frac{L}{I} \tag{2.3}$$

where C is the misclassification cost of false and true lead (FL, TL), false and true ice (FI, TI), which can be adjusted depending on which error should be penalized more. This study uses misclassification cost of $C_{FL} = C_{FI} = 1$ and $C_{TL} = C_{TI} = 0$, making the slope simply $S = \frac{L}{I}$. L and I are number of the 'actual' lead and

ice samples seen by the ground truth ($L = TL + FI, I = TI + FL$). This slope is used to create a straight line from the upper left corner of the ROC plot, until it intersects a point in the ROC curve, depicting the optimal operating point (Metz, 1978). Supervised machine learning classifiers then apply the trained model and the threshold found by OOP to the testing data set.

The training phase for the unsupervised machine learning algorithms (UNSUP) consists of applying the algorithm to find the clusters, then users labelling a class to each cluster. During the testing phase, K-Nearest Neighbour algorithm is applied to the testing waveform data such that each sample finds its closest labeled cluster.

The off-nadir lead (ONL) analysis (D-05) takes a unique approach. This analysis uses a manually generated ground truth including an additional class, "off-nadir leads". This was done by only converting the original sea ice class of ground truth to ONL class, if ONL was present. The point was identified as ONL only if the lead was within the across track footprint of SRAL (1.64km) (Kittel et al., 2021). In this analysis, the results of 3-class model are compared among the different classifiers but also to the results obtained by the binary model. The ground truth generation resulted in the 3-class model to have 2,537 ice class, 590 lead class, 151 ONL class, and a total of 3,278 waveforms. Whereas the binary class contains 2,683 ice class, 590 lead class and a total of 3,273 waveforms. Because of the limited data size, a 5-fold cross validation is used to evaluate their performances. Note that this analysis is conducted only for supervised machine learning as it aims to find unknown waveform features, and this is not possible with the unsupervised learning since a user-based selection must be made. It is expected that by adding another class for ONLs, points which were previously falsely detected as leads (but ground truth is ice with presence of ONL) in the binary model can be predicted as ONL class. If this is successful, the number of false lead can be reduced.

2.4. Results and Discussions

2.4.1. Ground Truth Data

Applying the method described in Section 2.2.4, ground truth data for all the study areas presented in Section 2.2.2 have been generated. A total of 14,723 waveforms were generated, in which 11,762 sea ice and 2,961 leads were found.

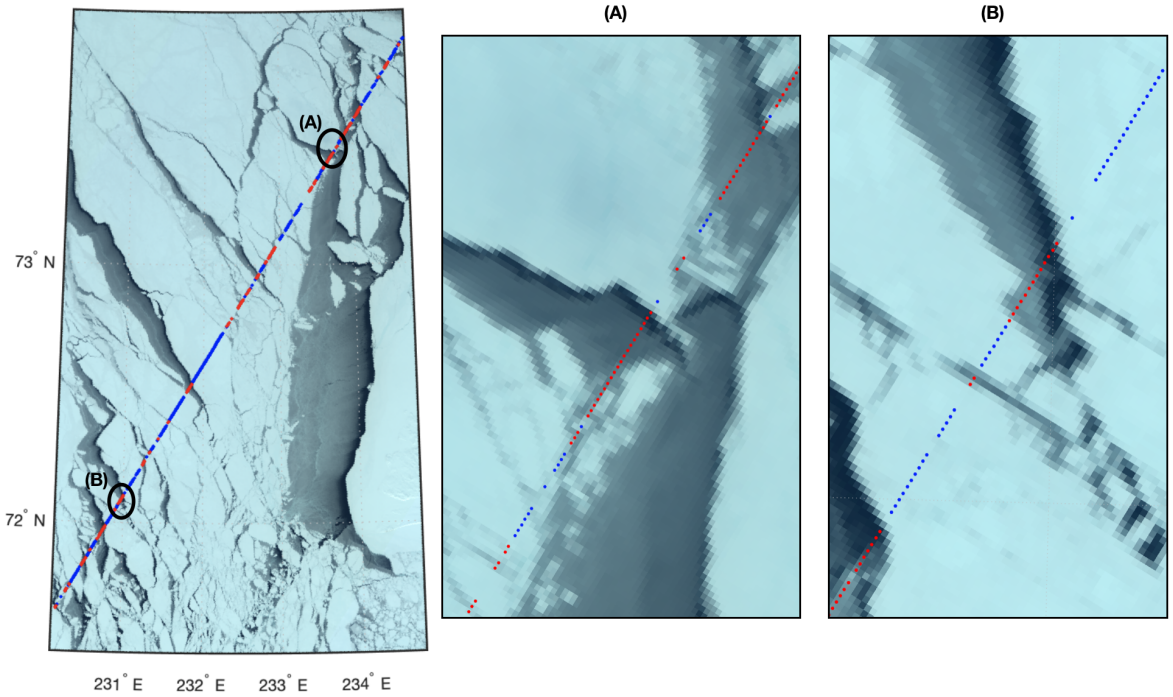


Figure 2.6: Validation data generated by the presented validation process, combining the image segmentation method and the analysis radiance changes. Red points depict leads whereas the blue points depict sea ice. (A) and (B) shows the zoomed view of the areas shown in the left image.

An example of the final product of the ground truth data is shown in Figure 2.6, where red and blue points depict leads and sea ice, respectively. It can be seen that the points are successfully labeled according to the OLCI image. The zoomed figures (A) and (B) also show further details of the image and the corresponding labeled data. The track show some sparse areas; these are the points which were rejected and unlabelled due to either the presence of (small) clouds or points that the two ground truth data generation methods did not agree with the class label (see Section 2.2.4 for more details). Through the ground truth generation process, a total of 3,519 waveforms out of the original 18,242 waveforms have been rejected. These generated validation data will be used for training the supervised learning machine learning classifiers and also be used as the validation data to assess the classifiers.

2.4.2. Selection of Waveform Features

As described in Section 2.3.3, the individual classification potential of the waveform features were studied. The resulting accuracies of all classifiers trained with each of the waveforms are given in Table 2.3. The produced accuracies are generally high, with most of the waveform features producing more than 80% accuracy for most of the classifiers. However, the unsupervised learning classifiers (K-medoid and HC) obtained substantially lower accuracies when using LeW, sigma0, PPL, PPR, and NrPeaks. HC classifier also did not perform well when using kurtosis. It is seen that using WW, PP, PPloc, skewness, and MAX, produces high accuracies for all of the classifiers. Though TeW has achieved a high average accuracy, it has not been selected due to the relatively low accuracy produced by the K-medoid classifier.

Additionally, most of the supervised learning classifiers produced very similar results when using NrPeaks (79.8%). In fact, the classifiers which produced this accuracy predicted all of the waveforms to be sea ice, i.e. obtained 0% TLR. Because NrPeaks can only have discrete integer values, it is not a suited feature for machine learning algorithms. Previous studies which used NrPeaks were thresholding based classifications (Bij de Vaate, 2019; Schulz and Naeije, 2018).

Based on this result, the waveform parameters which will be utilized in the following analyses are *WW, PP, PPloc, skewness, and MAX*, as they resulted in relatively high overall accuracies for all of the classifiers. Though other waveform feature combinations may be beneficial for some of the classifiers, it has been seen that the relative performance of the classifiers remains approximately constant (see Appendix A.3 for more details).

2.4.3. Tuning Machine Learning Classifiers

Once the waveform features to be used in this study were selected, the classifier models were finalized by selecting their algorithm input settings. Though this tuning process varied per classifier, ROC graphs were mainly used to help visualize the correctness of the classifiers. Table 2.4 shows the list of (hyper)parameters used in each of the classifier.

2.4.4. Training Results and Model Selection

The preliminary assessment of the classifiers were done by analysing the training results of their general winter performances, given by the D-01 data set. The result of the training phase are evaluated by the 5-fold cross validation and the resulting ROC graphs are shown in Figure 2.7. For visual clarity, the ROC curve of the classifiers are divided into tree-based classifiers (left) and non tree-based classifiers (right) in Figure 2.7. From the tree-based classifiers, it can be seen that Ada Boost, Bagged and RUS Boost classifiers show very similar results throughout the curve. However, DT classifier show lower TLR values throughout the curve, showing that it performs worse than the other tree-based classifiers. Consulting the ROC curves obtained by the non-tree based classifiers, it is seen that most classifiers produce very similar results in the region where FLR values are 3 to 5%. However, for lower values of FLR, the classifiers performances differ. SVM classifier clearly performs worse compared to the rest, producing very low TLR values for a given FLR value. KNN and ANN classifiers have higher TLR values for the very low FLR regions, outperforming the NB and LD classifiers. Given this preliminary analysis of the supervised learning classifiers, this study will focus on *Ada Boost, Bagged, RUS Boost, KNN, ANN, NB* and *LD* supervised learning classifiers in the further analysis, as DT and SVM clearly showed worse performances in this preliminary analysis.

Table 2.3: Training accuracies (in %) of classifiers trained with a single waveform parameter. The bold numbers show the waveform parameters with the best performances.

	AdaBoost	Bagging	KNN	SVM	DT	NB	LD	ANN	RUSBoost	K-medoid	HC	Average
MAX	88.76	88.76	88.84	88.96	88.18	88.92	88.93	88.92	88.76	83.41	82.86	**87.85**
Kurt	87.84	87.78	87.83	88.06	87.06	88.07	88.05	88.07	87.85	86.57	49.50	84.55
Skew	89.04	89.05	88.95	89.08	88.40	89.05	89.06	89.06	89.00	87.59	88.39	**88.81**
PP	90.79	90.82	90.90	91.01	90.25	90.97	90.98	90.98	90.82	90.68	90.67	**90.82**
WW	91.02	91.00	91.05	91.10	90.99	91.10	91.10	91.10	91.03	90.93	90.84	**91.03**
LeW	85.82	85.82	85.82	85.82	85.82	79.88	85.82	85.82	85.82	20.91	28.46	75.14
TeW	88.38	88.38	88.38	88.38	88.38	88.38	88.38	88.38	88.38	71.28	86.45	86.79
sigma0	81.58	81.97	81.36	81.50	81.63	81.51	81.50	81.52	81.89	23.72	20.60	71.69
PPL	86.35	86.45	86.50	86.52	85.65	86.52	86.55	86.47	86.38	43.60	20.18	77.30
PPR	87.53	87.56	87.72	87.79	86.96	87.76	87.80	87.82	87.61	70.48	20.17	80.58
PPloc	90.32	90.48	90.41	90.49	89.89	90.49	90.49	90.46	90.50	89.79	90.26	**90.34**
NrPeaks	79.8	79.8	79.8	79.8	79.8	79.8	79.8	79.8	44.3	59.10	70.22	73.8

Table 2.4: Final input settings of the classification method.

Classifier	**Settings**
DT	Maximum number of splits = 100, Split criterion = Gini index
AdaBoost	Maximum number of splits = 100, Number of learners = 30, Learning rate = 0.1
Bagging	Maximum number of splits = 11777, Number of learners = 30
RUSboost	Maximum number of splits = 20, Number of learners = 30, Learning rate = 0.1
ANN	Fully connected layers =1, Layer size = 10, Activation = Relu, No regularization
KNN	Number of neighbors = 100, Distance metric = Euclidean
LD	Covariance structure = Full
NB	Predictor distribution = Gaussian
SVM	Kernel function = Gaussian, Kernel scale = 0.56, Box constrained level = 1
Kmed	Cluster size = 15
HC	Cluster size = 40, Linkage = farthest distance
Threshold	Classify as leads if MAX>4000, PPloc > 0.55, WW<40, PP>0.3, skew>7. Else: sea ice.

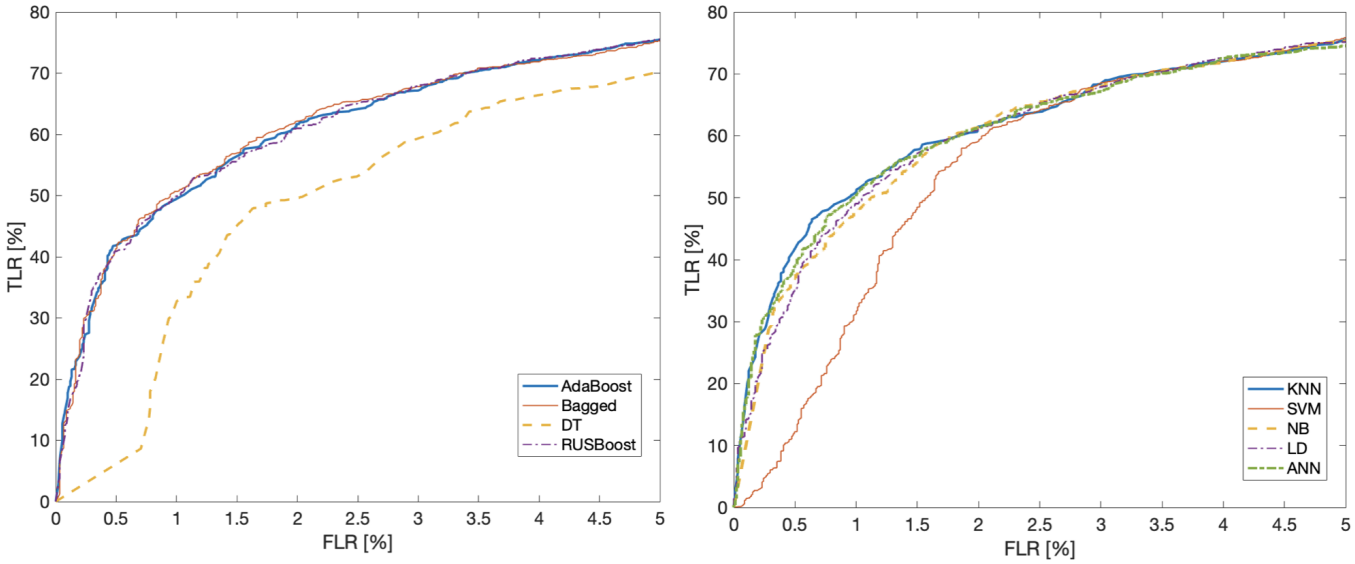


Figure 2.7: ROC graph of 9 supervised learning classifiers after 5-fold cross validation during the training phase. Tree based algorithms (left) and the other algorithms (right) are plotted separately for better visual interpretation.

As described in Section 2.3.3, unsupervised learning classifiers produce clusters of waveforms, where the users must manually label the cluster with a class. An example of the resulting cluster of K-medoid cluster with K=15 is shown in Figure 2.8. It is well known that waveform returns of leads show single-peak waveforms with high power and narrow shape, whereas returns from sea ice have more diffuse scattering, resulting in weaker power and show no clear peaks (Müller et al., 2017). Based on this knowledge, the users can label the clusters. However, it must be noted that some clusters show rather ambiguous returns such as in cluster number 10 and 12. Their dominant feature is narrow specular returns, but weak and noisy returns can also be observed. If a very conservative result is desired, one must select clusters that are clearly showing lead class, such as cluster number 2, 11 and 13. The more clusters that seem 'ambiguous' are selected, higher TLR may be achieved, but at a cost of increasing the FLR. This could be one of the biggest disadvantage of unsupervised learning in the context of this application, that the outcome is highly dependent on human decision-making.

For unsupervised learning classification methods, ROC graph has been produced by changing the number of cluster sizes as it is the dominant influential parameter. For both K-medoid and hierarchical clustering classifiers, their resulting TLR and FLR are shown in Figure 2.9, depending on the input clustering size. One can observe that the results differ significantly depending on the cluster size for both classifiers. If a small cluster size is selected, the number of samples per cluster increases and the probability for correct classification decreases (Dettmering et al., 2018), suggesting that the clustering is too coarse. This is for example seen for K=5 for both of the classifiers, where it resulted in very high FLR values, indicating an over-detection of leads. This in fact is in line with the findings of Dettmering et al. (2018). For larger cluster sizes (K>5), the results change less significantly in the region of FLR values between 3 and 6 %, and there is no clear correlation between the cluster sizes and the classifier performance.

Interestingly, HC classifier produces exact same results for some of the clustering sizes (e.g, K=5,10 or K=15,20,25,30), as seen in Figure 2.9. As HC algorithm creates a cut on a fixed dendrogram to produce the given number of clusters, it is possible that after manual selection of the clusters, the selected waveforms are exactly the same as to other cluster sizes. This results in producing the same FLR and TLR values even with different input of number of clusters, which makes this classifier inflexible. On the other hand, K-medoid classifier shows more diverse results with different range of FLR and TLR.

Finally, considering the AUC of the ROC curves presented in Figure 2.9, K-medoid classifier show better results consistently obtaining higher TLR values compared to HC classifier. Therefore, *K-medoid classifier* will be the focus for the following analyses.

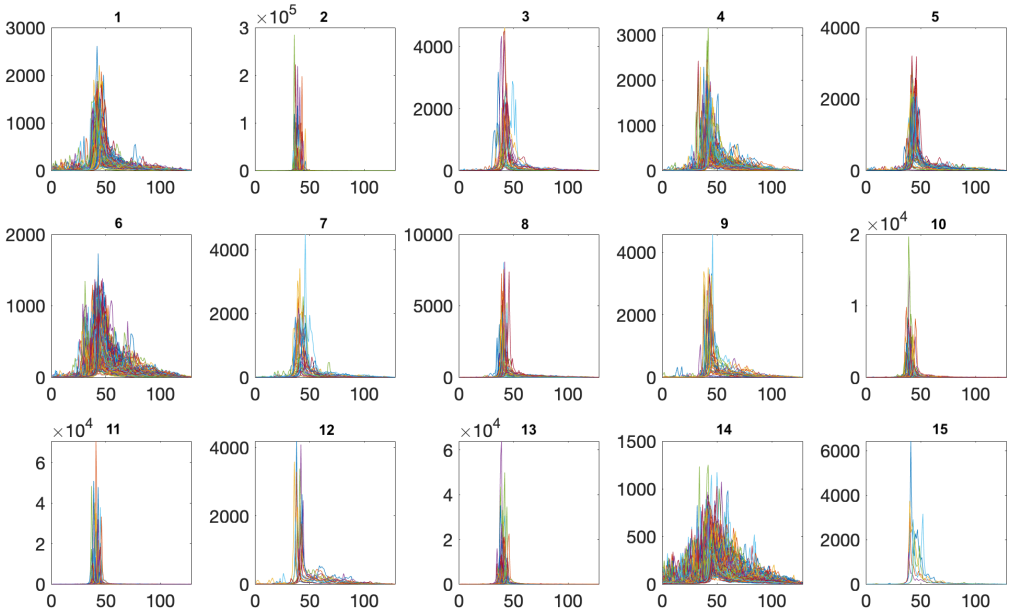


Figure 2.8: Example of waveform clusters provided with K-medoid classification, K=15.

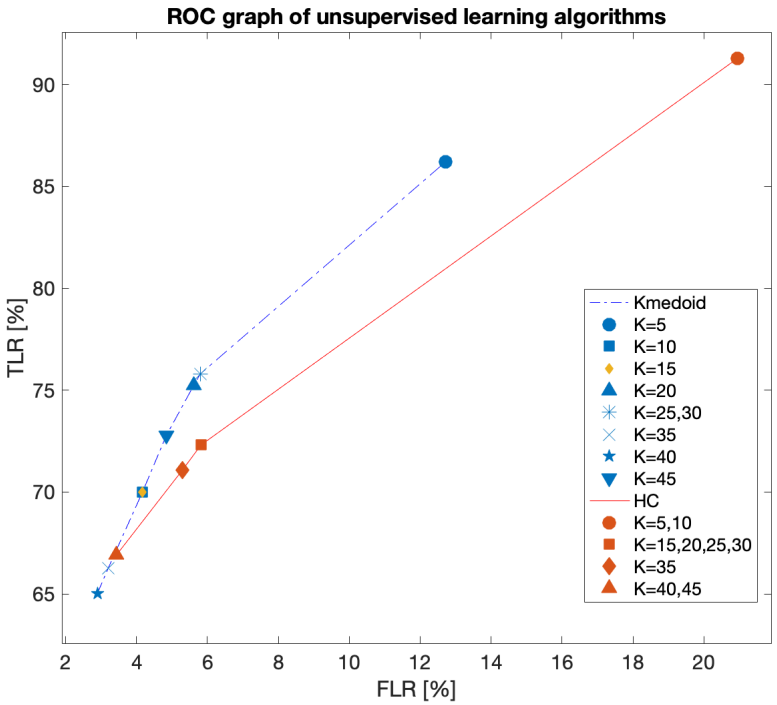


Figure 2.9: ROC graph of two unsupervised learning classifiers (K-medoid and Hierarchical clustering) with different number of cluster sizes, during the training phase.

2.4.5. Classification Performances during the Winter Months (D-01 to D-03)

The classification performances obtained with test data set of the winter months are provided in Table 2.5, showing their specific overall accuracy, TLR, and FLR values. These values are also presented in an ROC graph, as seen in Figure 2.10. Result from each classifier can be distinguished by the different shapes, whereas the black, blue and green colors depict the results of the general performance (D-01), the analysis of using training data from another year (D-02), and the analysis of using training data from different areas (D-03), respectively. As discussed in Section 2.3.2, points that are closer to the top-left corner (FLR=0% and TLR=100%) in an ROC graph are favored. To better understand and visualize the relative performances of the classifiers, the best performing points are connected, to create an optimal front.

Table 2.5: Accuracies, TLR and FLR (in %) of classifiers tested in different data sets to assess its general performance (D-01), influence of training with data form other year (D-02) or other study area (D-03), and using data set from summer months (D-04)

	D-01 (general performance)			D-02 (yearly bias)			D-03 (spatial bias)			D-04 (summer months)		
	Accuracy	TLR	FLR	Accuracy	TLR	FLR	Accuracy	TLR	FLR	Accuracy	TLR	FLR
AdaBoost	91.75	71.79	3.31	91.39	67.8	2.55	91.57	61.58	2.9	67.61	13.6	6.51
Bagged	90.86	68.89	3.69	89.92	63.97	3.41	91.23	63.42	3.65	65.9	33.3	18.49
RUSboost	89.68	83.25	8.73	87.52	86.77	12.29	90.34	79.19	7.6	58.24	61.84	43.49
ANN	91.51	70.06	3.31	91.3	68.66	2.88	91.65	62.08	2.9	67.47	29.82	14.5
KNN	91.85	72.65	3.39	84.86	64.18	9.83	91.99	63.76	2.81	67.9	11.84	5.25
NB	91.38	75.04	4.58	91.58	77.33	4.75	92.04	70.3	3.96	64.76	25.58	15.74
LD	91.82	67.69	2.2	91.48	66.9	2.19	91.83	57.4	1.82	68.62	16.74	5.56
Kmed	91.51	67.35	2.5	90.64	60.18	1.56	92.74	70.21	3.33	41.48	93.42	83.4
Threshold	89.78	55.38	1.69	89.84	56.59	1.62	91.08	52.52	1.82	49.57	81.14	65.55

In order to understand the relative performances, it is useful to see how close the points are to this optimal front. Furthermore, one could also compare the TLR of the classifiers which have similar FLR values. For example, consulting the general performance (D-01) results, it can be seen that the result obtained by KNN classifier outperforms AdaBoost, ANN and Bagged classifiers, as KNN obtains higher TLR than these classifiers, while they all obtain similar FLR values.

First, the differences between the classification performances among the different case studies are discussed. It can be seen that the general performance (D-01) of the classifiers are mostly closer to the optimal front, especially when compared to the other case studies (D-02 and D-03). For example the general performance results (D-01) of KNN and LD are part of the optimal front, while Ada Boost and K-medoid classifiers are also producing results very close to the optimal front. Interestingly, most of the supervised learning classifiers do not suffer significantly when they are trained with data set from another year (D-02). Classifiers such as NB and RUS Boost are even part of the optimal front, while LD also remains very close to it. However, KNN performs poorly when trained with data set from another year, as the resulting point is very far from the optimal front especially when comparing to its general performance result. Lastly, despite the the overall accuracies remained approximately the same from the general performance (see Table 2.5), the supervised classifier's result from the spatial bias analysis (D-03) generally moved away from the optimal front, especially when comparing to the other case studies (D-01, D-02). This suggests that the classifiers obtain worse performance; higher FLR and lower TLR values, when they are trained with data set from different study areas.

Through this comparison between the three case studies, it is seen that the supervised learning based classifiers perform better if the training data is from the same data set as the testing data. The effect of using training data from a different year does not necessary worsen the performances, but using data set from a different study area may worsen the performances slightly. Though the worsening of these performances are not significant, using K-medoid classifier would be favorable in case ground truth is not available, as its performance is not affected by the change of the training data set.

RUS Boost classifier obtains results that are close to the optimal front for the three case studies, however

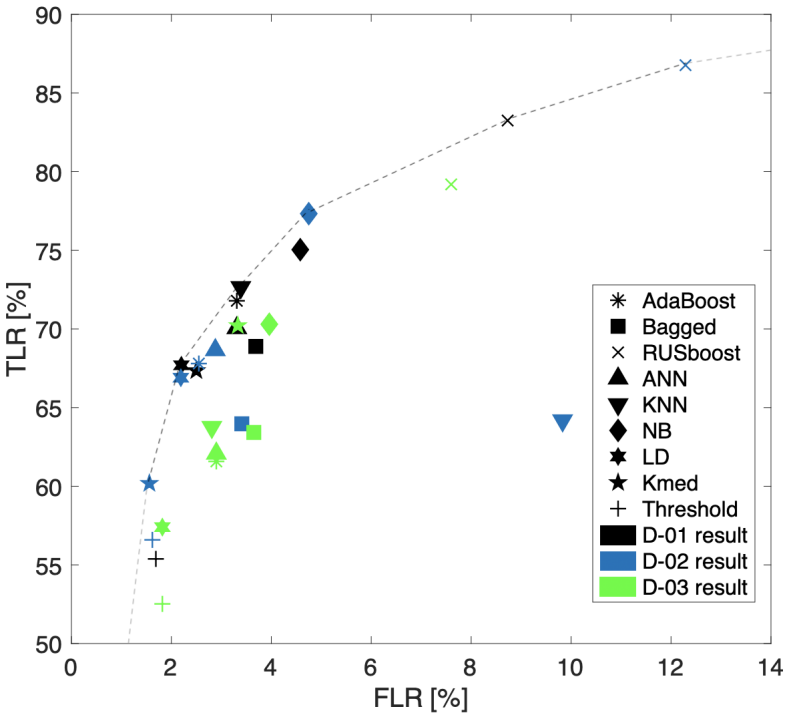


Figure 2.10: ROC graph showing the results of the classification performances in the winter months. Each classifier can be distinguished by the different shapes, whereas the black, blue and green colors depict the results of the general performance (D-01), the analysis of using training data from another year (D-02), and the analysis of using training data from different areas (D-03), respectively.

its very high FLR value is undesirable for lead detection for SSH estimation purposes. RUS Boost classifier obtains these high FLR and TLR values as this algorithm opts to increase the correct classifications for the minor class (Seiffert et al., 2008), leading to higher TLR but at the cost of increasing the FLR. It is also interesting that the thresholding method shows a very conservative results (low FLR). It has been seen that the threshold values derived by Laxon et al. (2013) tend to over-detect leads (Lee et al., 2016). However, as long as the ground reference is known, the thresholding classifier can be a very flexible classifier by consulting the class distribution for different waveform parameters and adapting different thresholding values. Furthermore from this result, KNN classifier produced very unpredictable results. While KNN classifier produced one of the best results for the general performances and showed marginal bias for spatial analysis (D-02), the performance for temporal bias analysis worsened significantly. Therefore, KNN classifier might not be the best suited classifier for lead detection as its performance could be significantly influenced by different data sets. Furthermore, it is seen from the ROC graph that threshold classifier and bagged classifier consistently performed worse compared to the others as they do not produce high enough TLR for a given FLR value, for all of the case studies.

It is not straightforward to conclude which classifier is best suited for lead detection in the winter seasons, however with the help of the different case studies, it is found that some classifiers are preferred over others for different circumstances. K-medoid classifier consistently performed well, producing results that were close to the optimal front for all of the case studies. Because k-medoid classifier does not rely on the ground truth label and only considers the altimetry data, any spatial and temporal biases that may be associated with supervised learning can be fully avoided. Therefore, if the ground truth data is unavailable for the testing area, using k-medoid classifier will be the most suited. If the ground truth is available, supervised learning classifiers such as AdaBoost, LD or ANN may be preferred over the k-medoid classifier, as their resulting points were slightly better. However, this distance is very minimal and the results may change depending on the data set.

2.4.6. Performance during the Summer Months (D-04)

The last three columns in Table 2.5 show the overall accuracies, TLR and FLR obtained by all of the classifiers for the summer months, described by the data set D-04. These values are also presented in an ROC graph, as seen in Figure 2.11. The figure also presents all the results from D-01 and D-03 (as shown in Figure 2.10) in gray markers, for comparison. The diagonal line in the ROC graph connecting (FLR,TLR) = (0,0) and (FLR,TLR) = (100,100) depicts a line for AUC=0.5, which is equivalent to random guessing (Fawcett, 2006) (see Section 2.3.2). Therefore, any points lying right-bottom compared to this line is considered to be worse than a random classifier.

The results of the classifiers all show very high FLR values (above the 5% limit). The classifiers either do not detect enough leads (too low TLR) or over-detect them (too high FLR). It is also immediately clear from the ROC graph that these classifiers are not producing results as accurately as they did for the winter seasons (D-01 to D-03). The points also lie very close to the diagonal line depicting a random guessing classifier.

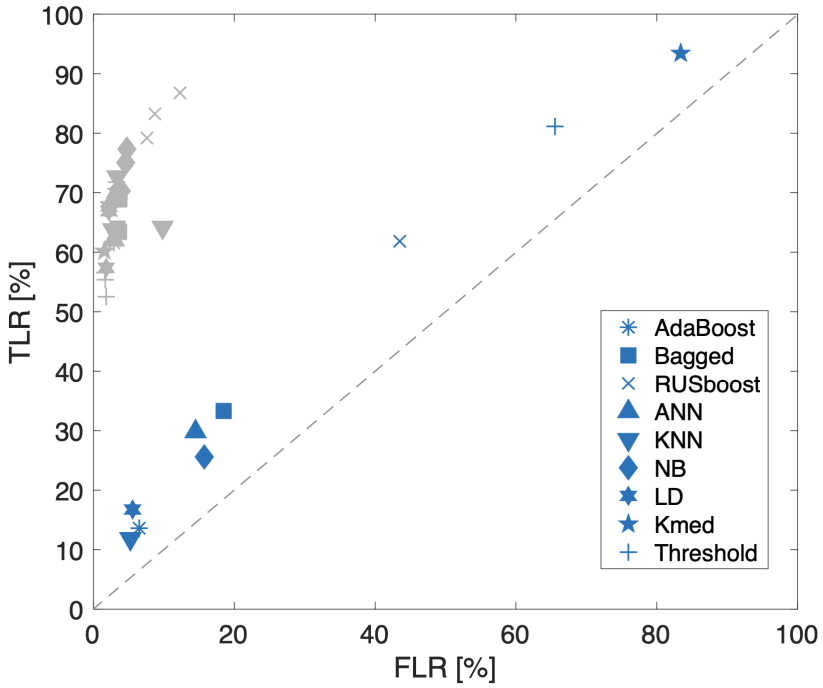


Figure 2.11: ROC graph showing the results of the classification performances in the summer months, in blue markers. Results from the winter months (D-01 to D-03) are also presented in gray markers, for comparison. Each classifier can be distinguished by its marker shape.

The worsening performances in the summer months is believed to be due to two reasons. Firstly, the majority of the altimetry waveform returns show specular returns even if the ground truth suggests that they are sea ice. The presence of melting ponds or smoothed layer on top of the ice due to the melting can be the cause of this increase in specular returns during the summer months. Secondly, the distinction between leads and sea ice from the OLCI images are more difficult during the summer. This is also thought to be caused by the presence of melt ponds and the layer of water on top of the sea ice which could influence the radiance retrieved by OLCI images.

An example of the OLCI image and the waveform returns from their corresponding points seen in Figure 2.12 support these arguments. The points which are classified as sea ice on the OLCI image show very specular reflections (a,b,c). The ground truth considers (i,ii) as leads, in which the waveform return of (i) agrees to. However, waveform from (ii) is not specular but rather diffuse. This could be due to the presence of waves on the water or small ice floes with sizes of below the resolution of OLCI image. The waves on wide leads were uncommon in the winter seasons, but the thinner surrounding ice and more inter-connected open water could allow larger waves to be in the open water, resulting in diffuse returns during the summer season. Furthermore, OLCI image show unevenness in color on the sea ice. This can cause the ground truth to be inaccurate as well. The slightly darker areas may suggest that there is a melt pond, but this must be confirmed with images or sensors with higher resolution.

In conclusion, detection of leads in summer using waveform altimeter or images from OLCI performs poorly and an alternative method must be found. The sole usage of OLCI image to detect the sea ice openings could possibly be done, however further elaboration on the image segmentation algorithm is needed. Even if the usage of OLCI image for lead detection during summer can obtain high accuracy, the lead detection in summer with Sentinel-3 still comes with difficulty as the Sentinel-3 latitude coverage is limited while the ice sheets extent continues to shrink and exists only in higher latitudes.

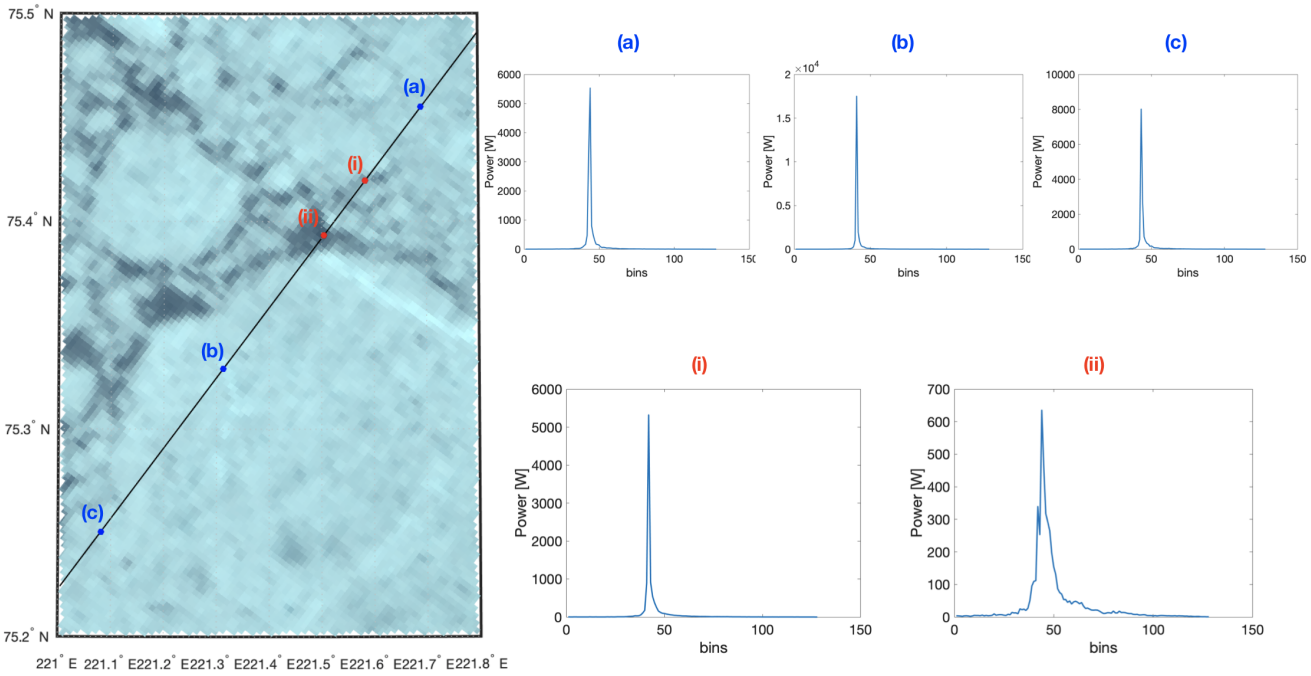


Figure 2.12: Left image: Sentinel-3B ground track on OLCI image taken on 07/07/2020. The validation data consider red points (i, ii) as leads and blue points (a,b,c) as sea ice. Right: L1-b waveforms from the corresponding points seen in the OLCI image.

2.4.7. Influence of Off-nadir Leads (D-05)

Finally, the influence of the presence of the off-nadir lead (ONL) is analyzed by adding another class to the analysis of supervised machine learning classifications. Figure 2.13 shows an example of a confusion matrix of RUS Boost classifier once ONL class is introduced. The numbers show the total prediction made per class over the given ground truth class. The % values show the fraction of the number of predicted class over the given ground truth class (each row sums up to 100%). The most important information that can be extracted from this confusion matrix is the True ONL Rate, TLR, and FLR_{ONL} (see Section 2.3.2 for definitions), and these are used to effectively compare the different classifiers. False leads (sum of cells with blue boundaries of Figure 2.13) and its fraction over the total number of waveforms (FLR_{ONL}) is the value that is most concerned in this analysis because by introducing the ONL class, it is aimed to reduce the total falsely predicted leads such that a potential bias in the SSH estimation can be reduced.

This analysis was conducted for other classifiers and their results are summarized in Table 2.6. Firstly, most of the classifiers were not able to detect majority of the ONLs. KNN and LD classifiers did not predict a single ONL. RUS Boost classifier was an exemption, detecting 60.26% of the ONLs. This is due to its algorithm which aims to alleviate the problem of class imbalance as discussed in subsubsection 2.2.3.1, hence it can detect more ONLs than others. However, this comes at a cost of lowering the TLR values. Although NB classifier did not detect as many ONL as RUS Boost, it also obtained a relative high ONL rate reaching 44.68% of the total ONLs.

When comparing the binary class models and the 3-class models, the performance differences varied depending on the classifier. The difference in TLR and FLR were also different per classifier, either effecting the results positively (green) or negatively (red). Interestingly, the two classifiers which had a higher detection rate of the ONL (RUS Boost and NB classifiers), were the only two classifiers that decreased their FLR_{ONL} as

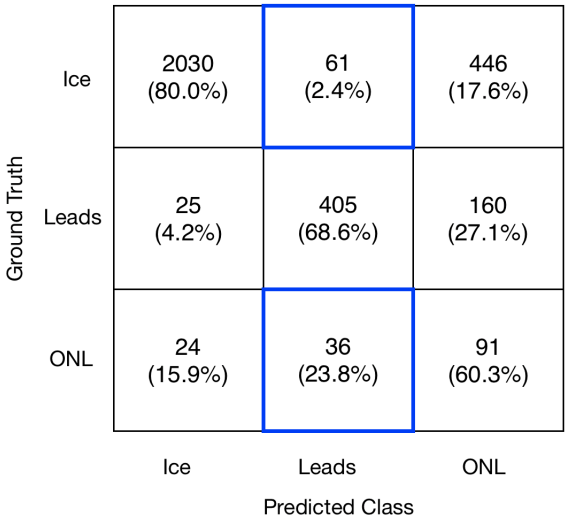


Figure 2.13: Confusion matrix of RUS boost classifier, 3-class model.

compared to the binary model. The most significant change in results between binary and 3-class model was also seen in RUS Boost and NB classifiers, where the TLR dropped up to 17.66% and the FLR_{ONL} also dropped 2-3%.

In conclusion, many classifiers are not able to detect ONLs when applying an additional class in the training data set of the supervised machine learning classification scheme. Simply adding another ONL class does not imply that the number of falsely predicted leads can be reduced. However, it has been seen that the two classifiers which had high ONL detection rate had a significant drop in the number of falsely predicted leads. It is unknown whether this relations always stands as the sample size of this study is too small, however this is an interesting point for future studies.

Since RUS Boost can best predict the ONLs, it could also be used to reject the ONLs in the beginning of the analysis if a more conserved lead detection is desired. However, there needs to be more research into the general waveform features of the ONLs and to improve its prediction accuracy.

Table 2.6: Results of classification performances of the supervised learning classifiers for the off-nadir leads analysis, given by D-05 data set. Results from both 3-class model and the binary model are shown, where the difference of their TLR and FLR_{ONL} are also presented.

	3-class model						Binary class model				Difference	
Classifier	Number of detected ONL	True ONL Rate [%]	Number of True Lead	TLR [%]	Number of False Lead	FLR_{ONL} [%]	Number of True Lead	TLR [%]	Number of False Lead	FLR_{ONL} [%]	TLR (3class-binary) [%]	FLR_{ONL} (3class-binary) [%]
AdaBoost	5	3.31	458	77.63	125	3.81	438	74.2	95	2.90	3.39	0.91
Bagged	10	6.63	446	75.59	129	3.94	419	71.0	100	3.06	4.58	0.88
RUSboost	91	60.26	405	68.64	97	2.96	497	84.2	206	6.29	-15.59	-3.33
ANN	15	9.93	437	74.07	116	3.54	466	79.0	101	3.09	-4.92	0.45
KNN	0	0	450	76.27	119	3.63	435	73.7	77	2.35	2.54	1.28
LD	0	0	416	69.22	62	1.89	392	61.73	45	1.40	7.49	0.49
NB	63	44.68	369	61.40	48	1.46	502	79.06	125	3.89	-17.66	-2.42

2.4.8. Comparison to Other Studies

The results obtained in this paper cannot be directly compared to the results from the previous studies due to their different input data (e.g, different satellite, instruments, study dates and study areas), different settings of the classifiers, or the different ground truth that was used for validation. However, it is still interesting to compare the similarities and differences to the findings of this paper.

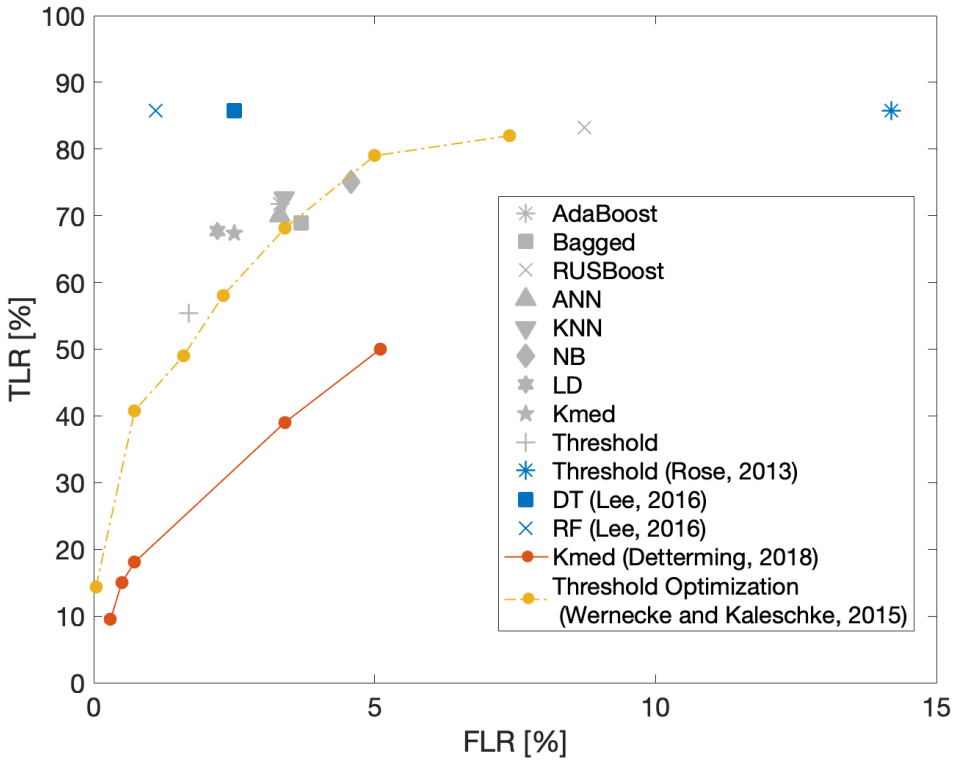


Figure 2.14: ROC graph showing classification results from previous studies and general performance (D-01) results obtained by this paper.

For example, Lee et al. (2016) also applied tree-based supervised machine learning classifications; decision trees (DT) and random forest (RF), on SAR altimetry data from CryoSat-2 for lead detection, and validated them with MODIS images taken by the Terra and Aqua satellites. These results are presented in a ROC graph (in blue square and x markers) together with the winter general performances obtained from this paper (gray markers) in Figure 2.14. The classification results obtained by Lee et al. (2016) show extremely high accuracies and high TLR values compared to results obtained in this paper. However thee TLR and FLR values are validated only with 239 waveforms, hence it is possible that the classifiers have been over-fitted to this small study area. Their findings also show that ensemble tree classifier outperforms DT classifier as it obtained lower FLR for similar TLR values, which agrees to the findings of this study.

Dettmering et al. (2018) studied the classification performance of unsupervised K-medoid classifier, using CryoSat-2 altimetry data and validated with images taken from the NASA Operation Ice Bridge mission. The ROC curve is also shown in Figure 2.14 in red. The resulting accuracies found by them were significantly lower than the result of the K-medoid classifier seen in this study. This may be due to the fact that the resolution of the validation image was 1 m (Dettmering et al., 2018), as compared to the 300 m resolution of OLCI that was used in this paper. This allows for more narrow leads to be found, which can possibly increase the number of falsely predicted leads.

Furthermore, the results obtained by Wernecke and Kaleschke (2015) who classified CryoSat-2 altimetry data using threshold optimization and validating them with MODIS images taken by Terra and Aqua are also presented in Figure 2.14. Note that Figure 2.14 show the ROC curve of the best performing classifier, using MAX. Their TLR and FLR values are comparable to the results from this study, however they obtain slightly lower TLRs in the region of 2 - 5 % FLR. Their results are also slightly worse compared to the result obtained by

the threshold classifier from this paper, however this difference is minor, and it could be due to the differences in the study areas (Wernecke and Kaleschke (2015) only validated the data in the Beaufort Sea region).

By simply consulting the results from the past studies shown in Figure 2.14, one may conclude that RF classifier is the best and K-medoid classifier is not suitable for lead detection based on their TLR / FLR values. However, in order to fairly assess these classifiers, they must be compared with the same input and validation data, and this paper has successfully compared the classification performance of three different types of classifiers (supervised-, unsupervised machine learning and thresholding method) in order to gain broader understanding of the performance difference in the classifiers and identified the most suited classifiers for lead detection.

2.5. Conclusions and Future Work

This paper aimed to assess different classification methods for lead detection in the Arctic Ocean using Sentinel-3 SRAL altimetry data. The paper assessed nine supervising machine learning algorithms, two unsupervised machine learning algorithms and a threshold classifier to study areas from a wide range of latitudes and longitudes. This provided a complete understanding of the performance of different classifiers, and their behaviours depending on the different case studies. The paper also proposed an automatic validation process using OLCI images to effectively analyze these large study areas. The method made use of K-means image segmentation and analysis of the changes in radiance values of OLCI image pixels along track of Sentinel-3 satellite. This novel approach of using OLCI image for validating SRAL altimetry waveform classification in the context of lead detection, ensures a perfect temporal resolution between the two data.

This study demonstrated successful classifications from March and April in the years of 2017 to 2020. Classification results varies on many factors and hence the classifier shall be selected considering the condition and the purpose of lead detection. Overall, AdaBoost, LD and ANN classifiers from the supervised machine learning showed the most robust and excellent result throughout the analysis. However, supervised learning requires labeled training data hence ground truth of the study area must be readily available. This study showed that applying a training data from another study area or year has slightly worsened the performances, hence this is not recommended. The unsupervised machine learning K-medoid classifier on the other hand, does not require the ground truth data and consistently showed excellent results. This makes this classifier extremely attractive, as producing ground truth data can be time consuming. A disadvantage of K-medoid classifier could be that it requires a user input when labeling the clusters, making the classifier not fully automatic and result can be affected by user bias. It is also not suited for analysing samples which users do not know the cluster differences, such as the analysis of off-nadir leads. Thresholding method did not perform as well as the machine learning based methods. However, due to its simplicity in application, one may still prefer this method over the others. It is also easy to adjust its result by simply changing the threshold values, but this is only effective when the ground truth class distribution is known.

Application of these classifiers on SAR altimeters in the summer months is not recommended, as most of the waveform returns, including sea ice, show specular reflection. Therefore, classification using solely SAR altimetry is deemed unsuitable and auxiliary information is required. Usage of high resolution optical data may aid for improving the accuracies for these seasons. OLCI image can serve for this purpose by implementing a better algorithm to distinguish the melt ponds, uneven color in ice, and leads. However, due to the rapid melting of sea ice, Sentinel-3 ground track coverage will not be optimal for this analysis in the near future.

This paper also assessed whether supervised machine learning can distinguish between leads, sea ice and ONLs. Many classifiers failed to detect the ONLs, with an exemption of the RUS Boost and NB classifiers. Though it was with a cost of decreasing TLR, RUS Boost and NB classifiers decreased the number of false lead compared to their binary model. This technique can be used to initially reject ONLs for more conservative analysis, however, there must be more study on this to achieve a higher True ONL Rate, to minimize the number of rejected points.

The author sees room for improvement in this research in the following ways. Firstly, SSH shall be computed using the leads detected by the proposed classifiers in this study, and validate this with other altimetry measurements for instance from the dedicated aircraft campaigns. Similar work has for example been seen in Lee et al. (2016), where the derived ice thickness estimations were compared to the Airborne Electromagnetic (AEM)-bird data collected during the CryoSat Validation experiment (CryoVex) field campaign. This way, the validation method is not only biased towards the OLCI image and the derived product can be directly compared among the classifiers. Secondly, a better algorithm shall be developed for OLCI image classification,

especially for cloud rejection. Cloud rejection in this study has been done manually, as the existing software with cloud rejection capabilities, such as the Sentinel Application Platform (SNAP) (European Space Agency), did not always manage to identify the smaller sized clouds. This manual rejection is time-consuming and limits the areas that could be validated. With an improvement in the algorithm for cloud rejection, the ground truth generation as well as the validation process would be more efficient, allowing the analysis to cover more areas in the Arctic Ocean. Fully automated ground truth generation can also benefit the off-nadir leads analysis, if there are more reference samples of ONLs, the classification accuracy could also be further improved.

Acknowledgements

We would like to thank the European Space Agency (ESA) data hub for the publicly accessible Sentinel-3 OLCI imagery and SRAL data (https://scihub.copernicus.eu/).

3

Conclusions and Recommendations

3.1. Conclusions

The aim of this research was to assess different waveform classification methods in order to ultimately find the best one for SSH estimation purposes in the Arctic Ocean. The main research question and its sub-questions were formulated in Section 1.5, in order to accomplish this research goal. This section provides answers to these questions by referring to the main findings from the journal article presented in Chapter 2, which leads to a final conclusion of this research.

- **SQ-1: How can SAR altimetry waveform classifications be validated in an effective way?**

  This study used the images derived by OCLI on board Sentinel-3 as the ground truth. The main advantage of this was to achieve a perfect temporal alignment with the measurements from SRAL SAR altimeter, also on board Sentinel-3. Therefore, no ice-drifting models had to be employed in this research. Because this study aimed to conduct waveform classification for a large part of the Arctic Ocean for over several years, it required a very large amount of validation data to be produced. Therefore, the ground truth generation process as well as the validation process were aimed to be fully automated. Unfortunately, due to the complex nature of clouds affecting the hyperspectral images, publicly available software with cloud rejection capabilities such as the Sentinel Application Platform (SNAP) (European Space Agency), could not completely reject the smaller sized clouds. Therefore it has been decided that the clouds will be rejected manually in this study. Aside from the cloud rejection, the ground truth generation and validation process has been automated by using K-means image segmentation and analysis of radiance changes in OLCI image pixels nadir to the satellite.

- **SQ-2: How do the overall accuracy, TLR, and FLR compare between the classifiers?**

  From this study, it has been seen that the classification results vary significantly depending on the application and there is no single classifier that is always better than the others. The results are also not straightforward to interpret. The overall accuracy alone cannot provide any insight to how the classifiers falsely classify the leads. The TLR and FLR values are also difficult to compare among classifiers, unless they are compared for a fixed FLR or TLR. Therefore, this study made use of ROC graphs and its area under the curve. With the help of ROC graphs, it is clear to see which classifiers are more suited for this problem. In the preliminary analysis, it has been seen that DT, SVM and HC classifiers performed worse compared to the rest and were discarded for further analyses. From the different case studies conducted with the remaining classifiers, it has been seen that LD, Ada Boost, ANN and K-medoid classifiers consistently produced excellent results, in terms of overall accuracy, TLR and FLR.

- **SQ-3: What is the combination of waveform parameters that results in best accuracy?**

  In the article presented in Chapter 2, the best combination of waveform parameters per classifier were not looked for, but rather a combination of waveform parameters which was beneficial to all the classifiers were found. This was not only to provide a fair comparison between the classifiers but also to avoid having too many dependent variables to control during the analyses. There are also a number of other

(hyper)parameters to be tuned, which are also dependent on the combination of waveform parameters, hence it has been decided to fix the combination of waveform parameters to be used. This study analyzed the individual "classification potential" of the waveform features, by examining the classification performance when trained with a single waveform feature. As seen in Table 2.3, MAX, skew, PP, WW, and PPloc all achieved overall accuracies with an average of above 85%. WW had the best average overall accuracy, achieving 91.03 %. TeW also achieved an overall accuracy of above 85%, however because K-medoid classifier performed poorly with this waveform it has not been used for further analyses.

Because this selection of waveform parameters could have influenced the final result of the classifiers, further analysis was conducted. This is given in Appendix A.3, showing the influence of adding different waveform features. The classification performances vary with different combination of waveform parameters, but the relative performance of the classifiers remains approximately constant. This additional analysis justifies the selection of a specific combination of waveform features to compare the classification performances between the classifiers.

- **SQ-4: How do the hyperparameters of machine learning algorithm influence the classifiers performance?**

  The hyperparameters of each classifier influenced the classification performance very differently. On average, a change in one hyperparameter in 0.1% to 0.5 % change in the overall accuracy, but also some resulted in a few % change. Because this can influence the result significantly, the classifiers were carefully tuned before being applied to the test data.

- **SQ-5: How well do classifiers perform when trained with data from another year?**

  The effect of classifiers being trained with data from another year has been studied using the D-02 data set, as presented in the article (Chapter 2). The results were plotted in ROC graph together with the general performance results (D-01), for comparison. It has been seen that generally the results obtained by the supervised classifiers do not suffer when trained with data from another year. LD classifier even showed better results compared to its general result. KNN classifier was an exception, as its performance suffered significantly. In contrary, the unsupervised learning K-medoid classifier is not affected by the change of training data, as it is only dependent on the altimetry data.

- **SQ-6: How well do classifiers perform when trained with data from another study area?**

  The effect of classifiers being trained with another study area has been tested with the D-03 data set. Though the overall accuracies remained approximately the same from the general performance, the classifiers suffered more when compared to results obtained with D-02 data set (trained with data from another year) when considering the TLR and FLR values. The reason to this could be due to the fact that the trained data taken from southern areas compared to the testing data. This suggests that more ice melting could have taken place in the training data, which affects the waveform returns. This has also seen in the seasonal analysis (D-04 or SQ-7) that when sea ice melting occurs, classification using only waveform returns becomes more difficult.

- **SQ-7: What is the seasonal influence on the performance of classifiers?**

  It has been seen that the presented classifiers perform poorly during the summer seasons, with classifiers not detecting enough leads (too low TLR) or over-detecting them (too high FLR). By observing the waveform returns from an example study area, it has been clear that the waveforms did not corresponds to the ground truth. Most waveforms had specular reflection, even on sea ice. A point which the ground truth considers as a lead showed a diffuse return, which is probably due to the large open water area connecting to other open water areas, making it easier for the waves to form.

  It is known that the ice melting occurs more rapidly in the sea ice edges as they are more exposed to the dark ocean surface, making them more prone to melting (Xia et al., 2014). Since the ground track latitude is limited for Sentinel-3 satellite, during the summer season it measures the points closer to the sea ice edges. Therefore, it was speculated that this could be the reason why most returns were specular and the results may be different for sea ice closer to the North pole. Therefore, an additional study using CryoSat-2 satellite has been conducted. This material is provided in Appendix B.2. This additional study disproved the hypothesis; the specular returns are dominant even in areas closer to the North pole. Therefore, it is concluded that classification using solely the SAR altimetry is unsuitable in summer, not only for the studied areas but also areas closer to the North pole.

- **SQ-8: What is the impact of the presence of off-nadir leads on the performance of the classifiers?**

  In order to understand the impact of the presence of off-nadir leads (ONL) on the performance of the classifiers, this study has added another class to be trained in the supervised learning classifiers. This was called as 3-class model, with waveforms being labeled as either leads, sea ice, or ONL. This was introduced since it was assumed that the with the presence of ONL, the classifier falsely predicted leads when ground truth is actually sea ice. The study aimed to reduce these falsely predicted leads by adding another class for ONLs. From all the classifiers that were studied, RUS Boost and NB were the only classifier which reduced the number of false leads in the 3-class model compared to the binary class model. Although 3-class model NB classifier reduced its number of false lead compared to its binary model, binary model LD classifier obtained better result with higher TLR and lower FLRONL Therefore, it can be understood that the 3-class model is not necessarily better than the binary model, even if some classifiers can significantly reduce the number of false lead. Finally, because RUS Boost best predicted the ONLs, this can be used to reject the ONL points in the beginning for more conservative lead detection.

Finally, the main research question can be answered:

- **RQ: How do empirical thresholding methods, supervised and unsupervised machine learning based classifiers compare in their performances, and what advantages/disadvantages do they have when detecting leads in the Arctic Ocean using SAR altimetry data?**

  It cannot be concluded simply stating that one type of classifier is better than the other, as this study showed how different application can affect the performances of the classifiers very differently. Overall, the supervised learning AdaBoost, ANN, and LD classifiers showed the most robust and excellent result throughout the analysis. However, supervised learning requires labeled training data hence ground truth of the study area must be readily available. This study showed that applying a training data from another study area has slightly worsened the performances, hence this is not recommended. The unsupervised machine learning K-medoid classifier on the other hand, does not require the ground truth data and consistently showed excellent results. This is makes this classifier extremely attractive, as producing ground truth data can be time consuming and computationally heavy. A disadvantage of K-medoid classifier could be that it requires a user input when labeling the clusters, making the classifier not fully automatic and result can be affected by user bias. It is also not suited for analysing samples which users do not know the cluster differences, such as the analysis of off-nadir leads. Thresholding method did not perform as well as the machine learning based methods. However, due to its simplicity in application, one may still prefer this method over the others. It is also easy to adjust its result by simply changing the threshold values, but this is only effective when the ground truth class distribution is known.

  This research successfully provided more insight on the behavior of the different classifiers. Previously, performances of different classifiers were known from different studies, which used different input data and validation data. Therefore, the classifiers could not be directly compared, and it was for example still unknown whether unsupervised classifiers could outperform supervised classifiers. When only considering the results from the past studies (see Figure 2.14), it may seem like K-medoid classifier (from Dettmering et al. (2018)) performed significantly worse compared to the DT or RF classifiers (from Lee et al. (2016)). However, this significant difference is believed to be mainly coming from the difference in their validation process; Dettmering et al. (2018) used validation data with extremely high resolution (1m) and Lee et al. (2016) only validated 239 samples. As this study applied both classifiers on the same input and validation data, it has become clear that the K-medoid classifier produce similar or better results than the DT and ensemble tree classifiers (Ada Boost, Bagged, RUS Boost). Therefore this study also showed the importance of assessing the classifiers with the same input and validation process for a fair comparison.

3.2. Recommendations

During this thesis, interesting ideas and additional tasks have been identified that were left unexplored, mainly due to time constraints. These points are believed to improve the research and are valuable additions to this topic. This section briefly describes these points for future research.

- **Computation of SSH using the detected leads**

  This research focused on detection of leads, such that these instantaneous water level references can be used to estimate the SSH. Therefore, it is of high interest to compute the SSH using the leads detected by the proposed classifiers. These shall be validated with other altimetry measurements such as the dedicated aircraft campaign (e.g. Operation Ice Bridge), laser measurement from ICESAT-2, tide gauges, or existing sea surface height models. For example Poisson et al. (2018) compared its derived SSH to the DTU 2013 mean sea surface (Andersen et al., 2015) to compute the sea level anomaly. Lee et al. (2016) also used the Airborne Electromagnetic (AEM)-bird data collected during the CryoSat Validation experiment (CryoVex) field campaign to compare the derived ice thickness. In this way, the validation method is not only biased towards the OLCI image and the derived product can be directly compared among the classifiers.

- **Cloud rejection in OLCI images**

  One of the bottlenecks in this study was the cloud rejection procedure of the OLCI images. Cloud rejection in this study has been done manually, as the publicly available software with cloud rejection capabilities such as the Sentinel Application Platform (SNAP) (European Space Agency) was not able to detect the smaller sized clouds very well. This manual rejection is time-consuming and limits the areas that could be validated. With an improvement in the algorithm for cloud rejection, the ground truth generation as well as the validation process would be more efficient, allowing the analysis to cover more areas in the Arctic Ocean.

  For example, an ongoing cloud making development project "Sentinel 3 Synergy Cloud Mask Development" by EUMETSAT, which aims to develop a new atmospheric mask product primarily focused on clouds, by exploiting the spectral synergistic capabilities of OLCI and SLSTR instruments on-board the Sentinel-3 satellites (EUMETSAT, 2021), may allow for better cloud masking for the future studies. Furthermore, an interesting study conducted by Giuffrida et al. (2020) proposed a Convolutional Neural Network (CNN) to be integrated to the hyper-spectrometer carried by a nanosatellite to reject cloud-covered areas before transmitting them to the ground. The transmitted images are those which present less than 70% of cloudiness in the frame. This has been already installed on the Hyperscout-2, as part of Phisat-1 ESA mission which was launched in September 2020. This novel and innovative technology of using artificial intelligence on board the satellite can also be advantageous in the future for selecting study areas of optical or hyperspectral images that are affected by the presence of clouds.

- **Automation of identifying off-nadir leads in OLCI images**

  Another major manual process conducted in this study was labeling the ONL points in the D-05 data set. The labelling process involved close inspection of the individual points, manually. This study did not propose an automatic detection of these ONL points due to the time constraint and also because it was not the main focus of the research. However, by developing an algorithm which can identify points which contains off-nadir leads within the SRAL cross track resolution (1.64 km), more insights on ONLs can be gained. Because there are no specific waveform shape of the ONLs that are known, it will be interesting to see how the waveform shape changes with respect to the surrounding environment (size of the ONL, location of the ONL, number of surrounding leads in the location, etc.).

- **Additional classifiers to be studied**

  In this study, 13 classifiers were analysed. They were mainly selected due to the promising results seen in the previous studies in the context of waveform classification and also considering the software package availability on MATLAB. Due to time constraint, not all possible classifiers could be explored. Especially there were less unsupervised learning classifiers in this study, compared to the number of supervised learning classifiers. Based on the results obtained by this research, application of self-organizing maps (SOMs) may be an interesting addition. SOMs are a type of artificial neural network (ANN), which are trained using an unsupervised, competitive learning as opposed to error-correction learning used

in conventional ANNs (Miljković, 2017). SOMs are also frequently used in the field of remote sensing, especially in land cover and agricultural classifications (A. Filippi, 2010). Because ANN classifier produced excellent results throughout this research, an unsupervised learning classifier using ANN may outperform K-medoid classifier and show more robust outputs.

- **Validation of OLCI images**

  This study made use of OLCI images as the validation data. As stated in the article, this validation data does not depict the reality as it cannot detect leads narrower than its resolution, but rather act as a common ground to assess the SAR altimetry waveform classifications. Therefore, the classification performance measure is always limited by the accuracy of the validation data. In order to understand the accuracy of this validation data, it is necessary to compare the OLCI image with higher resolution images to understand its accuracy and see how the radiance of a pixel capture leads that are smaller than its resolution. Unfortunately, this is not an easy process since as there are temporal delay between OLCI image measurements and the available higher resolution images from aircraft campaigns such as CryoVex and Operation IceBridge. Because the speed of the current of 0.1km/h (Quartly et al., 2019), ice drift models must be employed for such analysis. An additional study was conducted to attempt validating the OLCI images. See Appendix B.4 for more information.

- **Consideration for ocean class**

  This study focused on a binary classification (leads and sea ice), as all the study areas which were selected could be classified to those two classes. However, in reality the study areas may include surfaces of the ocean. In such a case, those areas must be manually rejected before proceeding to the binary classification proposed by this study, as the training data does not include waveform returns from the ocean. In the future studies, including another class for the ocean surfaces will be more effective for the users. Furthermore, addition of the ocean class may also resolve the ambiguity between the waves seen on the leads (see Section 2.4.7) and the sea ice return.

- **Lead detection during summer seasons with other methods**

  This study has demonstrated that, unfortunately, the SAR altimetry waveform classification during summer is not suitable as almost all waveform returns show specular reflection due to ice melting. However, detecting open water during summer is crucial for understanding the quantitative impact of the ice melting in the Arctic Ocean. Therefore, an alternative method must be used during this season for lead detection. The article showed that OLCI image is capable of doing this, if the lead detection algorithm can be improved and tailored for the summer season as uneven sea ice color hinders its accuracy. However, it was also mentioned that due to the limitation of the ground coverage of the Sentinel-3 satellite, this might not be the best option.

  Because CryoSat-2 also uses the SAR altimeter, the same issue will be faced in the summer months, which is also discussed in Appendix B.2. On the other hand, studies have shown that ICESAT-2 can successfully detect leads by analyzing the photon rate (number of photon returns per laser pulse, i.e. the apparent reflectivity of the surface) from its lidar called Advanced Topographic Laser Altimeter System (ATLAS) (Petty et al., 2020). The results show that the leads are distinguishable even during summer months. Additionally, Tilling et al. (2020) characterized the complex photon backscatter signals from the melt ponds, which serves as the first step to automating melt pond detection and improving the sea ice height products. However, more research is needed to gain further understanding of the complex nature of the photon returns from the melt ponds.

  Finally, aircraft campaigns can take extremely high resolution images and hence leads and open waters will be able to be detected accurately during the summer seasons. However, these measurements are limited to local areas and continuous measurements cannot be made.

- **Use of SLSTR sensor for summer lead detection**

  Sentinel-3 satellites carry Sea and Land Surface Temperature Radiometer (SLSTR), which is a thermal radiometer providing land and sea surface temperatures (EUMETSAT, 2018). Exploiting this data might also be beneficial for further analysis in lead detection. This can be used to assess how the specular returns become more dominant (even over sea ice) when melting occurs, and relate this to the sea ice surface temperature. It has been seen that SAR altimetry waveform classification suffers significantly during the summer seasons. With the help of SLSTR measurements, the relation between the surface

temperature and classification accuracies could be derived. This could be useful as the waveform classification can be avoided once the surface reaches a certain temperature. Moreover, using SLSTR will demonstrate the synergy of the Sentinel-3 payloads and allowing the all the measurements to have a temporal match.

3.3. Other Applications

Based on the methodology and processes which were implemented in this thesis, there were arising ideas to apply these to other applications, despite not being directly tied to the topic of this research. These are listed in the following.

- **Detection of still water bodies such as lakes and canals**

  This study has presented the successful classification between specular and diffuse waveform returns. Specular returns are derived from smooth surfaces and these are not only limited to the leads in the Arctic Ocean. This classification may be for example applied to detect lakes and canals, which is useful for estimating the inland water level estimation (Kleinherenbrink et al., 2020) .

- **Use of image segmentation method for object detection**

  The image segmentation method used in this research acted as a powerful tool in distinguishing the darker leads from the lighter sea ice in the optical image. This image segmentation can be used number of other applications, such as object detection or surface anomaly detection, as long as the instrument can pick up on the differences in the radiance.

A

Appendix

This appendix serves as the appendix to the jounral article presented in Chapter 2. This appendix provides important supporting materials of this research which were not part of the article as they were deemed less important compared to the presented material. However, these additional materials provide important information to completely understand the research as they justify decisions made in the methodology or gain concrete understandings of the results.

A.1. Distribution of waveform features

Waveform features played an important role throughout this research as they provided the measure of dissimilarities between the waveforms. Figure A.1 shows the complete histogram of all waveform features class distribution. It can be observed that the waveform features which performed the best in the individual analysis for all classifiers; MAX, skew, PP, ww, PPloc, show a clear distinction between the two classes. On the other hand, the waveform features which performed poorly in the individual analysis such as NrPeaks, sigma0, LeW and PPL show otherwise as their overlapping area of the two classes are larger.

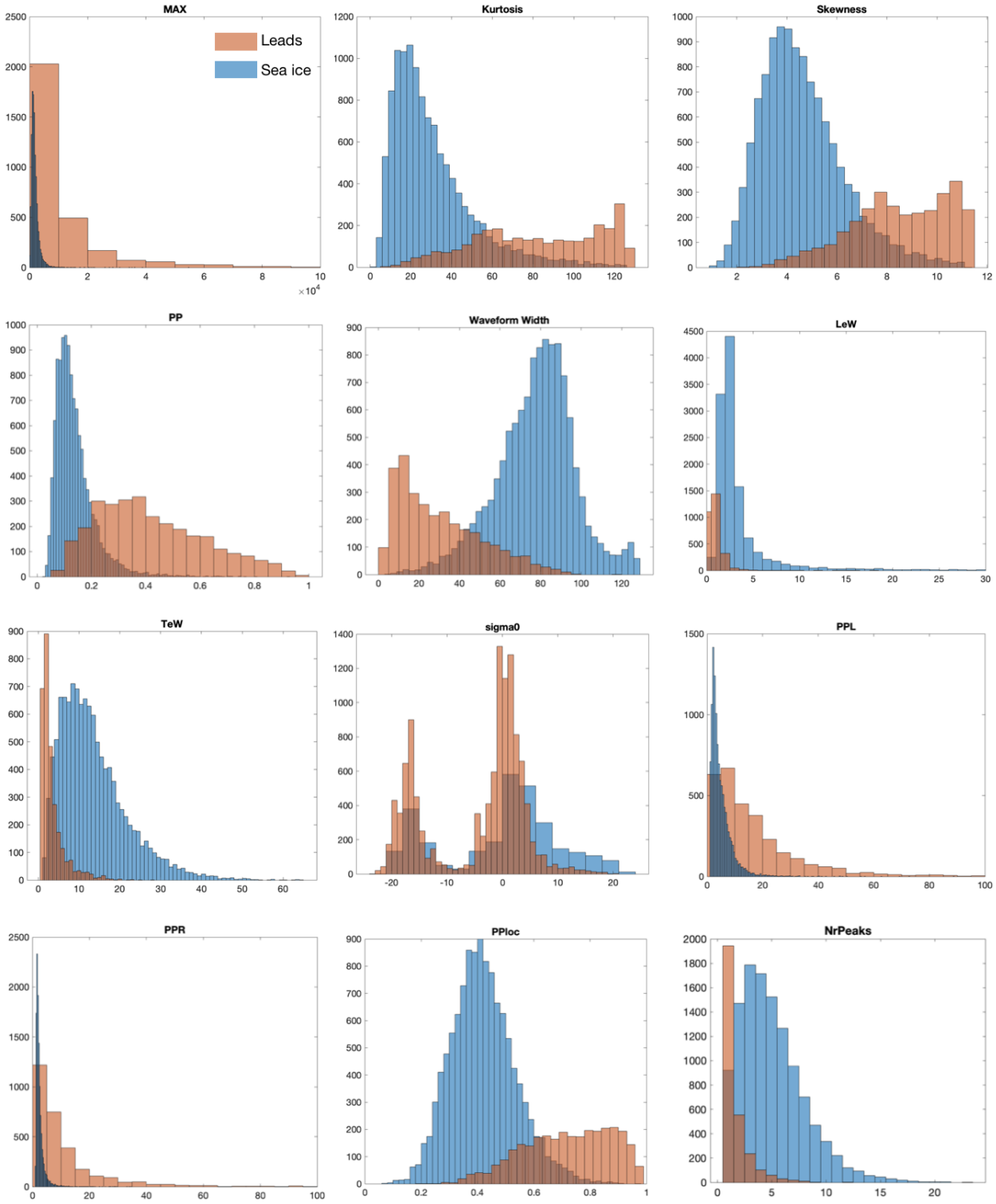


Figure A.1: Distribution of waveform features, data from March/April 2017 to 2020.

A.2. Selection of thresholding values

Analyzing the waveform distribution from the histograms as seen in Figure A.1, the thresholding values were found by using the selected features (MAX, PP, skewness, WW and PPloc), in accordance with the machine learning classifiers. First, the range of waveform feature that had majority of lead counts and could be distinguished from ice distributions were identified; 3,000 < MAX < 7,000, 7 < skew < 8, 0.15 < PP < 0.35, 0.5 < PPloc < 0.7, 25 < WW< 50. These ranges were used for each waveform feature during the random search, where 50 random value from each waveform feature were selected to find the outputting result. The result of this random search is seen in Figure A.2, where each point represents the classification result (in TLR and FLR) found by using the random combination of waveform features.

From Figure A.2, the points that clearly perform better than the rest form an optimal front. This can be seen as an optimization problem as this study aims to maximize TLR while minimizing FLR. One of the points in the front is selected, and this selected point is marked in yellow in Figure A.2. This point was selected since its performance is most comparable to other machine learning classifiers. This selected point was a result of thresholding values of the following; lead class if MAX>4000, PPloc > 0.55, ww<40, PP>0.3, skew>7, else: sea ice.

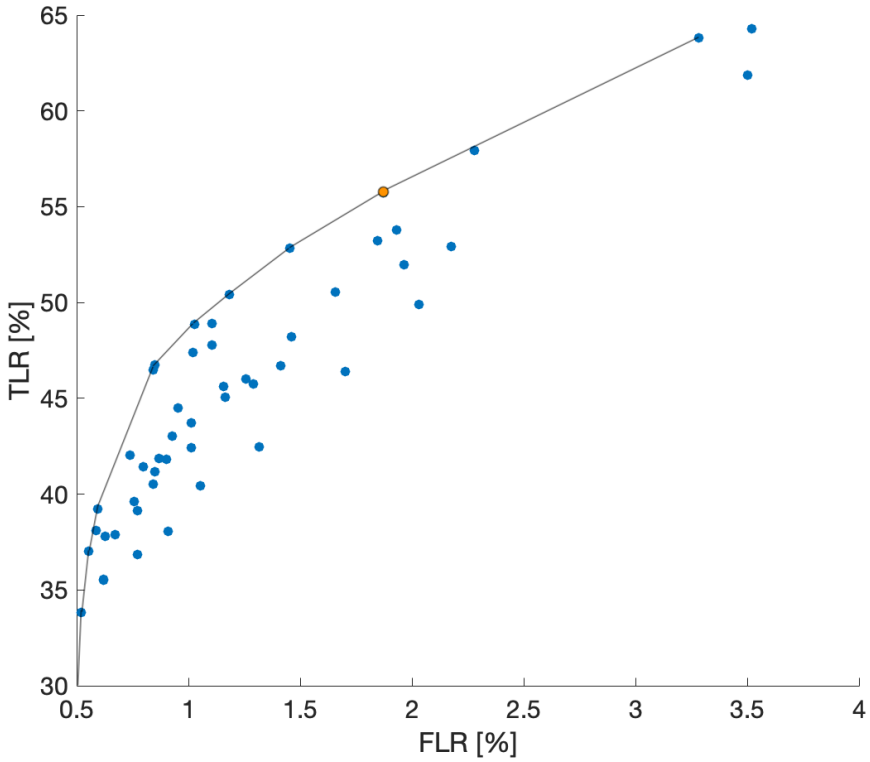


Figure A.2: ROC graph showing the results of random grid search of thresholding values. The line shows the pareto front of this analysis, whereas the yellow point depicts the final choice of the thresholding values used for this thesis.

A.3. Sensitivity analysis: influence of waveform features

In this section, the sensitivity of the classifiers are analyzed. The classifiers are dependent on the input settings including (hyper)parameters and waveform features. In this study, the waveform features were selected by running the classifiers with single features to find the classification potential of each feature. Because the presented paper aimed to use a combination of waveform features that is beneficial to all of the classifiers, there were some features that were not used in the analysis even if it was beneficial to some of the classifiers.

Table A.1 shows six different combinations of waveform feature combination which were tested in this analysis. Combination 1 was the combination that was selected in this study. Combination 2 to Combination 4 adds another waveform feature which performed well in the individual analysis for the supervised learning classifiers; namely kurtosis, TeW and PPR. Combination 5 adds these three features to the original one. Finally, Combination 6 uses all of the waveform features presented in this study, except from NrPeaks. It has been discussed in Section 2.4.2 that NrPeaks was not capable of distinguishing the leads and sea ice due to its nature of having discrete integer values.

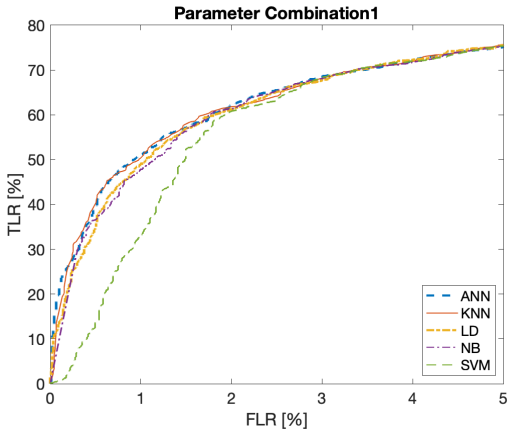
Table A.1: Combinations of waveform features to be used in the sensitivity analysis

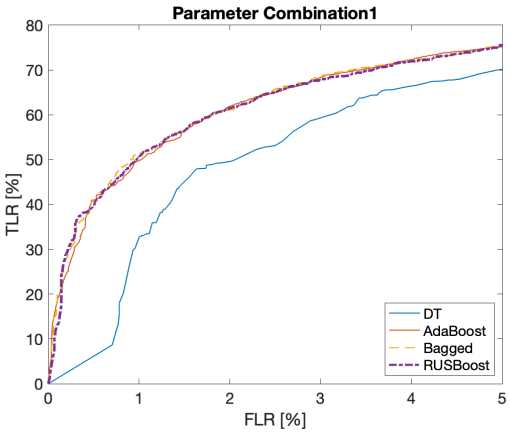
Combination	Waveform Features
Combination 1	MAX, skew, PP, ww, Pploc
Combination 2	MAX, skew, PP, ww, Pploc, kurt
Combination 3	MAX, skew, PP, ww, Pploc, TeW
Combination 4	MAX, skew, PP, ww, Pploc, PPR
Combination 5	MAX, skew, PP, ww, Pploc, kurt, TeW, PPR
Combination 6	MAX, skew, PP, ww, Pploc, kurt, TeW, PPR, LeW, sigma0, PPL

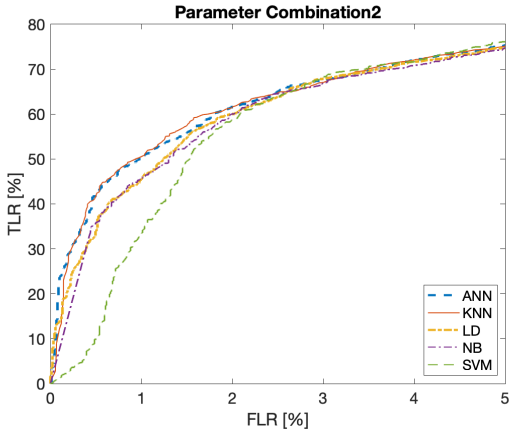
These combinations of waveform features are then applied to the classifiers. The ROC graphs of the supervised classifiers are produced by using 5-fold cross validation (see Figure A.3) and the cluster sizes were changed for unsupervised classifiers to generate their ROC graphs (see Figure A.4).

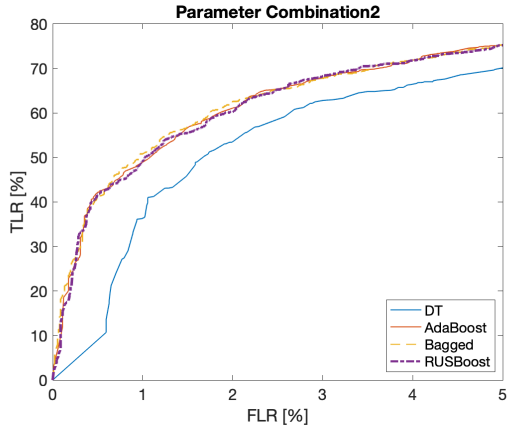
One can observe from these ROC graphs that for different combination of input waveform features, the classifiers behave very differently. Firstly, the results from the supervised learning classifiers show that relative performances of the presented classifiers do not change significantly. SVM and DT consistently performs worse compared to the rest of the classifiers, for all combination of the waveform features. The ensemble tree classifiers (Ada Boost, Bagged and RUS Boost) have very similar curves and their differences remain minimal for all waveform feature combinations. Though ANN, KNN, LD and NB experience some variation with different waveform combinations, their differences remain approximately constant. The ROC graphs obtained by the unsupervised classifiers have slightly more variability. There are some waveform combinations in which HC classifier outperforms the K-medoid classifier (e.g. Combination 4 and Combination 6), as it "crosses" the curve of K-medoid classifier. However, for all waveform feature combinations, K-medoid classifier outperforms HC classifier in most of the points.

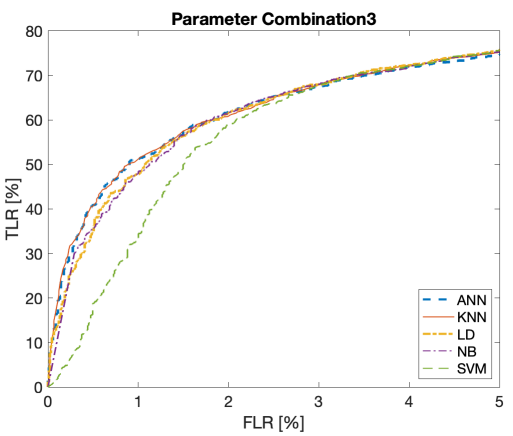
This analysis justifies the decision of initially selecting a specific combination of waveform features, as the relative performance of the classifiers do not vary significantly, when adding extra features that could have been beneficial to some of the classifiers.

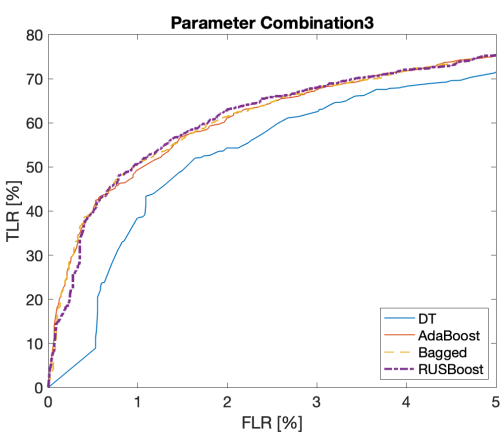

Parameter Combination1
TLR [%]
FLR [%]
ANN
KNN
LD
NB
SVM

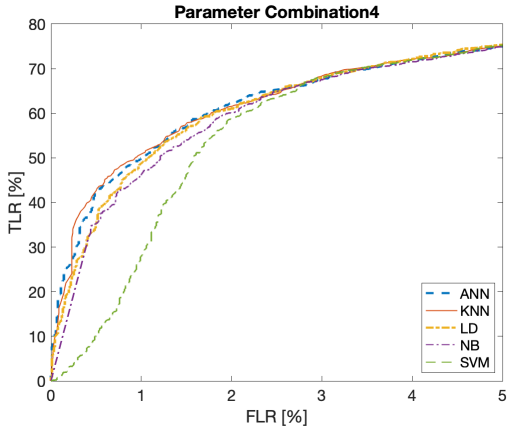

Parameter Combination1
TLR [%]
FLR [%]
DT
AdaBoost
Bagged
RUSBoost

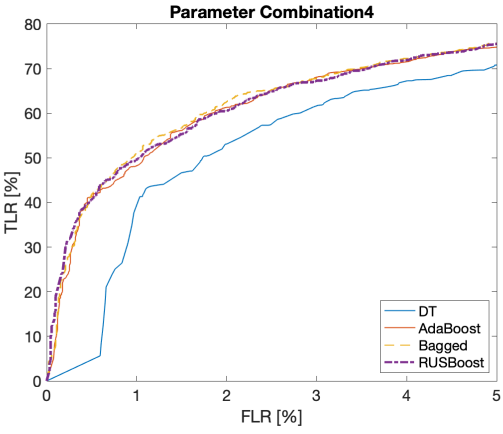

Parameter Combination2
TLR [%]
FLR [%]
ANN
KNN
LD
NB
SVM


Parameter Combination2
TLR [%]
FLR [%]
DT
AdaBoost
Bagged
RUSBoost


Parameter Combination3
TLR [%]
FLR [%]
ANN
KNN
LD
NB
SVM


Parameter Combination3
TLR [%]
FLR [%]
DT
AdaBoost
Bagged
RUSBoost


Parameter Combination4
TLR [%]
FLR [%]
ANN
KNN
LD
NB
SVM


Parameter Combination4
TLR [%]
FLR [%]
DT
AdaBoost
Bagged
RUSBoost

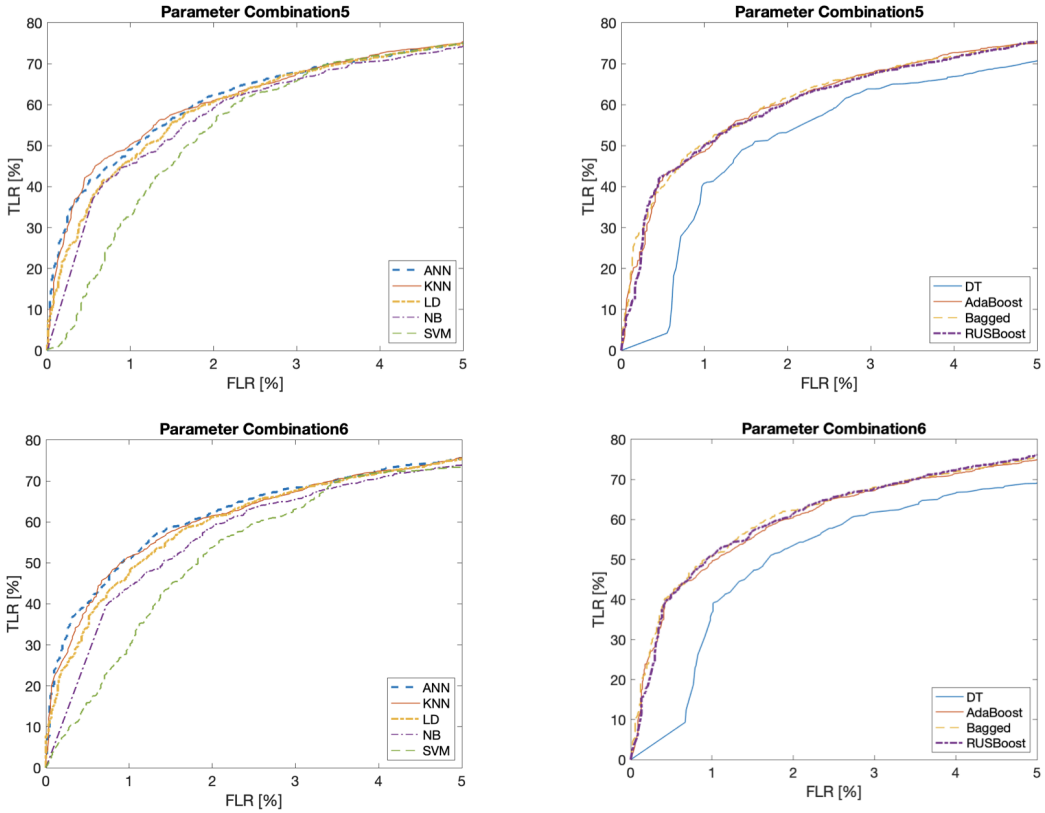


Figure A.3: Sensitivity analysis of supervised learning algorithms for different waveform features combinations.

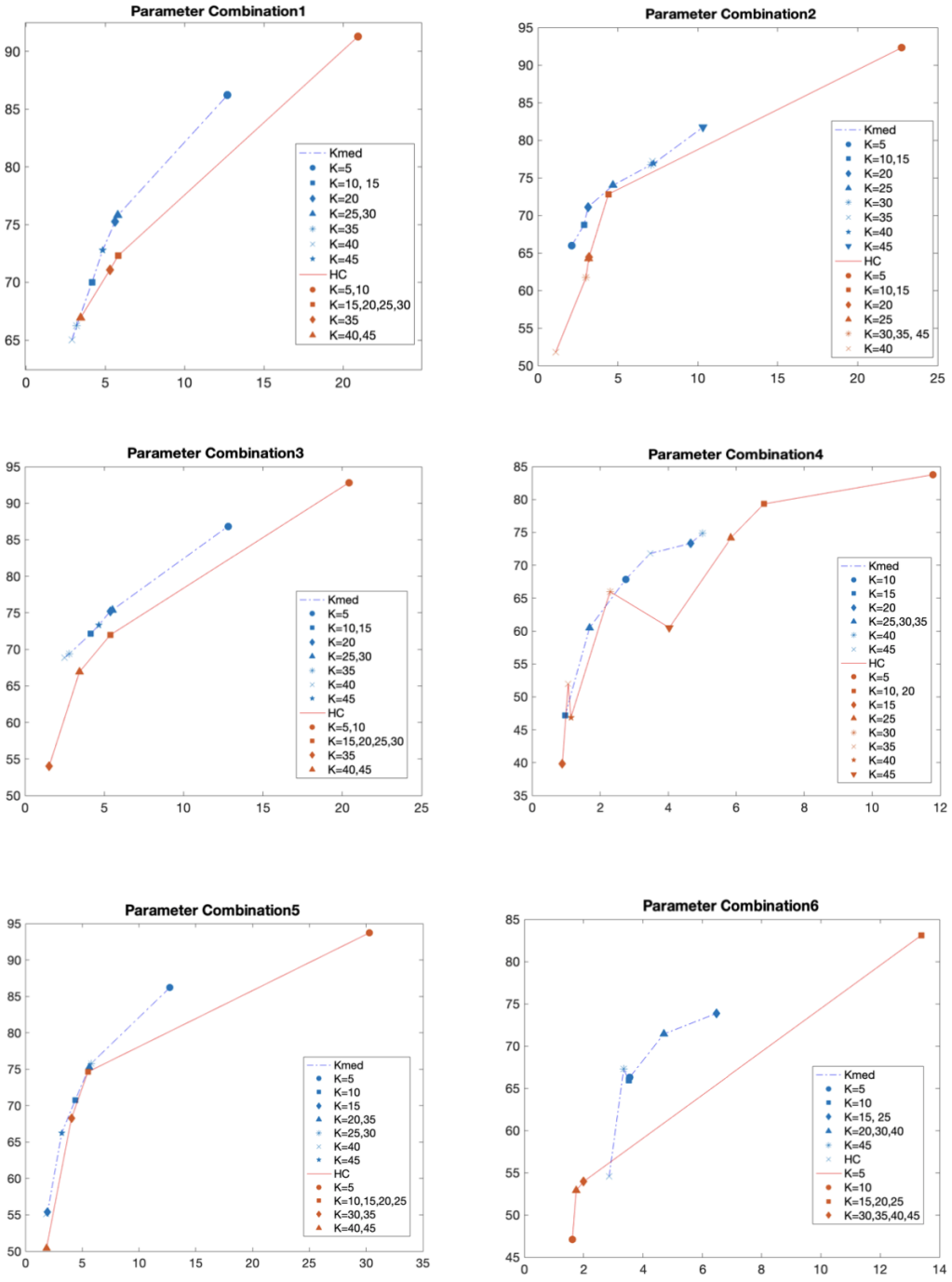


Figure A.4: Sensitivity analysis of unsupervised learning algorithms for different waveform features combinations.

B

Supporting Materials

In addition to the Appendix A, this appendix provides supporting materials which may not directly contribute to the research, but are deemed important and interesting to be presented.

B.1. Distribution of waveform features with off-nadir leads

Similarly to Appendix A.1, this section provides series of histograms showing the class distribution of the waveform features seen in the off-nadir leads (ONL) analysis (data set D-05). The histogram showing the class distribution of the ONL analysis is given in Figure B.1.

These histograms provide insight to why ONLs classification is not straightforward and many classifiers produced poor results in detecting them. From Figure B.1, it can be seen that for most of the waveform features, the distribution of the ONL class lies almost exactly where the lead and sea ice class overlap. This agrees with the fact that ONL waveform returns are practically a "superposition" of these two classes, therefore contain features from both of these classes. This supports the argument of ONL class not having a "unique" feature, therefore remains difficult to be detected.

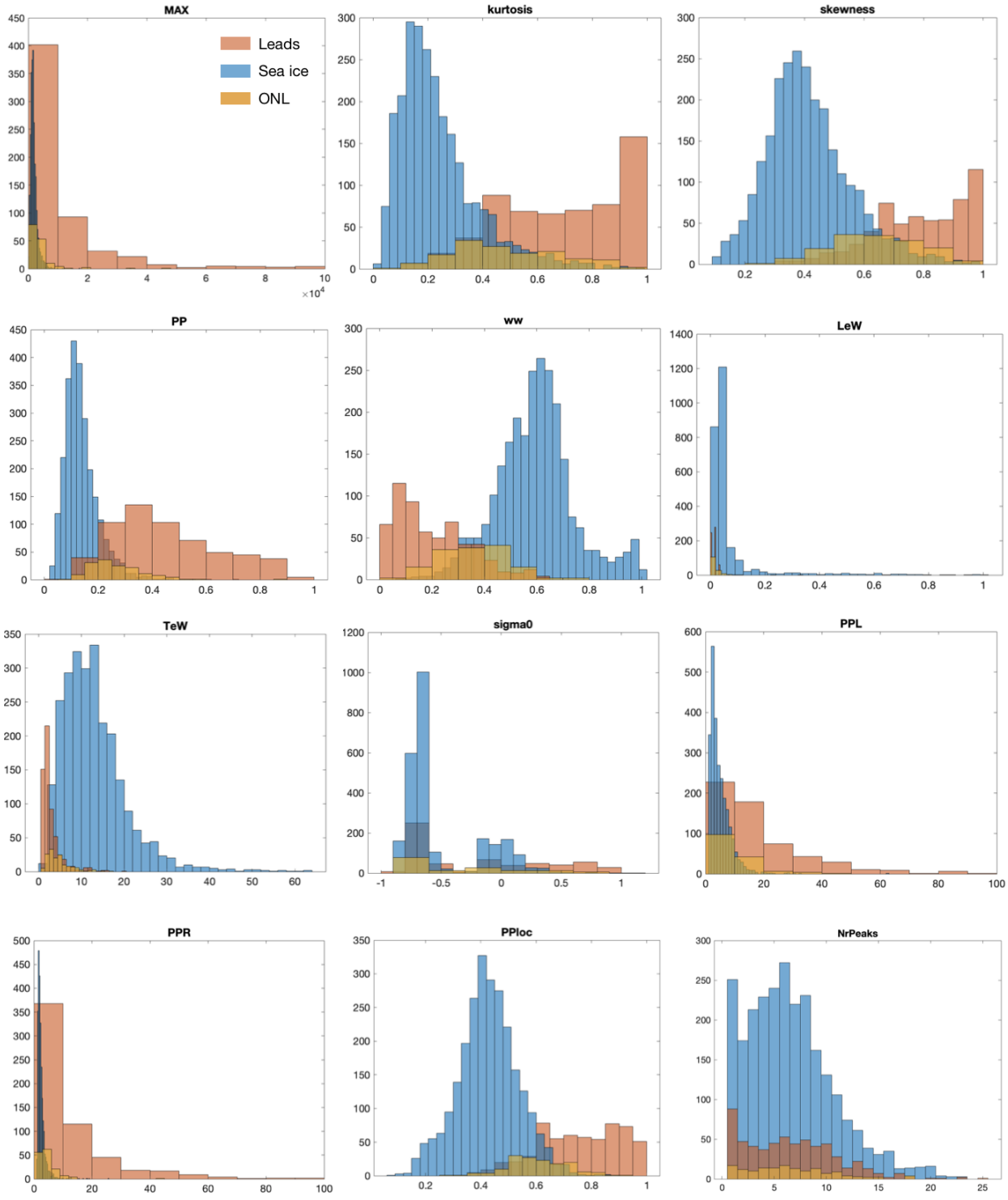


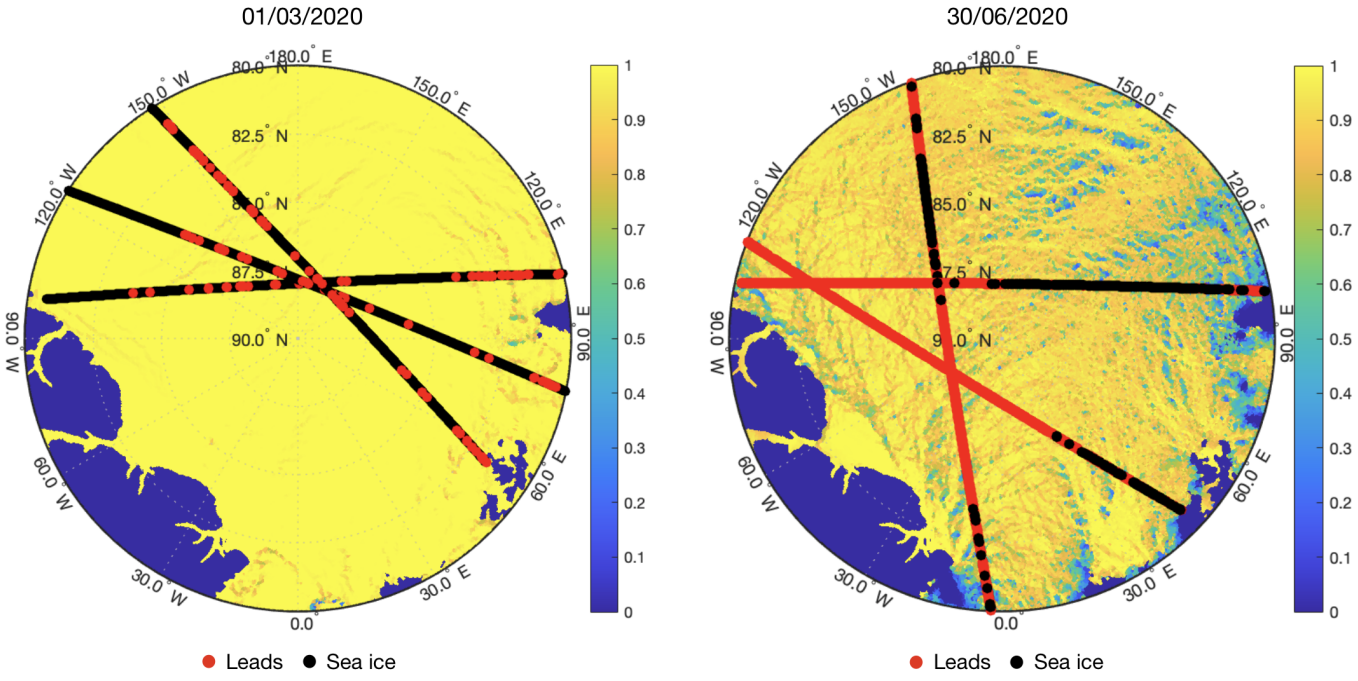
Figure B.1: Distribution of waveform features including off-nadir leads class, data from D-05 data set.

B.2. Dominance of specular waveform returns in summer

This section presents the additional study conducted using the SAR altimetry data acquired by CryoSat-2 in order to understand the location where specular returns are dominated in the Arctic Ocean. As stated in Section 3.1, it was speculated that the dominant specular returns in the summer months could have been due to the latitude of the study areas which could be more prone to ice melting. Because Sentinel-3 satellites are limited in its polar coverage, CryoSat-2 measurements were consulted such that the waveforms from higher latitudes could be studied, to test this hypothesis. For this analysis, the waveforms were classified to leads or sea ice by implementing the thresholding method used by Bij de Vaate (2019), since these thresholding values are more tailored to the CryoSat-2 waveforms.

Two images are presented in Figure B.2, showing the CryoSat-2 tracks from March (left) and June (right) in 2020. It is immediately clear that the waveform returns are dominated by sea ice in March, with occasional waveform returns suggesting leads. However, the image from June suggests otherwise. The waveform returns are mostly classified as leads, suggesting the dominance of specular returns. Furthermore, there are no clear correlation between the location where these specular reflections are dominating, which disproves the hypothesis. Therefore, it is concluded that using only SAR altimetry for lead detection in summer seasons is unsuitable.

Figure B.2: CryoSat-2 tracks in the Arctic region from 01/03/2020 (left) and 30/06/2020 (right). The red/black colors show the predicted class of the given point (leads, sea ice, respectively). The background image show the sea ice concentration model derived by the Arctic sea ice forecasting system (T. Williams, 2019)



B.3. Application of DBSCAN classifier

This study attempted to implement another unsupervised machine learning classifier called Density-Based Spatial Clustering of Applications with Noise (DBSCAN). However, due to its poor classification performances, it has been decided not to be considered for this research.

DBSCAN is a density-based clustering algorithm which is widely used in many applications. It groups together samples that are close together to each other, marking the outlier points that lie in the lower density regions (Ester et al., 1996). Once the algorithms finds the core samples of high density, the cluster is expanded. This expansion is dominated by ϵ which specifies how close the points should be to be included in the cluster. This algorithm works very good for data which contains clusters of similar density. However, if a data set has clusters of very different densities, the performance of the DBSCAN algorithm may suffer (Shah, 2012).

The test result of DBSCAN classifier for the winter months (D-01) is shown in Table B.1. The results are extremely poor, where almost all of the samples were classified as lead, producing both TLR and FLR close to 100%. This result did not improve when changing or optimizing the hyperparamter (ϵ). This suggests that the density of the clusters of leads and sea ice described by the waveform features differ significantly that the constant distance of ϵ is not able to pick up the clusters.

Table B.1: Results using DBSCAN classifier (Accuracy, TLR and FLR in %) for general winter performance (D-01)

Accuracy [%]	TLR [%]	FLR [%]
22.77	99.61	97.26

B.4. Verification & Validation of OLCI ground truth data

In order to verify the ground truth data generation method using the OLCI images, as described in Section 2.2.4, a manual verification has been conducted. In total, 1,982 labeled points which were produced from this data generation method are analyzed for this verification. These points were obtained by Sentinel-3A satellite in March/April of 2017 - 2020. From these selected study areas, false lead and false ice were counted. When the ground truth (based on OLCI image, seen by human eye) showed sea ice surfaces and OLCI ground truth label showed a lead, this was counted as false lead. In contrary, when the ground truth showed lead and OLCI ground truth label showed ice, this was counted as false ice.

Once the numbers of false leads and false ice are known, TLR, FLR and overall accuracies are computed by consulting the total number of ground truth ice and leads. These are presented in Table B.2. Most of the false leads and false ice were found at the edges of leads or locations where the radiance difference between leads and sea ice was not large. These areas are also hard to distinguish with human eye. Though the result is not perfect, the process showed very high accuracy and TLR values, and a very small FLR value. In conclusion, the ground truth generation method proposed in this study has produced successful results. It must also be noted that the classification performances based on the ground truth data (supervised learning and optimized thresholding) in this study will always limited by this accuracy of ground truth data.

Table B.2: Result of manual verification of OLCI validation process.

Total points	False Lead	False Ice	TLR [%]	FLR [%]	Overall Accuracy [%]
1982	18	4	95.51	1.12	98.89

This verification however does not provide how accurate OLCI image can depict the reality, as it is still limited by its own resolution. Rather, it provided insights on how accurate the OLCI ground truth data generation method was, with respect to if the points were to be labeled with manual inspection.

In order to understand the accuracy with respect to the reality, another data source must be used. This study attempted using the high resolution images obtained by the L1B Geolocated and Orthorectified Images taken by the Operation IceBridge (OIB) aircraft survey campaigns (Dominguez, 2018) for validation. These images have spatial resolution of from 0.015 m to 2.5 m (Dominguez, 2018).

In order to compare the two images, locations where OLCI image and images taken from OIB coincided with as little measurement time delay as possible were looked for. Figure B.3 shows an example of images taken by OIB on 19/4/2017 ~16:00 UTC and an image taken by OLCI on 19/4/2017 ~20:00 UTC, therefore these images have a measurement delay of approximately four hours. The colored squares shown in the OLCI image correspond to the exact locations where the OIB images (of the same color on the frame) were taken.

Figure B.3 shows that the leads found in the OIB images do not correspond to the leads seen in the OLCI image. If the four squares in the OLCI image in Figure B.3 were moved slightly towards the left-top corner, they would correspond to the location of the leads seen in the OLCI image. This suggests that even with a measurement time delay of four hours, the sea ice can be drifted significantly. An ice drifting model must be employed to analyze these data, however this is left unexplored in this study due to time constraints. This also highlights the importance of using measurements which have a temporal match (this study used OLCI and SRAL altimeter both on board of Sentinel-3 satellites to obtain a temporal match).

Furthermore, finding the coinciding locations of OIB and OLCI images are very difficult since the time difference between the two measurements could be simply too large and also many location had to be discarded due to the cloud covering in the OLCI image. Therefore, it is concluded that that images from OIB are not suited for validating the OLCI images. Other data such as space borne SAR images may be a better validation data source as the spatial coverage is much larger, and more coinciding measurement locations with OLCI images may be found.

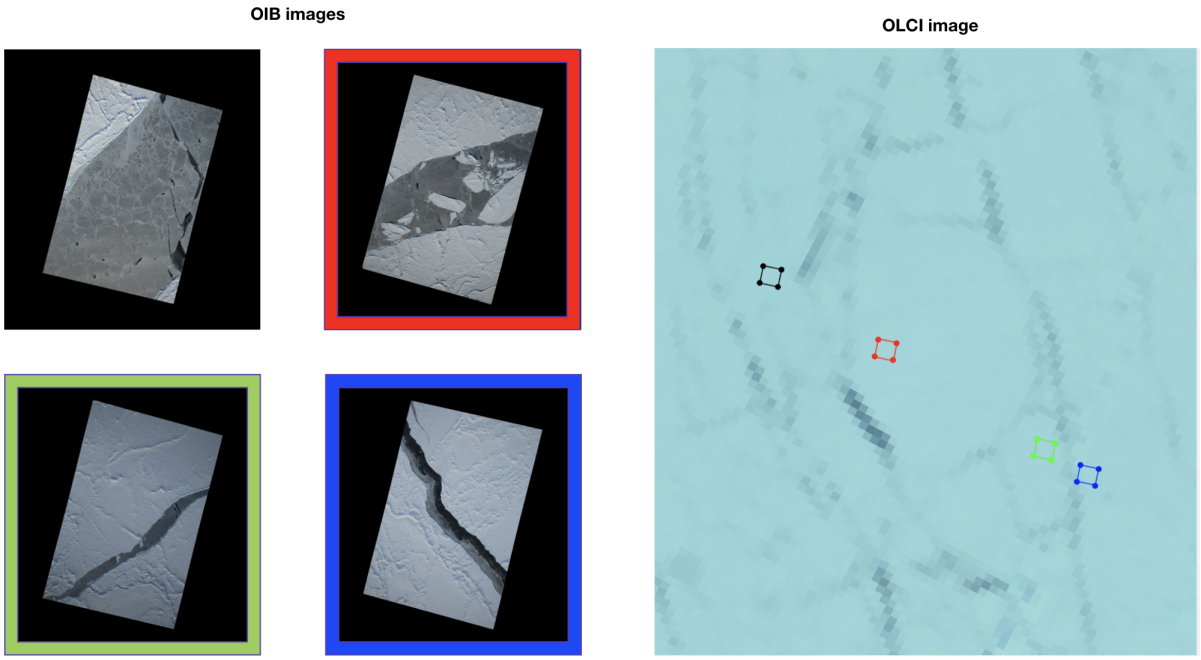


Figure B.3: Four images on the left show the L1B Geolocated and Orthorectified Images taken by OIB aircraft servery campaigns on 19/4/2017 ~20:00 UTC, whereas the right image taken by OLCI from Sentinel-3A satelltie on 19/4/2017 ~20:00 UTC. The colored squares shown in the OLCI image correspond to the exact locations where the four OIB images were taken.

Bibliography

A. Filippi, I. Dobreva, A.K.J.J., 2010. Self-Organizing Map-based Applications in Remote Sensing. Self-Organizing Maps doi:10.5772/9163.

Andersen, O., Knudsenand, P., Stenseng, L., 2015. The DTU13 MSS (mean sea surface) and MDT (mean dynamic topography) from 20 years of satellite altimetry. volume 144. in IGFS (International Associationof Geodesy Symposia), Cham, Switzerland: Springer. doi:https://doi.org/10.1007/1345_2015_182.

Armitage, W., Bacon, S., Ridout, A., Thomas, S., Aksenov, Y., Wingham, D., 2016. Arctic sea surface height variability and change from satellite radar altimetry and grace, 2003–2014. Journal of Geophysical Research: Oceans 121, 6762–6778. doi:10.1002/2015JC011579.

Barnhart, K.R., Overeem, I., Anderson, R.S., 2014. The effect of changing sea ice on the physical vulnerability of Arctic coasts. Cryosphere 8, 1777–1799. doi:10.5194/tc-8-1777-2014.

Bij de Vaate, I., 2019. Comparison between SAR altimeter-derived water levels and GTSM output for the Arctic region. Technical Report. Delft University of Technology.

Bij de Vaate, I., Vasulkaramd, A., Slobbe, D., Verlaan, M., 2021. The Influence of Arctic Landfast Ice on Seasonal Modulation of the M 2 Tide . Journal of Geophysical Research: Oceans 126. doi:10.1029/2020jc016630.

Bourg, L., Bruniquel, J., Morris, H., Dash, J., Preusker, R., Dransfeld, S., 2021. Copernicus Sentinel-3 OLCI Land User Handbook .

Breiman, L., 1996. Bagging Predictors. Machine Learning , 123–140.

Breiman, L., 2001. Random forests. Machine Learning , 5–32.doi:10.1201/9780367816377-11.

C. Lüpkes, T. Vihma, G.B.U.W., 2008. Influence of leads in sea ice on the temperature of the atmospheric boundary layer during polar night. Geophysical Research Letters 35, 2–6. doi:10.1029/2007GL032461.

De Swart, H.E., Zimmerman, J.T.F., 2009. Morphodynamics of tidal inlet systems. Annual Review of Fluid Mechanics 41, 203–229. doi:10.1146/annurev.fluid.010908.165159.

Dettmering, D., Wynne, A., F.L. Müller, M.P., Seitz, F., 2018. Lead detection in polar oceans-a comparison of different classification methods for Cryosat-2 SAR data. Remote Sensing 10. doi:10.3390/rs10081190.

Dinardo, S., Benveniste, J., 2013. Guidelines for the SAR (Delay-Doppler) L1b Processing. ESA document .

Dominguez, R., 2018. IceBridge DMS L1B Geolocated and Orthorectified Images, Version 1. Boulder, Colorado USA. NASA National Snow and Ice Data Center Distributed Active Archive Center. doi:https://doi.org/10.5067/OZ6VNOPMPRJ0.

Ester, M., Kriegel, H., Sander, J., Xu, X., 1996. A density-based algorithm for discovering clusters in large spatial databases with noise, in: Proceedings of the Second International Conference on Knowledge Discovery and Data Mining, AAAI Press. p. 226–231.

EUMETSAT, 2017. Sentinel-3 SRAL Marine User Handbook .

EUMETSAT, 2018. Sentinel-3 OLCI Marine User Handbook.

EUMETSAT, 2021. Summary of Expert Meeting for the study " Sentinel 3 Synergy Cloud Mask Development ". Technical Report v1.

European Space Agency, . SNAP - ESA Sentinel Application Platform v8.0.0 URL: http://step.esa.int.

European Space Agency, 2016. OLCI/Sentinel-3A L1 Full Resolution Top of Atmosphere Reflectance URL: https://cmr.earthdata.nasa.gov/search/concepts/C1286874966-LAADS.html. Accessed on 4/5/2021.

European Space Agency, CNES, 2020. Radar altimetry tutorial and toolbox; how altimetry works URL: http://www.altimetry.info/radar-altimetry-tutorial/how-altimetry-works/. accessed on 13/08/2020.

Fawcett, T., 2006. An introduction to ROC analysis. Pattern Recognition Letters 27, 861–874.

Freund, Y., Schapire, R., 1999. A Short Introduction to Boosting. Journal of Japanese Society for Artificial Intelligence , 771–780.

Friedman, N., Geiger, D., Goldszmidt, M., 1997. Bayesian Network Classifiers. Machine Learning , 131–163.

Giuffrida, G., Diana, L., de Gioia, L., Benelli, G., Meoni, G., Donati, M., Fanucci, L., 2020. CloudScout: A deep neural network for on-board cloud detection on hyperspectral images. Remote Sensing 12, 1–17. doi:10.3390/rs12142205.

Grossi, E., Massimo, B., 2007. Introduction to artificial neural networks. European Journal of Gastroenterology and Hepatology 19, 1046–1054. doi:10.1097/MEG.0b013e3282f198a0.

Hamada, M., Kanat, Y., Adejor, A., 2019. Sea ice drift in the Arctic since the 1950s. International Journal of Innovative Technology and Exploring Engineering 2, 1016–1019. doi:10.35940/ijitee.K1596.129219.

Hand, D., 1997. Construction and assessment of classification rules. Wiley Series in Probability and Statistics, Wiley.

Hastie, T., Tibshirani, R., Friedman, J., 2009. The Elements of Statistical Learning. volume 2. Springer-Verlagr, New York, NY, USA.

Jensen, J., 2005. Introductory Digital Image Processing: A Remote Sensing Perspective. volume 3. Pearson Education, Upper Saddle River.

Kaufman, L., Rousseeuw, P., 1987. Clustering by means of Medoids, in Statistical Data Analysis Based on the L1–Norm and Related Methods .

Kittel, C., Jiang, L., Tøttrup, C., Bauer-Gottwein, P., 2021. Sentinel-3 radar altimetry for river monitoring - A catchment-scale evaluation of satellite water surface elevation from Sentinel-3A and Sentinel-3B. Hydrology and Earth System Sciences 25, 333–357. doi:10.5194/hess-25-333-2021.

Kleinherenbrink, M., Naeije, M., Slobbe, C., Egidoc, A., Smith, W., 2020. The performance of CryoSat-2 fully-focussed SAR for inland water-level estimation. Remote Sensing of Environment 237.

Kokelj, S., Lantz, T., Solomon, S., Pisaric, F., Keith, D., Morse, P., Thienpont, J., Smol, J., Esagok, D., 2012. Using Multiple Sources of Knowledge to Investigate Northern Environmental Chan Regional Ecological Impacts of a Storm Surge in the Outer Mackenzie Delta, N.W.T. . Arctic , 257–272.

Kwok, R., Rothrock, D.A., 2009. Decline in Arctic sea ice thickness from submarine and ICESat records: 1958–2008. Geophysical Research Letters 36, 732–737.

Laxon, S., Giles, K., Ridout, A., Wingham, D., Willatt, R., Cullen, R., Kwok, R., Schweiger, A., Zhang, J., Haas, C., Hendricks, S., Krishfield, R., Kurtz, N., Farrell, S., , Davidson, M., 2013. CryoSat-2 estimates of Arctic sea ice thickness and volume. Geophysical Research Letters 40, 732–737. doi:10.1002/grl.50193.

Lee, S., Im, J., Kim, J., Kim, M., Shin, M., Kim, H., Quackenbush, L., 2016. Arctic Sea Ice Thickness Estimation from CryoSat-2 Satellite Data Using Machine Learning-Based Lead Detection. Remote Sensing 8. doi:10.3390/rs8090698.

Lee, S., Kim, H., Im, J., 2018. Arctic lead detection using a waveform mixture algorithm from CryoSat-2 data. Cryosphere 12, 1665–1679. doi:10.5194/tc-12-1665-2018.

Lindsay, R., Schweiger, A., 2015. Arctic sea ice thickness loss determined using subsurface, aircraft, and satellite observations. Cryosphere 9, 269–283. doi:10.5194/tc-9-269-2015.

Ludwig, V., Spreen, G., Pedersen, L., 2020. Evaluation of a New Merged Sea-Ice Concentration Dataset at 1 km Resolution from Thermal Infrared and Passive Microwave Satellite Data in the Arctic. Remote Sensing 3183. doi:10.3390/rs12193183.

MATLAB, 2020. Classification Learner App (R2020b) URL: https://www.mathworks.com/help/stats/classification-learner-app.html?s_tid=CRUX_lftnav.

Meier, W., Stroeve, J., 2020. State of the cryosphere: Sea ice URL: https://www.itl.nist.gov/div898/handbook/eda/section3/eda35b.htm. Accessed on 24/05/2020.

Metz, C., 1978. Basic Principles of ROC Analysis. Seminars in Nuclear Medicine VIII, 283–298.

Miljković, D., 2017. Brief Review of Self-Organizing Maps , 1061–1066.

Mohri, M., Rostamizadeh, A., Talwalkar, A., 2012. Foundations of Machine Learning. The MIT Press.

Morison, J., Kwok, R., Peralta-Ferriz, C., Alkire, M., Rigor, I., Andersen, R., Steele, M., 2012. Changing Arctic Ocean freshwater pathways. Nature 481, 66–70. doi:10.1038/nature10705.

Müller, F.L., Dettmering, D., Bosch, W., Seitz, F., 2017. Monitoring the arctic seas: How satellite altimetry can be used to detect openwater in sea-ice regions. Remote Sensing 9, 1–20. doi:10.3390/rs9060551.

Nielsen, F., 2016. Introduction to HPC with MPI for Data Science. Undergraduate Topics in Computer Science, Springer.

Notz, D., SIMIP Community, 2020. Arctic Sea Ice in CMIP6. Geophysical Research Letters 47. doi:10.1029/2019GL086749.

Passaro, M., Müller, F., Dettmering, D., 2018. Lead detection using Cryosat-2 delay-doppler processing and Sentinel-1 SAR images. Advances in Space Research 62, 1610–1625. doi:10.1016/j.asr.2017.07.011.

Peacock, N., Laxon, S., 2004. Sea surface height determination in the Arctic Ocean from ERS altimetry. Journal of Geophysical Research C: Oceans 109, 1–14. doi:10.1029/2001JC001026.

Petty, A., Kurtz, N., Kwok, R., Markus, T., T.A.Neumann, 2020. Winter Arctic Sea Ice Thickness From ICESat-2 Freeboards. Journal of Geophysical Research: Oceans 125, 1–28. doi:10.1029/2019JC015764.

Poisson, J., Quartly, G., Kurekin, A., P.Thibaut, Hoang, D., Nencioli, F., 2018. Development of an ENVISAT altimetry processor providing sea level continuity between open ocean and arctic leads. IEEE Transactions on Geoscience and Remote Sensing 56, 5299–5319. doi:10.1109/TGRS.2018.2813061.

Qin, A., Shi, S., Suganthan, P., Loog, M., 2005. Enhanced direct linear discriminant analysis for feature extraction on high dimensional data. Proceedings of the National Conference on Artificial Intelligence 2, 851–855.

Quartly, G., Rinne, E., Passaro, M., Andersen, O., Dinardo, S., Fleury, S., Guillot, A., Hendricks, S., Kurekin, A., Müller, F., Ricker, R., Skourup, H., Tsamados, M., Michel, 2019. Retrieving sea level and freeboard in the Arctic: A review of current radar altimetry methodologies and future perspectives. Remote Sensing 11. doi:10.3390/RS11070881.

Quinlan, J., 1986. Induction of decision trees. Machine Learning 1, 81–106. doi:10.1007/bf00116251.

Raney, R., 1998. The delay / doppler radar altimeter. IEEE Transactions on Geoscience and Remote Sensing 36, 1578–1588.

Ricker, R., Hendricks, S., Helm, V., Gerdes, R., 2015. Classification of cryosat-2 radar echoes. Towards an Interdisciplinary Approach in Earth System Science: Advances of a Helmholtz Graduate Research School , 149–158doi:10.1007/978-3-319-13865-7_17.

Ricker, R., Hendricks, S., Helm, V., Skourup, H., Davidson, M., 2014. Sensitivity of CryoSat-2 Arctic sea-ice freeboard and thickness on radar-waveform interpretation. Cryosphere 8, 1607–1622.

Rose, S., Forsberg, R., Pedersen, L., 2013. Measurements of sea ice by satellite and airborne altimetry. DTUSpace, National Space Institute URL: http://forskningsbasen.deff.dk/Share.external?sp=S2b70450d-62a5-4de2-b255-dec7fc9b5f8b{&}sp=Sdtu.

Savas, C., Dovis, F., 2019. The impact of different kernel functions on the performance of scintillation detection based on support vector machines. Sensors 19, 1–16. doi:10.3390/s19235219.

Schulz, A.T., Naeije, M., 2018. SAR Retracking in the Arctic: Development of a year-round retrackers system. Advances in Space Research 62, 1292–1306. URL: https://doi.org/10.1016/j.asr.2018.01.037, doi:10.1016/j.asr.2018.01.037.

Screen, J., Simmonds, I., 2010. The central role of diminishing sea ice in recent arctic temperature amplification. Nature 464, 1334–1337.

Seiffert, C., Khoshgoftaar, T., Van Hulse, J., Napolitano, A., 2008. RUSBoost: Improving classification performance when training data is skewed. International Conference on Pattern Recognition doi:10.1109/icpr.2008.4761297.

Shah, G., 2012. An improved DBSCAN, a density based clustering algorithm with parameter selection for high dimensional data sets. 3rd Nirma University International Conference on Engineering, NUiCONE 2012 , 6–8.doi:10.1109/NUICONE.2012.6493211.

Shen, X., Zhang, J., Zhang, X., Meng, J., Ke, C., 2017. Sea Ice Classification Using Cryosat-2 Altimeter Data by Optimal Classifier-Feature Assembly. IEEE Geoscience and Remote Sensing Letters 14, 1948–1952. doi:10.1109/LGRS.2017.2743339.

Shepherd, J., Bunting, P., Dymond, J., 2019. Operational Large-Scale Segmentation of Imagery Based on Iterative Elimination . Remote Sensing 658. doi:10.3390/rs11060658.

Shu, S., Zhou, X., Shen, X., Liu, Z., Tang, Z., Li, H., Ke, C., Li, J., 2020. Discrimination of different sea ice types from CryoSat-2 satellite data using an Object-based Random Forest (ORF). Marine Geodesy 43, 213–233. URL: https://doi.org/10.1080/01490419.2019.1671560, doi:10.1080/01490419.2019.1671560.

Slobbe, C., 2020. Fast4NL Project Description URL: https://www.fast4nl.nl/ProjectDescription.html. Accessed on 11/09/2020.

Stroeve, J., Holland, M., Meier, W., Scambos, T., Serreze, M., 2007. Arctic sea ice decline: Faster than forecast. Geophysical Research Letters 34, 213–233. URL: https://doi.org/10.1080/01490419.2019.1671560, doi:10.1080/01490419.2019.1671560.

Su, H., Ji, B., Wang, Y., 2019. Sea Ice extent detection in the Bohai Sea Using Sentinel-3 OLCI Data. Remote Sensing 11, 1–17. doi:10.3390/rs11202436.

T. Williams, A. Korosov, P.R.E., 2019. Presentation and evaluation of the Arctic sea ice forecasting system neXtSIM-F. The Cryosphere Discuss 37. doi:=10.5194/tc-2019-154.

Tangirala, S., 2020. Evaluating the impact of GINI index and information gain on classification using decision tree classifier algorithm. International Journal of Advanced Computer Science and Applications , 612–619doi:10.14569/ijacsa.2020.0110277.

Tilling, R., Kurtz, N., Bagnardi, M., Petty, A., Kwok, R., 2020. Detection of Melt Ponds on Arctic Summer Sea Ice From ICESat-2. Geophysical Research Letters 47, 1–10. doi:10.1029/2020GL090644.

Wadhams, P., 2016. The global impacts of rapidly disappearing arctic sea ice URL: https://e360.yale.edu/features/as_arctic_ocean_ice_disappears_global_climate_impacts_intensify_wadhams. Accessed on 23/09/2020.

Wang, M., Overland, J., 2012. A sea ice free summer arctic within 30 years: An update from CMIP5 models. Geophysical Research Letters 39, 1955–1968. doi:10.5194/tc-9-1955-2015.

Weeks, W.F., 2010. On Sea Ice. University of Alaska Press, Fairbanks. p. 664.

Wernecke, A., Kaleschke, L., 2015. Lead detection in Arctic sea ice from CryoSat-2: Quality assessment, lead area fraction and width distribution. Cryosphere 9, 1955–1968. doi:10.5194/tc-9-1955-2015.

Wingham, D., Francis, C., Baker, S., Bouzinac, C., Brockley, D., Cullen, R., de Chateau-Thierry, P., Laxon, S., Mallow, U., Mavrocordatos, C., Phalippou, L., Ratier, G., Rey, L., Rostan, F., Viau, P., Wallis, D., 2006. CryoSat: A mission to determine the fluctuations in Earth's land and marine ice fields. Advances in Space Research 37, 841–871. doi:10.1016/j.asr.2005.07.027.

Xia, W., Xie, H., Ke, C., 2014. Assessing trend and variation of Arctic sea-ice extent during 1979-2012 from a latitude perspective of ice edge. Polar Research 33:1. doi:10.3402/polar.v33.21249.

Xu, L., Li, J., Brenning, A., 2014. A comparative study of different classification techniques for marine oil spill identification using RADARSAT-1 imagery. Remote Sensing of Environment 141, 14–23. doi:10.1016/j.rse.2013.10.012.

Yang, J., Gong, P., Fu, R., Zhang, M., Liang, J.C.S., Xu, B., Shi, J., Dickinson, R., 2013. The role of satellite remote sensing in climate change studies. Nature Climate Change 3, 875–883. doi:10.1038/nclimate1908.

Zygmuntowska, M., Khvorostovsky, K., Helm, V., Sandven, S., 2013. Waveform classification of airborne synthetic aperture radar altimeter over Arctic sea ice. The Cryosphere 7, 1315–1324. doi:10.5194/tc-7-1315-2013.