DEVELOPMENT OF BENTHIC MONITORING APPROACHES FOR SALMON AQUACULTURE SITES USING MACHINE LEARNING, HYDROACOUSTIC DATA AND BACTERIAL eDNA

by © Ethan Gerald Armstrong

A thesis submitted

to the school of Graduate Studies in partial fulfillment of the

requirements for the degree of

Master of Science (Biology), Faculty of Science

Memorial University of Newfoundland

June 2019

St John’s, Newfoundland and Labrador

Abstract

Intensive caged salmon production can lead to localized perturbations of the seafloor environment where organic waste (flocculent matter) accumulates and disrupts ecological processes. As the aquaculture industry expands, the development of tools to rapidly detect changes in seafloor condition is critical. Here, we examine whether applying machine learning to two types of monitoring data could improve environmental assessments at aquaculture sites in Newfoundland. First, we apply machine learning to single beam echosounder data to detect flocculent matter at aquaculture sites over larger areas than currently achieved used drop camera imaging. Then, we use machine learning to categorize sediments by levels of disturbance based on bacterial tetranucleotide frequency distributions generated from environmental DNA. While echosounder data can detect flocculent matter with moderate success in this region, bacterial tetranucleotide frequencies are highly effective classifiers of benthic disturbance; this simplified environmental DNA-based approach could be implemented within novel aquaculture benthic monitoring pipelines.

Acknowledgments

I would like to thank Drs. Suzanne Dufour, Dounia Hamoutene and Flora Salvo as members of my supervisory committee for their trust and counsel over the past two years. Giving me the independence to follow my interests has led to opportunities I could not have imagined two years ago.

I would like to acknowledge the sources of funding that made this research possible including assistance from the Ocean Frontier Institute and the Program for Aquaculture Regulatory Research.

I extend my thanks to the benevolent strangers of the Stack Overflow community. Your willingness to share your expertise in statistics and machine learning is a demonstration of what is right about the internet.

I would like to thank my family and friends who encouraged and supported me and who never refrained from questioning my assumptions. I am a better person and more critical thinker because you took the time to listen and disagree with me. Finally, I would especially like to acknowledge my partner, Sherri Bowes for her patience, reassurance and unwavering strength of character. I love you and look forward to seeing the amazing things you will accomplish.

Table of Contents

List of Tables

List of Figures

List of Abbreviations and Symbols

DFO – Department of Fisheries and Oceans

SBE/SBEs – Single beam echosounder(s)

MBE/MBEs – Multibeam echosounder(s)

SML – Supervised machine learning

NL - Newfoundland

eDNA – environmental DNA

CTD – Current, temperature and depth oceanography instrument

TNF – Tetranucleotide frequency

CV – Cross validation

OTU – Operational taxonomic unit

IDE – Integrated development environment

Chapter 1. Introduction and overview

1.1 Aquaculture growth on global, national and provincial scales

Global aquaculture production has experienced significant growth in response to increased demand and collapses of conventional fisheries (Naylor et al., 2000; Asche et al., 2008). In Canada, the aquaculture industry underwent a 63% increase in production value between 2003 and 2013 (FAO, 2016). Much of this growth has focussed on finfish, and particularly Atlantic Salmon, of which Canada is the world's 3rd largest producer (Chopin, 2015; Agriculture and Agri-Food Canada, 2017). The growth of the aquaculture industry in Newfoundland (NL) has mirrored global and national trends, seeing a 100-fold increase in revenue over the last two decades (Government of Newfoundland and Labrador, 2016, 2017). NL salmon aquaculture is done in open water net pens along the south coast in deep coastal bays (depths > 30 m) with complex topography and predominately hard bottom substrates (Hamoutene et al., 2013, 2015; Hamoutene, 2014). The production cycle consists of a 1 to 2-year growth period where salmon are fed a controlled diet (Hamoutene et al., 2013, 2015; Hamoutene, 2013). After harvesting, aquaculture sites may be restocked immediately or remain unoccupied for some time, allowing the seafloor to recover from the effects of aquaculture production in a process known as fallowing (Fisheries and Oceans Canada, 2015).

1.2 Impacts of aquaculture production

Intensive finfish aquaculture is associated with several environmental impacts including disease transmission (Arechavala-Lopez et al., 2013; Torrissen et al., 2013), deleterious

interactions of escaped fish with wild populations (Jensen et al., 2010; Arechavala-Lopez et al., 2013; Baskett et al., 2013) and changes to seafloor condition in a process referred to as organic enrichment (Hargrave et al., 1997, 2008; Crawford, 2003). Organic enrichment occurs as flocculent matter, a gel-mud consisting of uneaten fish-feed, faeces, other organic compounds and heavy metals, is deposited on the seafloor (Sather et al., 2006; Salvo et al., 2015). Organic enrichment and its subsequent changes to seafloor macrofaunal and microbial communities is well documented, with previously existing species and diversity declining as opportunistic species tolerant to enriched and hypoxic conditions colonize impacted areas (Pearson & Rosenberg, 1978; Pohle et al., 2001; Caravalho et al., 2006; Borja et al., 2008). With sustained levels of organic enrichment, native infauna may be eliminated (Keeley et al., 2014; Stoeck et al., 2018). Opportunistic species function as indicators of organic enrichment when performing environmental assessments at aquaculture sites (Hargrave et al., 2008; Hamoutene et al., 2013; Hamoutene, 2014).

Aquaculture production is paused in a process known as fallowing which allows the benthos to recover from aquaculture production (Zhulay et al., 2015). Benthic recovery in this context refers to the return of either benthic faunal or microbial diversity to pre-aquaculture conditions. Determining an appropriate duration of fallowing and its effectiveness is an area of active study, with partial recovery of some benthic faunal communities reported to occur between six and twenty-four months post-production (Lin & Bailey-Brock, 2008; Keeley et al., 2015), and total recovery requiring over two years (Keeley et al., 2014). Contrasting studies examining microbial communities have found

little evidence of benthic recovery, even after 35 months (Verhoeven et al., 2018), suggesting that variable results may stem from differences in study site locations or the diverse methods used in these assessments. The reintroduction of fish to an aquaculture site can result in rapid deterioration in seafloor condition, making the calibration of fallowing periods and the reintroduction of aquaculture production challenging (Keeley et al., 2015).

1.3 Current monitoring practices

Canadian regulations for monitoring aquaculture-derived organic enrichment include drop camera surveys and grab sampling (DFO, 2018). Drop cameras capture a 0.5 m x 0.5 m field of view whereas grab sampling collects the top 2 cm of sediment to measure various physicochemical parameters including redox, free sulphides and organic matter content (DFO, 2018). Images and sediment data are collected at stations along sampling transects whose length and orientation differ by region (DFO, 2018). In NL, visual monitoring is conducted along a minimum of 6 transects which extend 100 m from the cage array, with stations in 20 m increments (DFO, 2018).

Current monitoring practices cannot feasibly provide regulators with information regarding the far-field effects (beyond 100 m from cages) or the spatial extent of organic matter deposition, which may extend more than a kilometer from cage sites (Broch et al., 2017). While image capture is useful in detecting large-scale changes in epibenthic communities (Hamoutene et al., 2015), it cannot resolve intermediate changes in seafloor condition, and evaluations of benthic health are restricted to the visible layer of sediment,

which may not reflect true conditions (Siwabessy et al., 2013). Sediment physicochemical analysis is more sensitive to small scale changes in condition, but sediment collection using grabs is difficult in NL due to the predominance of hard bottom seafloors (Hamoutene et al., 2015; Donnet et al., 2018). Additionally, the sampling of flocculent matter at aquaculture sites relies in part on operator expertise to determine likely areas where organic material has accumulated, which is made difficult by the complex oceanographic and bathymetric conditions in NL (Salcedo-Castro & Ratsimandresy, 2013; Hamoutene et al., 2015).

1.4 Alternative approaches to benthic monitoring

Alternative approaches to benthic monitoring such as the use of hydroacoustic data, bacterial communities, and eDNA have been investigated to address limitations inherent in current monitoring protocols (Hughes et al., 2002; Wildish et al., 2004; Keeley et al., 2018; Stoeck et al., 2018; Verhoeven et al., 2018). Hydroacoustic approaches use echosounders affixed to ships or small vessels to collect continuous data of the seafloor as the vessel moves along transects, or in a grid pattern. Echosounders function by emitting a ping, or a series of acoustic signals at defined frequencies through the water column (Clark, 2018). As the acoustic signal reaches the seafloor, a certain amount of energy is reflected and returned to a transducer, which converts this signal to an interpretable format which can be represented in an echogram. Substrate types can be classified by measuring differences in return energy and other echogram characteristics such as fractal dimension, which serves as a measure of seafloor roughness (Clark, 2018). Hydroacoustic approaches are potentially useful in the context of benthic monitoring at

aquaculture sites as organic matter deposits could affect the acoustic signal in a consistent manner. As hydroacoustic data are continuously collected while the vessel moves along transects or in grids, the spatial extents of impacted seafloor could be more accurately defined than is currently achievable using widely spaced point sampling techniques. Hydroacoustic approaches have been successfully used to detect organic matter deposition at aquaculture sites using multibeam echosounders (MBEs) (Wildish et al., 2004) and side-scan acoustic imagery (Hughes et al., 2002). Such approaches, however, have not been evaluated in NL, which is predominated by hard bottom substrates as opposed to the soft bottom seafloors characterizing in the aforementioned studies.

Bacterial communities are sensitive to environmental changes introduced by aquaculture production (Nogales et al., 2011; Keeley et al., 2018; Stoeck et al., 2018), with work conducted at aquaculture sites in NL further demonstrating their potential as organic enrichment bioindicators (Verhoeven et al., 2016, 2018). Bacterial approaches typically begin with the collection of sediment samples. Environmental DNA (eDNA) is extracted from sediment and sequenced fragments of bacterial DNA are compared to a reference database which aligns sample sequences with those on record to assign a taxonomic designation (Cristescu, 2014; Apothéloz-Perret-Gentil et al., 2017). Often however, eDNA sequences present in samples are not found in reference databases meaning that comparisons between related bacteria are restricted to the phylum level (Cristescu, 2014; Apothéloz-Perret-Gentil et al., 2017). Operational taxonomic units (OTUs) can be constructed to overcome this limitation by creating clusters of sequences with user defined similarities (Mahé et al., 2015; Kopylova et al., 2016). OTUs however, are

dataset dependent, meaning that OTUs derived from different data sources cannot be compared (Callahan et al., 2017). Machine learning applied to bacterial eDNA could help streamline data analysis and has been successfully used in previous studies examining aquaculture impacts (Cordier et al., 2017, 2018).

Hydroacoustic and bacterial approaches have distinct objectives when applied to environmental monitoring at aquaculture sites. Hydroacoustic approaches indicate the presence of organic matter deposits while bacterial approaches quantify benthic impacts and community level changes associated with aquaculture derived organic enrichment.

1.5 Thesis objectives

The objective of this thesis is to examine two new monitoring techniques to overcome present limitations in NL aquaculture monitoring protocols. The first technique involves the collection of hydroacoustic data and machine learning algorithms to improve the detection of aquaculture-derived organic matter deposition. The second technique uses bacterial eDNA sequences to quantify benthic impacts and intermediate changes in seafloor condition.

The first chapter of this thesis introduces a method of predicting flocculent matter deposition in proximity to aquaculture sites using hydroacoustic data collected with a single beam echosounder (SBE). SBE data is used to train machine learning models which generalize well to unseen data. The purpose of this chapter is to outline best practices of model creation, validation and testing in complex oceanographic conditions, enabling regulators to rapidly assess the spatial bounds of organic matter deposition.

Overall, this work intends to overcome scale restrictions inherent in point sampling techniques to more rapidly assess the spatial extent of aquaculture-derived organic matter deposition. Additionally, the developed method is intended to be easily deployed by both regulators and industry members to monitor aquaculture impacts and its change over time, as well as validating the use of SBE which is a cheaper alternative to more commonly used MBEs. SBEs have been used to characterized benthic habitats like seagrass beds (Komatsu et al., 2002; Preston et al., 2006; Quintino et al., 2009) and discriminate substrate types (Bates & Whitehead, 2001; Foster et al., 2009; Lee & Lin, 2018)

The second chapter investigates the use of oligonucleotide frequencies derived from bacterial eDNA sequences in the development of a benthic monitoring pipeline which uses machine learning to predict four levels of benthic disturbance. Taxonomy-based approaches have demonstrated that bacterial communities are sensitive indicators of aquaculture-derived organic enrichment in NL (Verhoeven et al., 2016, 2018**)** and other regions with comparable oceanographic conditions (Stoeck et al., 2018). Current machine learning approaches applied to eDNA to predict ecological condition are computationally demanding and restricted in their ability to generalize to new study sites. By examining oligonucleotide frequencies, the objective of this chapter is to develop a uniform feature set which reduces computation time and describe a modelling pipeline which quickly outputs disturbance classifications. The proposed method provides regulators with a fast and flexible tool that can monitor changes in seafloor condition.

1.6 References

Agriculture and Agri-Food Canada (2017). *Sector Trend Analysis - Salmon Trends in the European Union*. Retrieved from http://www.agr.gc.ca/eng/industry-markets-and-trade/international-agri-food-market-intelligence/europe/market-intelligence/sector-trend-analysis-salmon-trends-in-the-european-union/?id=1498663105849

Apothéloz-Perret-Gentil, L., Cordonier, A., Straub, F., Iseli, J., Esling, P., & Pawlowski, J. (2017). Taxonomy-free molecular diatom index for high-throughput eDNA biomonitoring. *Molecular Ecology Resources*, 17, 1231-1242.

Arechavala-Lopez, P., Sanchez-Jerez, P., Bayle-Sempere, J. T., Uglem, I., & Mladineo, I. (2013). Reared fish, farmed escapees and wild fish stocks—a triangle of pathogen transmission of concern to Mediterranean aquaculture management. *Aquaculture Environment Interactions*, 3, 153-161.

Asche, F., Roll, K. H., & Tveterås, S. (2008). Future trends in aquaculture: productivity growth and increased production. In M. Holmer et al. (Eds.), *Aquaculture in the Ecosystem* (pp. 271-292). Springer, Dordrecht.

Baskett, M. L., Burgess, S. C., & Waples, R. S. (2013). Assessing strategies to minimize unintended fitness consequences of aquaculture on wild populations. *Evolutionary Applications*, 6, 1090-1108.

Bates, C. R., Whitehead, E. J., & Castle, B. (2001). Echo Plus measurements in Hopavagen Bay, Norway. *Sea Technology*, 42, 34-43.

Callahan BJ, McMurdie PJ & Holmes SP (2017) Exact sequence variants should replace operational taxonomic units in marker-gene data analysis. *The ISME Journal*, 11, 2639–2643.

Chopin, T. (2015). Marine aquaculture in Canada: well-established monocultures of finfish and shellfish and an emerging Integrated Multi-Trophic Aquaculture (IMTA) approach including seaweeds, other invertebrates, and microbial communities. *Fisheries*, 40, 28-31.

Clarke, J. E. H. (2018). Multibeam echosounders. In A. Micaleff et al. (Eds.), *Submarine Geomorphology* (pp. 25-41). Springer, Cham.

Cordier, T., Esling, P., Lejzerowicz, F., Visco, J., Ouadahi, A., Martins, C., Cedhagen, T. & Pawlowski, J. (2017). Predicting the ecological quality status of marine environments from eDNA metabarcoding data using supervised machine learning. *Environmental Science & Technology*, 51, 9118-9126.

Cordier, T., Forster, D., Dufresne, Y., Martins, C. I., Stoeck, T., & Pawlowski, J. (2018). Supervised machine learning outperforms taxonomy-based environmental DNA metabarcoding applied to biomonitoring. *Molecular Ecology Resources*, 18, 1381-1391.

Crawford, C. (2003). Environmental management of marine aquaculture in Tasmania, Australia. *Aquaculture*, 226, 129-138.

Cristescu, M. E. (2014). From barcoding single individuals to metabarcoding biological communities: towards an integrative approach to the study of global biodiversity. *Trends in Ecology & Evolution*, 29, 566-571.

Donnet, S., Ratsimandresy, A.W., Goulet, P., Doody, C., Burke, S., & Cross, S. (2018) Coast of Bays Metrics: Geography Hydrology and Physical Oceanography of Aquaculture Area of the South Coast of Newfoundland. DFO. *Canadian Science Advisory Secretariat Research Document*. 2017/076. x+109 p.

Fisheries and Oceans Canada (2015). Aquaculture Activities Regulations Guidance Document. Retrieved from http://www.dfo-mpo.gc.ca/aquaculture/management-gestion/aar-raa-gd-eng.htm.

Foster, G., Walker, B. K., & Riegl, B. M. (2009). Interpretation of single beam acoustic backscatter using lidar-derived topographic complexity and benthic habitat classifications in a coral reef environment. *Journal of Coastal Research*, 16-26.

Government of Newfoundland and Labrador (2016). *Economic Impacts of the Newfoundland and Labrador Aquaculture Industry.* Retrieved from http://www.fishaq.gov.nl.ca/publications/pdf/Aquaculture_Macro_FINAL.pdf

Government of Newfoundland and Labrador (2017). *The Economic Review 2017.* Retrieved from https://www.economics.gov.nl.ca/pdf2017/theeconomicreview2017.pdf

Hamoutene, D. (2014). Sediment sulphides and redox potential associated with spatial coverage of *Beggiatoa* spp. at finfish aquaculture sites in Newfoundland, Canada. *ICES Journal of Marine Science*, 71, 1153-1157.

Hamoutene, D., & Mabrouk, G., Sheppard, L., MacSween, C., Coughlan, E. & Grant, C. (2013). Validating the use of *Beggiatoa* sp. and opportunistic polychaete worm complex (OPC) as indicators of benthic habitat condition at finfish aquaculture

sites in Newfoundland. *Canadian Technical Report of Fisheries and Aquatic Sciences*, 3028, 1-26.

Hamoutene, D., Salvo, F., Bungay, T., Mabrouk, G., Couturier, C., Ratsimandresy, A., & Dufour, S. C. (2015). Assessment of finfish aquaculture effect on Newfoundland epibenthic communities through video monitoring. *North American Journal of Aquaculture*, 77, 117-127.

Hargrave, B. T., Holmer, M., & Newcombe, C. P. (2008). Towards a classification of organic enrichment in marine sediments based on biogeochemical indicators. *Marine Pollution Bulletin*, 56, 810-824.

Hargrave, B. T., Phillips, G. A., Doucette, L. I., White, M. J., Milligan, T. G., Wildish, D. J., & Cranston, R. E. (1997). Assessing benthic impacts of organic enrichment from marine aquaculture. In B. Kronvang et al. (Eds.), *The Interactions Between Sediments and Water* (pp. 641-650). Springer, Dordrecht.

Jensen, Ø., Dempster, T., Thorstad, E. B., Uglem, I., & Fredheim, A. (2010). Escapes of fishes from Norwegian sea-cage aquaculture: causes, consequences and prevention. *Aquaculture Environment Interactions*, 1, 71-83.

Keeley NB, Forrest BM, Macleod CK (2015) Benthic recovery and re-impact responses from salmon farm enrichment: Implications for farm management. *Aquaculture*, 435, 412–423

Keeley, N. B., Macleod, C. K., Hopkins, G. A., & Forrest, B. M. (2014). Spatial and temporal dynamics in macrobenthos during recovery from salmon farm induced

organic enrichment: When is recovery complete?. *Marine Pollution Bulletin*, 80, 250-262.

Keeley, N., Wood, S. A., & Pochon, X. (2018). Development and preliminary validation of a multi-trophic metabarcoding biotic index for monitoring benthic organic enrichment. *Ecological Indicators*, 85, 1044-1057.

Komatsu, T., Igararashi, C., Tatsukawa, K. I., Nakaoka, M., Hiraishi, T., & Taira, A. (2002). Mapping of seagrass and seaweed beds using hydro-acoustic methods. *Fisheries Science*, 68, 580-583.

Kopylova, E., Navas-Molina, J. A., Mercier, C., Xu, Z. Z., Mahé, F., He, Y., Zhou, H., Rognes, T., Caporaso, G. & Knight, R. (2016). Open-source sequence clustering methods improve the state of the art. *mSystems*, 1, e00003-15.

Lee, W. S., & Lin, C. Y. (2018). Mapping of tropical marine benthic habitat: Hydroacoustic classification of coral reefs environment using single beam (RoxAnn™) system. *Continental Shelf Research*, 170, 1-10.

Lin, D. T., & Bailey-Brock, J. H. (2008). Partial recovery of infaunal communities during a fallow period at an open-ocean aquaculture. *Marine Ecology Progress Series*, 371, 65-72.

Naylor, R. L., Goldburg, R. J., Primavera, J. H., Kautsky, N., Beveridge, M. C., Clay, J., Folke, C., Lubchenco, J., Mooney, H., & Troell, M. (2000). Effect of aquaculture on world fish supplies. *Nature*, 405, 1017.

Nogales, B., Lanfranconi, M. P., Piña-Villalonga, J. M., & Bosch, R. (2011). Anthropogenic perturbations in marine microbial communities. *FEMS Microbiology Reviews*, 35, 275-298.

Pearson, T. H., & Rosenberg, R. (1978). Macrobenthic succession in relation to organic enrichment and pollution of the marine environment. *Oceanography and Marine Biology Annual Review*, 16, 229-311.

Preston, J., Inouchi, Y., & Shioya, F. (2006). Acoustic classification of submerged aquatic vegetation. In *Proceedings of the Eighth European Conference on Underwater Acoustics*, ECUA (p. 317e322).

Mahé, F., Rognes, T., Quince, C., de Vargas, C., & Dunthorn, M. (2015). Swarm v2: highly-scalable and high-resolution amplicon clustering. *PeerJ*, 3, e1420.

Salcedo-Castro, J., & Ratsimandresy, A. W. (2013). Oceanographic response to the passage of hurricanes in Belle Bay, Newfoundland. Estuarine, *Coastal and Shelf Science*, 133, 224-234.

Salvo, F., Hamoutene, D., & Dufour, S. C. (2015). Trophic analyses of opportunistic polychaetes (Ophryotrocha cyclops) at salmonid aquaculture sites. *Journal of the Marine Biological Association of the United Kingdom*, 95, 713-722.

Sather, P. J., Ikonomou, M. G., & Haya, K. (2006). Occurrence of persistent organic pollutants in sediments collected near fish farm sites. *Aquaculture*, 254, 234-247.

Siwabessy, P. J. W., Daniell, J., Li, J., Huang, Z., Heap, A. D., Nichol, S., Anderson, T.J., & Tran, M. (2013). Methodologies for seabed substrate characterisation using

multibeam bathymetry, backscatter and video data: A case study from the carbonate banks of the Timor Sea, Northern Australia. Geoscience Australia.

Stoeck, T., Frühe, L., Forster, D., Cordier, T., Martins, C. I., & Pawlowski, J. (2018). Environmental DNA metabarcoding of benthic bacterial communities indicates the benthic footprint of salmon aquaculture. *Marine Pollution Bulletin*, 127, 139-149.

Torrissen, O., Jones, S., Asche, F., Guttormsen, A., Skilbrei, O. T., Nilsen, F., Horsberg, T., E. & Jackson, D. (2013). Salmon lice–impact on wild salmonids and salmon aquaculture. *Journal of Fish Diseases*, 36, 171-194.

Quintino, V., Freitas, R., Mamede, R., Ricardo, F., Rodrigues, A. M., Mota, J., Pérez-Ruzafa, Á., & Marcos, C. (2009). Remote sensing of underwater vegetation using single beam acoustics. *ICES Journal of Marine Science*, 67, 594-605.

Verhoeven JTP, Salvo F, Hamoutene D & Dufour SC (2016) Bacterial community composition of flocculent matter under a salmonid aquaculture site in Newfoundland, Canada. *Aquaculture Environment Interactions*, 8, 637–646.

Verhoeven, J. T. P., Salvo, F., Knight, R., Hamoutene, D., & Dufour, S. (2018). Temporal bacterial surveillance of salmon aquaculture sites indicates a longlasting benthic impact with minimal recovery. *Frontiers in Microbiology*, 9, 3054.

Zhulay, I., Reiss, K., & Reiss, H. (2015). Effects of aquaculture fallowing on the recovery of macrofauna communities. *Marine Pollution Bulletin*, 97, 381-390.

Co-authorship Statement

The following people and institutions contributed to the work undertaken in this thesis:

Ethan Armstrong, Memorial University
Dr. Suzanne Dufour, Memorial University, Co-supervisor
Dr. Dounia Hamoutene, Fisheries and Oceans Canada, Co-supervisor
Dr. Flora Salvo, Fisheries and Oceans Canada, Supervisor
Dr. Joost Verhoeven, Memorial University, Co-author

Chapter 2, Investigating the use of machine learning and single beam echosounders to detect a benthic aquaculture footprint on hard bottom substrates:

I am the primary author, with co-authors Dufour, Hamoutene and Salvo contributing to overall study design and assisting with manuscript preparation and revisions. Hamoutene and Salvo collected data used in the study; funding for field work was obtained by Hamoutene. I created the dataset and predictive models and analysed the results of predictive modelling. I created all figures present in this chapter.

Chapter 3, Machine learning mediated benthic aquaculture impact assessment using oligonucleotide frequencies

This chapter is co-authored by myself and Verhoeven. I contributed to study design, created predictive models and figures, and wrote the majority of the text. Verhoeven performed the bioinformatics analysis (generated oligonucleotide frequency data from bacterial eDNA sequences) and organized the data that I used in predictive models. He contributed text for the introduction and approximately 40% of the methods section and revised the chapter. Dufour, Hamoutene and Salvo participated in study design and provided editorial comments.

Chapter 2. Investigating the use of machine learning and single beam echosounders to detect a benthic aquaculture footprint on hard bottom substrates

2.1 Introduction

2.1.1 Overview of aquaculture production and industry growth

The combined effects of fisheries collapse and increased demand for seafood products has resulted in the significant expansion of global aquaculture production (Naylor et al., 2000; Asche et al., 2008). Aquaculture grew an average of 3.2% per year between 1961 and 2013 (FAO, 2016). In 2013, global aquaculture production exceeded that of conventional fisheries, validating predictions made regarding the source of consumed seafood (Costa-Pierce, 2010; FAO, 2016). The growth of the Canadian aquaculture industry reflects global trends, with production focussed on finfish, especially Atlantic salmon (Chopin, 2015; DFO, 2016).

In Newfoundland (NL), the expansion of aquaculture is perhaps even more pronounced: it has grown from a 3 to 300-million-dollar industry in just over two decades, and the provincial government has introduced plans to double production by 2020 (Government of Newfoundland and Labrador, 2016, 2017). Most aquaculture production in NL occurs on the island's southern coast and involves the cultivation of Atlantic salmon in open water net-pens. The NL south coast contains deep water fjords, deep and wide bays, and shallow regions with exposure to the open ocean. Overall, the south coast is characterized by a predominately hard but heterogenous seafloor (Anderson et al., 2005; Hargrave et al., 2008; Hamoutene, 2014; Hamoutene et al., 2015). The depth of

aquaculture sites in this region exceeds 30 m, and depth range within a site can be broad (e.g. 50-100 m) given the complex coastal bathymetry (Donnet et al., 2018).

2.1.2 Benthic impacts of aquaculture production

The principal benthic impact associated with finfish aquaculture is organic enrichment, which occurs as flocculent matter is deposited on the seafloor (Crawford, 2003; Holmer et al., 2008; Salvo et al., 2015). Flocculent matter is composed of uneaten fish-feed pellets, expelled faeces, microbes and other organic compounds derived from open water cage sites (Salvo et al., 2015). Based on its appearance during benthic imaging or in grab samples, flocculent matter differs physically from natural sediments: it is less compact, has a fluffy or gelatinous texture, and can be easily resuspended. Flocculent matter-linked organic enrichment results in changes to seafloor communities near aquaculture sites (Pohle et al., 2001; Caravalho et al., 2006; Borja et al., 2008) primarily driven by increased biological oxygen demand (Wildish and Pohle, 2005; Fodelianakis et al., 2015). Organic enrichment also results in changes in seafloor geochemistry (Holmer et al., 2005), further decreasing oxygen availability (Mazzola et al., 2000). The development of hypoxic conditions driven by enriched deposits can negatively affect pre-existing benthic communities (Jusup et al., 2009; Pochon et al., 2015). As conditions deteriorate, species tolerant of enriched and hypoxic conditions colonize these areas and are used by regulators as visual indicators of organic enrichment (Hargrave et al., 2008; Hamoutene, 2014; Salvo et al., 2014; Hamoutene et al., 2015). In NL, regulation is based on the proportion of drop-camera sampling stations that show visual indicators (bacterial

mats or opportunistic polychaetes), or that are barren (i.e. where epibenthic organisms were observed pre-aquaculture but are no longer present) (DFO, 2018).

2.1.3 Justification for the development a new monitoring technique

Monitoring techniques currently used in Canadian regulation to detect organic enrichment at aquaculture sites consist of drop camera surveys where substrates are predominantly hard, and grab sampling to measure redox, free sulphide and organic matter content where substrates are softer (DFO, 2018). Both are point sampling methods that assess small areas of the seafloor along transects: drop cameras must show a 0.5 m x 0.5 m field of view (DFO, 2018), and a comparable area of the seafloor area is typically sampled by most sediment grabs. Both the application of these methods and the length and spacing of stations along sampling transects differ by region (DFO, 2018).

Current point sampling methods return discrete, small-scale snapshots of benthic conditions rather than a broader view of the seafloor. Point sampling methods are time intensive, costly and do not give a continuous measure of the footprint. Moreover, sampling is restricted to the length of predetermined transects and cannot provide information on far-field organic deposition or the maximal spatial extent of aquaculture production impacts, which may extend a kilometer or more from cage sites (Broch et al., 2017). Point sampling techniques do, however, provide high-resolution data in relation to epibenthic taxon richness and have been shown to resolve large scale changes in seafloor condition (Hamoutene et al., 2015). Continuous imaging of the seafloor along transects using ROVs could enable the visualization of a greater seafloor area, but ROV

use has been problematic in NL due to the risk of entanglement in cage site anchor lines or other submerged hazards (Mabrouk et al., 2014, Salvo pers. comm.). There is currently no single, standard protocol in Canada that applies to every region, and the development of a universally applicable method would provide a more robust assessment of seafloor condition by allowing regulators to directly compare conditions in different parts of the country. Therefore, a method that could survey a wider area, cheaply and in less time would benefit both regulators and those required to perform environmental assessments.

2.1.4 Single beam echosounders (SBEs)

Single beam echosounders (SBEs) are a potential solution to the aquaculture environmental monitoring problem as they can survey larger areas more rapidly than currently used point sampling methods and can be deployed from small vessels. SBE transducers are either attached to a pole or to the hull of a vessel and continuously collect acoustic data from a series of points while the ship moves along transects. At each point, SBEs emit a single ping at a set frequency and beamwidth and receive one return measurement. Measurements of energy return can be used to discriminate various water column components such as schooling fish (Reid et al., 2000), seafloor features such as seagrass beds (Komatsu et al., 2002; Preston et al., 2006; Quintino et al., 2009) and different substrate types (Bates & Whitehead, 2001; Foster et al., 2009; Li et al., 2011; Lee & Lin, 2018), as harder substrates reflect more acoustic energy than soft sediments. Further, the return measurement is a complex signal that can be decomposed into spectral moments using a Fast Fourier Transform algorithm. The SBE frequency influences the maximum survey depth, the ability of the signal to penetrate hard substrates, and signal

resolution. Signals at higher frequencies, such as ~200 kHz, attain a maximal water depth of ~100 m, reflect off hard substrates and possess higher signal resolutions, making them ideal for mapping vegetation in shallow water (Preston et al., 2006; Quintino et al., 2009). Conversely, lower frequencies travel through hard substrates and can survey in deeper water. As beamwidth relates to the size of the acoustic footprint, narrower beamwidths are usually preferred as they will have finer spatial resolution. This is especially important when surveying in deep water as the footprint increases in size as sound attenuates through the water column (Galloway & Collins, 1998). Wide beamwidths result in deteriorated signal quality as sound returns at sub-optimal angles for transducer conversion (Snellen et al., 2011). SBEs have been used for seafloor classification (Bates & Whitehead, 2001; Hutin et al., 2005; Parnum et al., 2009; Snellen et al., 2011; Lee & Lin, 2018) and are more affordable than other acoustic remote sensing techniques such as multibeam echosounders (MBEs) (Parnum et al., 2009). The use of MBEs (Wildish et al., 2004) and side-scan acoustic imagery (Hughes et al., 2002) have been successfully used to detect aquaculture-derived organic matter deposition over soft bottom substrates. However, to our knowledge, there has been no assessment of the usefulness of SBE data for detecting flocculent matter deposits beneath aquaculture sites located over hard bottom or complex seafloors. In NL, flocculent matter deposition can change the nature of the substrate from hard to soft (Salvo et al. 2015).

2.1.5 Machine learning and related applications

Machine learning is a subset of artificial intelligence that combines computer science and statistics, giving machines the ability to learn from previous experience by improving upon a task without specified instruction (Bishop, 2006). From its genesis in the 1950s, machine learning and its applications have expanded as computing power became increasingly accessible. Machine learning is increasingly used by oceanographers and marine biologists to characterize seafloor habitats (Parnum et al., 2009; Snellen et al., 2011; Li, 2013; Tulldahl et al., 2013; Lee & Lin, 2018; Montereale-Gavazzi et al., 2018).

Machine learning is divided into supervised and unsupervised learning. In supervised learning, features are tied to a labelled response (James et al., 2013). In this way, supervised algorithms can learn from examples that serve as 'gold standards' of different classes to be predicted (Mohri et al., 2012). After training, models can be applied to data without labels to predict class membership based on decision boundaries derived from fitting labelled data (James et al., 2013). If the response is categorical, the supervised learning task is referred to as a classification whereas predicting continuous response variables is referred to as a regression (James et al., 2013). In unsupervised learning, features are not associated with labels (James et al., 2013). Instead, unsupervised learning detects clusters or underlying structure in data or performs dimensionality reduction as is done in principal component analysis (James et al., 2013).

A supervised learning approach to seafloor classification is attractive because data collection (e.g. collecting hydroacoustic data with SBEs or MBEs) can be far cheaper

than categorizing the small (<0.25 m^2), discrete portions of the seafloor assessed using grab sampling or image capture. By training algorithms on a relatively small set of labelled examples, boundaries that separate classes can be defined and applied over the entire survey area; assessing an equivalent area of the seafloor by grab sampling or drop camera imaging would be prohibitively expensive.

2.1.6 Objective

The objective of this work is to evaluate the ability of SBE data to detect flocculent matter deposited on the seafloor, using drop camera imaging data for groundtruthing. We define a method of best practices for detecting flocculent matter by applying machine learning to hydroacoustic data collected at NL aquaculture sites with steep slopes, predominately hard substrates, and variable depth profiles.

2.2 Materials and methods

2.2.1 Process overview

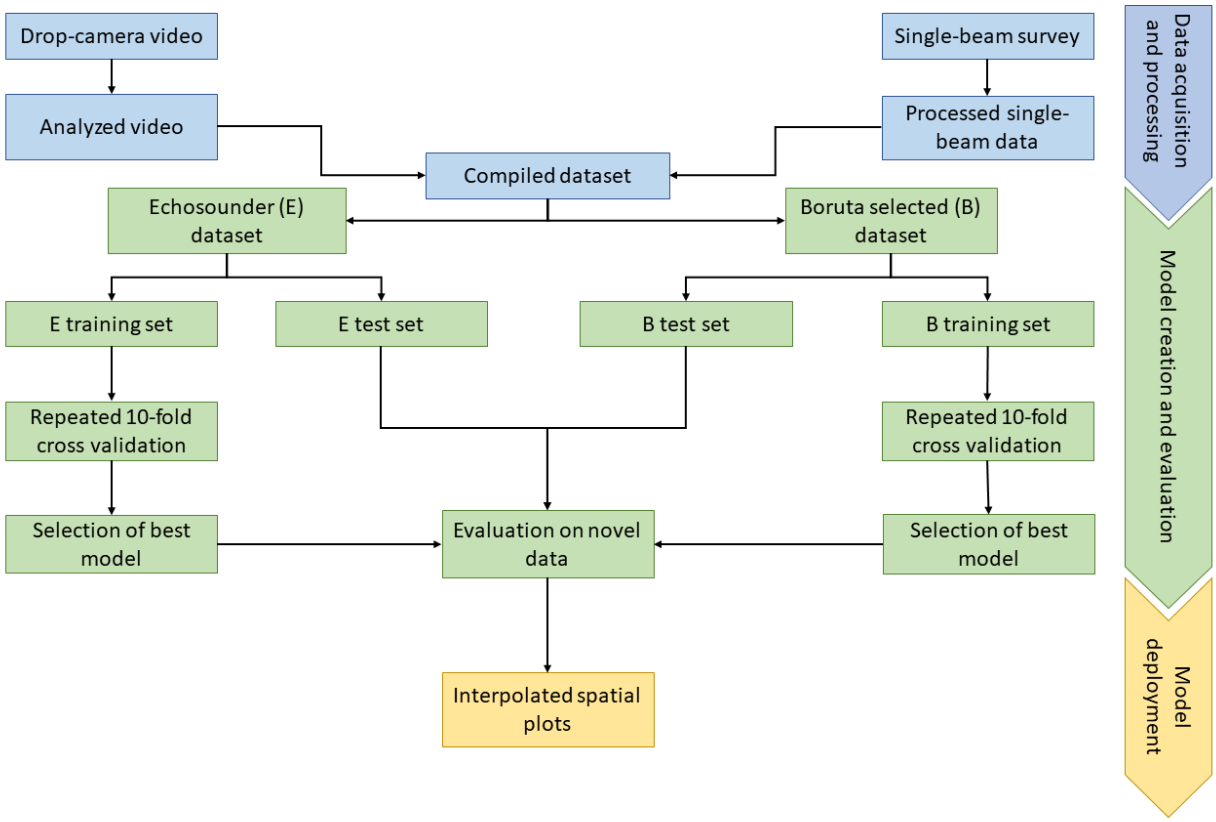

Figure 1. Project workflow illustrating data collection and merging before predictive modelling, evaluation and deployment. Echosounder dataset contains features derived from echosounder post processing. Bortua selected dataset contained echosounder features and their 2nd order interactions passed to the Boruta algorithm which filtered features based on their importance compared to permuted, uninformative features.

2.2.2 Study sites and data collection

We collected data within four coves in the Coast of Bays area of southern NL. At two sites, aquaculture operations had recently ceased (sites A and B) while the other two sites (C and D) were in production. A BioSonics DT-X single beam echosounder (rented from Hoskin Scientific) with a 204.8kHz transducer and a 9° beam width was used to collect hydroacoustic data, which were recorded using BioSonics Visual Acquisition software. The transducer was mounted to the gunnel of a 30 ft vessel with a triangular bracket, and vessel speed was maintained at 4 knots to provide continuous data as recommended by the manufacturer. Due to the high frequency emitted by the echosounder, survey depth was restricted to <100 m. Water temperature and conductivity (to determine salinity) were measured using data from CTDs on collection day to calibrate the speed of sound. For this project, two sites (C, D) were sampled following a grid pattern intersecting much of the cove to capture a maximum of spatial information. SBE data were collected and a primary classification of substrate type based on echosounder features was performed with Visual Habitat software to identify 5 bottom type clusters and create a reference map. Drop camera sampling was performed along transect lines in relation to the reference map and the position of the boat during SBE acquisition to obtain video recordings at five discrete locations corresponding to each of the five bottom type clusters at sites C (N = 24) and D (N = 25). At the two other sites (A, B), SBEs and video were sampled over transects designed for another project. For all sites, the same underwater video system was used with a camera and light mounted on a stainless-steel frame (see Salvo et al., 2015 for details on camera setup). Video was analysed by still

capture with ImageGrab (5.0.6) using the DFO photographic guide for video monitoring to characterize sediment types (Salvo et al., 2018). In total, 59 groundtruthing samples were collected at site A, 38 at site B, 24 at site C and 25 at site D. There were approximately 4 replicates per sample (i.e. images collected in proximity to one another, collected after briefly raising and lowering the camera system at each station along transects) bringing the total number of reference datapoints to 648.

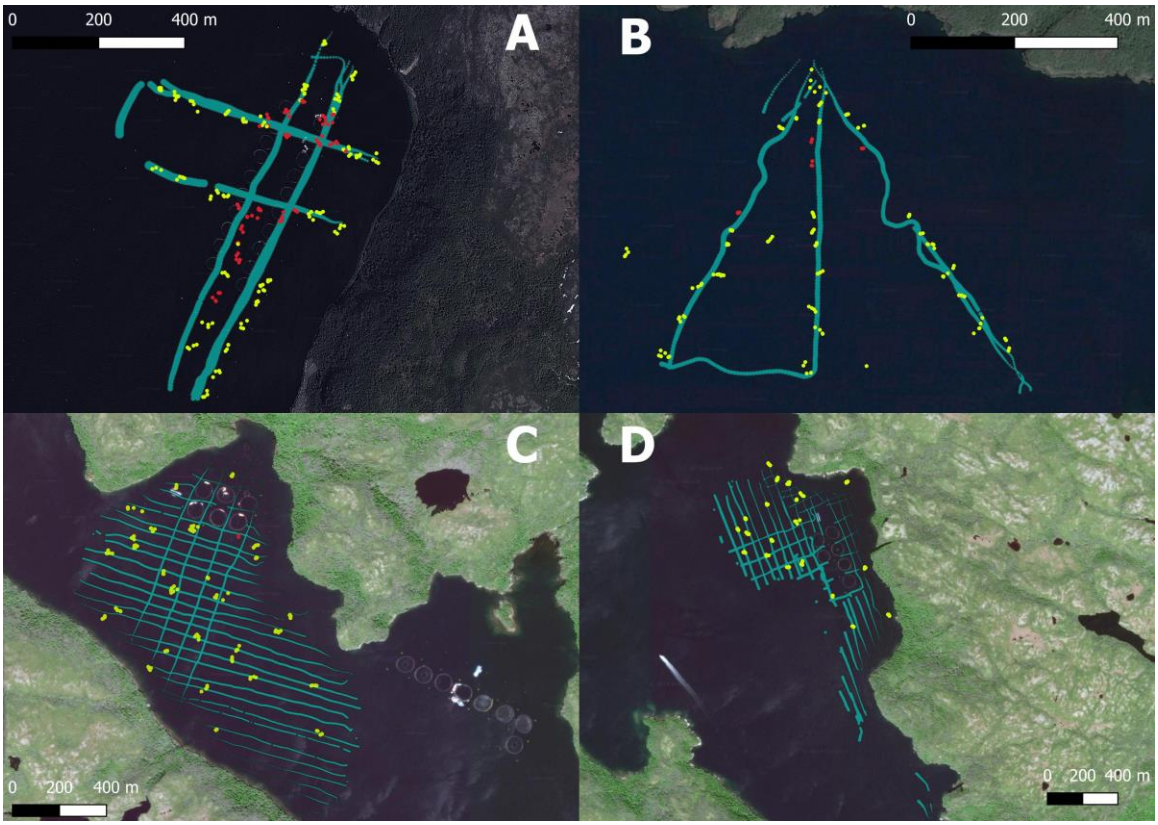

Figure 2. Outline of single beam survey path (teal) for each sample site. Line width scaled to seafloor area surveyed and changes as a function of depth. Red points indicate groundtruthed locations of flocculent matter. Yellow points indicate groundtruthed locations for substrates which are not flocculent matter. Sample location points (red and yellow) are not to scale.

Discrimination between flocculent matter and other soft sediments was done by examining the degree to which the camera system became submerged in sediment, the visual appearance of the sediment, and the presence of offgassing.

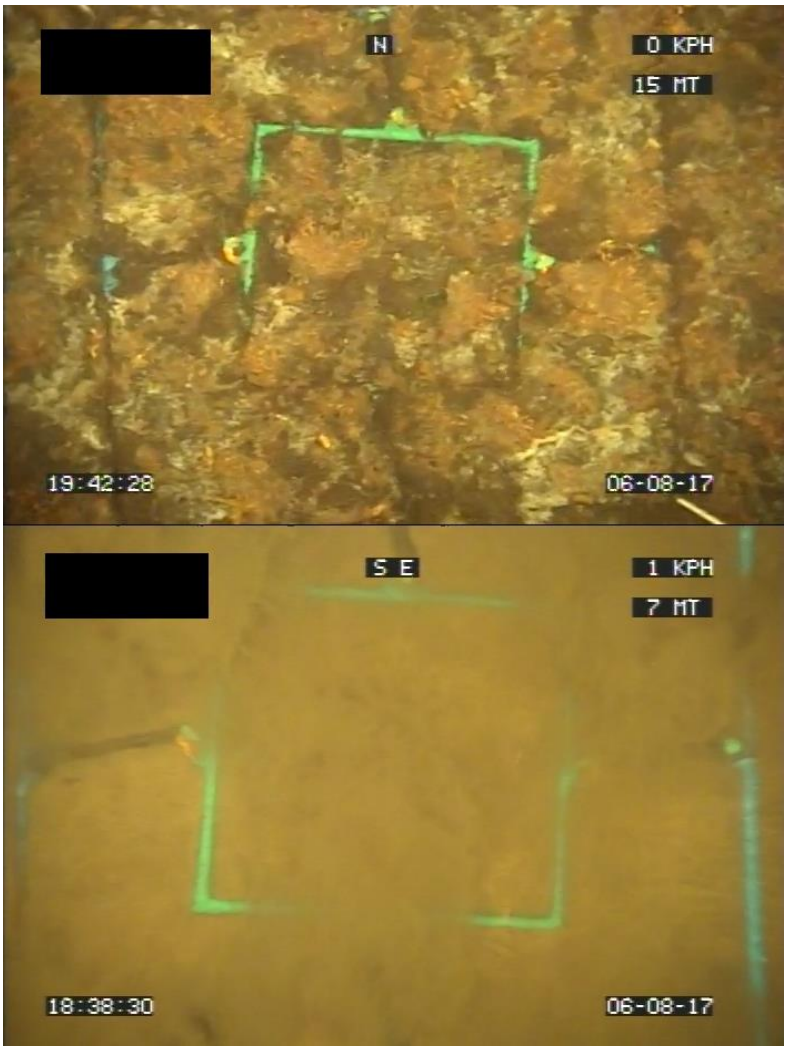

Figure 3. Drop camera images highlighting visual differences between flocculent matter (top) and soft sediments (bottom).

2.2.3 SBE post-processing and intermediate feature extraction

We used Visual Habitat (version 2.0.3.9824, 64-bit) post-processing software integrated with the BioSonics DT-X Echosounder to extract intermediate features. Bottom detection, feature extraction and bottom type algorithms were used to process raw echosounder data. Feature extraction algorithm parameters included rising edge bottom detection with a domain (level of time varied gain) of 30LogR, a rising edge threshold of -30 dB and length criterion set to 10 cm. The search window was 100 cm from the previous rising edge. The first and second bottom echoes were 1- and 3-times the pulse length, respectively. Depth normalization was enabled to normalize the acoustic envelope for depths greater than reference (the latter set as the median depth of each sample site). The fractal dimension threshold was set to -60 dB. For bottom typing analysis, the energy filter width was set to 10 pings and the energy filter threshold at 75%. Intermediate results from these analyses were exported as .csv for use in R (R Core Team, 2018).

2.2.4 Dataset creation, division and preparation

R (version 3.5.1, "Feather Spray", 64-bit) and RStudio (version 1.1.456) were used to wrangle, aggregate and tidy data, and to create predictive models (R Core Team, 2018). Before performing a supervised classification, raw echosounder and groundtruthing data (images) needed to be paired adequately. GPS data from the echosounder and from each benthic image were collated (note: GPS video data corresponds to vessel, not video position). Data were joined using nearest neighbor analysis (k = 1) in QGIS (version

3.4.1, ‘Madeira’, 64-bit), with a maximum distance threshold of 10 m between points selected. Datapoints from the four sites which met this criterion were combined into a merged dataset. Using all 4 sites (A-D) in a single training/validation set instead of one per site was done to ensure there were enough examples of flocculent matter available for algorithms to generate stable decision boundaries. The merged dataset contained n = 266 datapoints, of which 48 (18%) represented locations where flocculent matter was observed, while the remaining 218 (82%) were locations where no flocculent matter was seen. Seafloor classifications were binned into two categories, ‘flocculent’ and ‘not flocculent’. Fifteen predictors (single beam echosounder data) were included as potential features for the training procedure. These fifteen predictors consisted of measurements of fractal dimension, energy of the first and second echo and 12 spectral decompositions via a fast Fourier transform algorithm. A second dataset included 2^{nd} order interactions for the 15 single beam features. These 2^{nd} order interactions were subsequently filtered using the Boruta algorithm, as detailed below. Spatial coordinates were not included as potential features for the training procedure given that instances of flocculent matter were tightly clustered, and the resulting spatial autocorrelation would provide unrealistic estimates of model performance.

Caret’s *createDataPartition* function (Kuhn, 2008) was used to divide samples into training/validation and test sets using stratified random sampling to maintain class balances. A training/validation set is used to fit models to data, with the validation component used to measure the performance of models with varying hyperparameters. After dividing the dataset, the training/validation set was standardized via z-score

transformation setting $\bar{x}_j = 0$ and $s_j = 1$. These parameters were applied to features in the test set, as it is assumed that samples drawn from the same population share an underlying distribution (Friedman et al., 2001).

For the merged dataset containing raw echosounder features, highly correlated features were defined by identifying pairwise correlations ≥ 0.9. In total, 9/15 features had pair-wise correlations approaching 1 and were removed to decrease training time and potentially improve model performance (Li et al., 2011a, b; Chandrashekar & Sahin, 2014). For the dataset containing 2nd order interactions of raw echosounder features, the Boruta algorithm was used for feature selection (Kursa & Rudnicki, 2010). The Boruta algorithm adds shadow attributes not correlated with the response and calculates their maximum Z score (Kursa & Rudnicki, 2010). Boruta assigns features with importance measures significantly better than the maximum shadow attribute Z score as important, and those that do not as unimportant (Kursa & Rudnicki, 2010). From the 435 features created by including all 2nd order interactions and spatial features, 13 were selected.

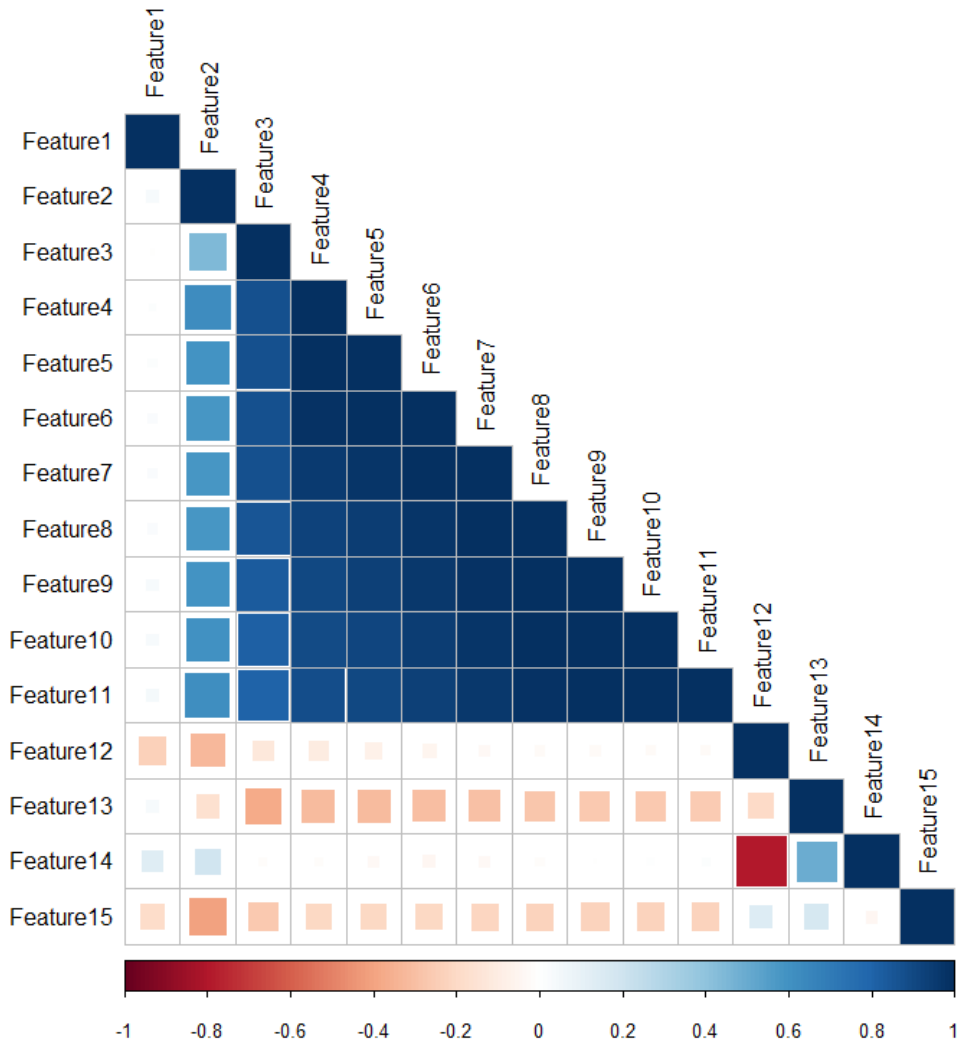

Figure 4. Pearson correlation plot used to filter features in the merged training dataset, with color scale indicating degree of correlation from -1 to +1. Size of squares included for illustrative purposes indicating strength of correlation, with size increasing as correlations approach |1| and decreasing as correlations approach 0.

Class imbalance can negatively affect performance and was accounted for during model creation using the Synthetic Minority Over-sampling Technique (SMOTE) (Chawla et al., 2002). SMOTE oversamples the minority class (flocculent matter) by synthesizing new feature vectors based on the values of k-nearest neighbors while undersampling the majority class (Chawla et al., 2002). The inclusion of SMOTE in model training has resulted in increased performance for other classification problems (Chawla et al., 2002; Karabulut & Ibrikci, 2014; Ren et al., 2015). Models using filtered echosounder and Boruta selected datasets were created with and without SMOTE to compare performance. Oversampling, undersampling and k-nearest neighbors were left at default values of +200%, -200% and 5, respectively. SMOTE was used for each resampling fold through its implementation in the *caret* package.

2.2.5 Algorithm selection

The No-Free Lunch Theorem states that there is no best algorithm for all problems (Wolpert and Macready, 1997). Since it cannot be determined a priori which algorithm will return an optimal fit to data, five were selected based on their distinct methods of generating decision boundaries and their successful use as classifiers.

1) **Random forest** (Breiman, 2001) is a popular algorithm due its easy implementation, robustness to noisy data and its reduction of overfitting common with single decision trees (Breiman, 2001; Friedman et al., 2001). Random forests work by creating n-number of decision trees with a random subset of data and averaging their results in a process known as bagging (Breiman, 2001). By

selecting random features at each split, decision trees are decorrelated from one another (Breiman, 2001). From these random features, the one which improves the purity (Breiman, 2017) or gain (Quinlan, 2014) of the resulting node is selected.

2) **Xgboost** (Chen & Guestrin, 2016) and specifically the xgbTree implementation also uses decision trees as a base learner but uses boosting rather than bagging (Friedman, 2001). Boosting works by creating weakly correlated decision trees and adjusting the weights of observations depending on their correct classification (Friedman et al., 2001). These weights are then used in the next iteration with difficult to classify observations receiving additional attention (Friedman et al., 2001). This process continues for a specified number of rounds, after which all models are combined by weighted majority vote, creating a single strong learner from many weak learners (Friedman et al., 2001).
3) **LogitBoost** is another tree-based algorithm which uses boosting (Friedman et al., 2000) but differs in that it constructs decision stumps (decision trees with a single split) and minimizes log-likelihood loss rather than exponential loss (Friedman et al., 2001). By using the log-likelihood loss function, the LogitBoost algorithm is made more robust to noise and outliers (Friedman et al., 2001; Feng et al., 2005).
4) **Support Vector Machines** (SVMs) are binary classifiers which function by identifying support vectors (observations nearest the dividing hyperplane) that maximize class separation (Cortes & Vapnik, 1995). Originally, SVMs performed classifications where classes were linearly separable, but with kernel

functions SVMs can create non-linear decision boundaries by projecting data into higher dimensional space with the radial basis function kernel being commonly used (Chang et al., 2010; Boughorbel et al., 2017).

5) **k-Nearest Neighbors (k-NN)** is an algorithm which functions by classifying observations based on the memberships of k-number of nearest points in the feature space (Cover & Hart, 1967; Keller et al., 1985). Distances between neighbors can be measured in different ways with Euclidean distance typically used for continuous variables (Keller et al., 1985). k-Nearest Neighbors is one of the simplest machine learning algorithms and does not make assumptions regarding the distribution of data (Keller et al., 1985).

2.2.6 Model creation, validation and hyperparameter tuning

Models were created and evaluated using the *caret* package (Kuhn, 2008) in R. Repeated k-fold cross validation with 10 folds and 100 repeats was used to train and evaluate model performance. k-fold cross validation divides a dataset into k-number of subsamples. Models are trained on k-1 sets, with the last fold withheld to evaluate performance. Each of the folds has a turn being the validation set, after which, the original dataset is resampled, and the process is repeated. Cross validation permits a thorough examination of the hyperparameter space while reducing the risk of model overfitting (Chicco, 2017). One hundred repetitions were used to account for covariate drift during stratified random sampling and to ensure performance estimates were stable (Moreno-Torres et al., 2012; López et al., 2014; Li et al., 2016). Seeds were set to ensure the same training/validation folds were used for all candidate models.

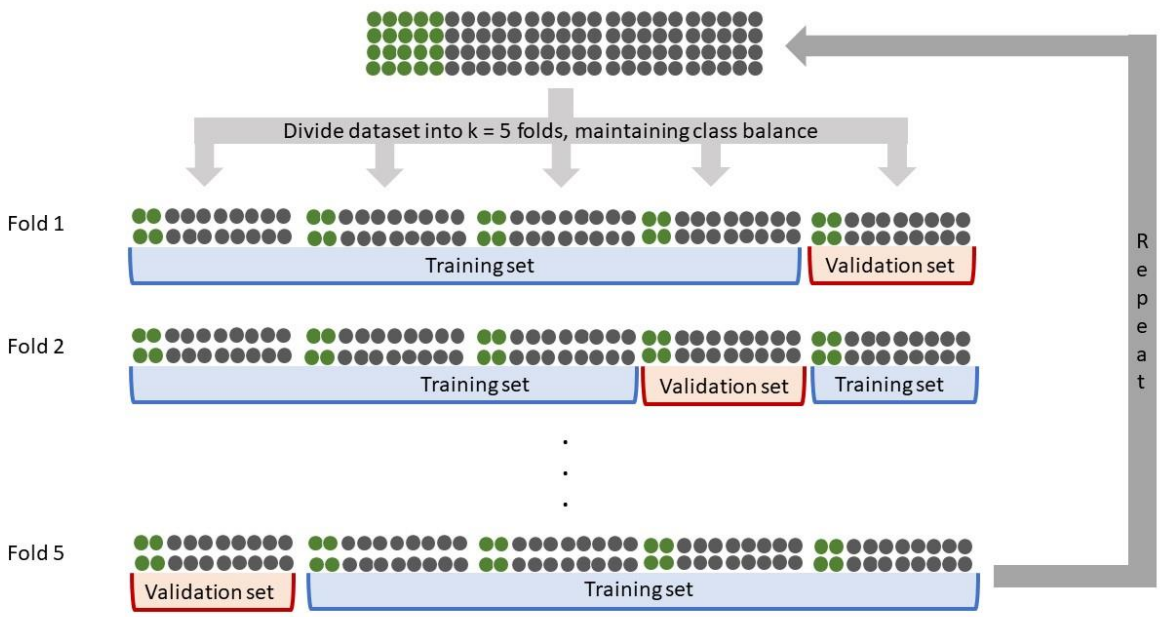

Figure 5. Example of 5-fold repeated cross validation. Green dots represent cases of flocculent matter. Grey dots represent substrates which are not flocculent matter.

Hyperparameters are parameters set before model training which cannot be generalized from data. Hyperparameters significantly effect algorithm behavior and must be tuned to create optimally performing models. Grid search was used for algorithms with a single hyperparameter. For algorithms with 2 or more hyperparameters, a random search with 60 attempted tuples was used to find optimal values. **Random search** has been demonstrated to be more efficient than grid or manual search, finding comparable or better solutions while reducing computation time (Bergstra & Bengio, 2012). The random search length of 60 tuples provides 95% certainty that one of the tuples lies within the top 5% of optimal solutions. This is shown by the following equation where p is the probability of our result being in a certain q or quantile:

$$1 - q^n \geq p \Rightarrow n \geq \frac{\log(1-p)}{\log(q)}$$

Accuracy is a common metric of model performance but should not be used when classes are imbalanced or when the cost of type I vs type II error differ (Chawla et al., 2002). Instead, curve of the **receiver operator characteristic** (ROC) is used as a performance metric because it is insensitive to class imbalance, weighs the costs between sensitivity and specificity (Fawcett, 2006) and is implemented as an objective function in the *caret* package. A ROC score of 1 represents 100% sensitivity and specificity while 0.5 indicates performance no better than chance (Fawcett, 2006). **Matthew's correlation coefficient (MCC**) is used to evaluate model performance on novel holdout data (Matthews, 1975). MCC is a desirable metric in this context because it accounts for sensitivity and specificity in model evaluation, is similarly insensitive to class imbalances

(Boughorbel et al., 2017) and is easier to interpret with those familiar with correlation coefficients. Values of MCC range between -1 and 1; a coefficient of 1 represents complete agreement between predicted and observed cases, 0 represents agreement no better than chance and -1 a perfectly incorrect prediction (Boughorbel et al., 2017).

Twenty models were created using treatments of the training set for each of the five algorithms. These treatments included:

- echosounder features only (**E**),
- echosounder features only with SMOTE **(ES)**,
- Boruta selected features (**B**),
- Boruta selected features with SMOTE **(ES)**.

2.2.7 Statistical measures

Confusion matrix

		Observed	
		Floc	Not
Predicted	Floc	TP	FP
	Not	FN	TN
		P	N

True positive	TP
True negative	TN
False positive (Type I error)	FP
False negative (Type II error)	FN
Sensitivity	$\frac{TP}{P}$
Specificity	$\frac{TN}{N}$
Matthews Correlation Coefficient (MCC)	$\frac{TP \times TN - FP \times FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}}$

2.2.8 Model summary

Table 1. Summary of algorithms, training sets (sb: single beam features only; fe: feature engineered single beam features including 2nd order interactions selected by Boruta importance score), hyperparameter search methods and data treatments (CS: centered and scaled using z-score transformation; HCF: Removal of highly correlated features; SMOTE: Synthetic minority oversampling technique; Boruta: Boruta variable importance wrapper for feature selection) used in the creation of predictive models.

Model	Algorithm	Training set	Hyperparameter search	Treatment
lb.1	LogitBoost	sb	grid	CS, HCF
lb.2	LogitBoost	sb	grid	CS, HCF, SMOTE
lb.3	LogitBoost	fe	grid	CS, Boruta
lb.4	LogitBoost	fe	grid	CS, Boruta, SMOTE
xgb.1	xgbTree	sb	random	CS, HCF
xgb.2	xgbTree	sb	random	CS, HCF, SMOTE
xgb.3	xgbTree	fe	random	CS, Boruta
xgb.4	xgbTree	fe	random	CS, Boruta, SMOTE
rf.1	ranger	sb	random	CS, HCF
rf.2	ranger	sb	random	CS, HCF, SMOTE
rf.3	ranger	fe	random	CS, Boruta
rf.4	ranger	fe	random	CS, Boruta, SMOTE
knn.1	knn	sb	grid	CS, HCF
knn.2	knn	sb	grid	CS, HCF, SMOTE
knn.3	knn	fe	grid	CS, Boruta
knn.4	knn	fe	grid	CS, Boruta, SMOTE
svm.1	svmRadial	sb	random	CS, HCF
svm.2	svmRadial	sb	random	CS, HCF, SMOTE
svm.3	svmRadial	fe	random	CS, Boruta
svm.4	svmRadial	fe	random	CS, Boruta, SMOTE

2.2.9 Flocculent matter presence probability mapping and interpolation

Trained and evaluated models were used to predict probabilities of flocculent matter presence for single beam echosounder points which exceeded the 10 m threshold required to be merged with groundtruthing video. Predicted probabilities and the spatial coordinates of single beam points were merged and exported as a .csv. QGIS (v 3.6.3 'Noosa' with GRASS 7.6.1) was used to create flocculent matter probability maps with TIN interpolation used to create a continuous surface between data points. TIN

interpolation works by triangulating a set of observations with edge connections that form an interpolation surface (Mitas & Mitasova, 1999). TIN interpolation was selected because it does not interpolate beyond data extents and performs well with variable distance points (Yang et al., 2004). Areas with large numbers of points can be rendered in higher resolution and in lower resolution where observations are sparse. Due to variable distances between single beam sampling transects, TIN pixel size was set to 10 m^2 to maintain projection quality and minimize the introduction of interpolation artifacts.

2.2.10 Statistical analysis of incorrect classifications

Correct and incorrect predictions from model evaluation on holdout test data were treated as binary values in a logistic regression to determine contributions of specific single beam features to misclassifications of flocculent matter presence. A type III ANOVA was run using the *Anova* R package to account for the sequential nature of logistic regression and provide a more robust assessment of features associated with classification errors.

2.3 Results

2.3.1 Model performance

The best performing model for each treatment is examined against withheld test data (N = 66) to evaluate their performance on unseen data. Like training data, test data stems from all four sites. Here we present only the best output of each model type.

Single beam only models (E, ES). Echosounder only models reported sensitivities ranging from 0 (no detection of flocculent matter) to 0.6667 (1 corresponding to the correct classification of all examples of flocculent matter). All models were significantly improved by SMOTE, with the best model (knn.2) having an MCC of 0.482 and detecting 2/3 cases of flocculent matter in the test set. SMOTE resulted in increased sensitivity at the expense of specificity for all models.

Table 2. Confusion matrix between predicted and observed values of knn.2 on novel holdout data examining only single beam features (p = 9). 'Floc' is known cases of flocculent matter determined by image capture. 'Other' include all substrates where flocculent matter was not detected during image analysis.

		Observed	
		Floc	Other
Predicted	Floc	8	12
	Other	4	42
	Total	12	54
	Sensitivity	0.6667	---
	Specificity	---	0.7778

Boruta selected models using 2nd order interactions (B, BS). 2nd order interactions selected by the Boruta algorithm were examined to determine if basic feature engineering and selection improved performance. Compared to echosounder only models, feature engineering and selection resulted in models with poorer performance and longer computation times. The best performing Boruta selected model (rf.3) recorded a sensitivity, specificity and MCC of 0.16667, 0.96296 and 0.121 respectively.

Table 3. Confusion matrix between predicted and observed values of model rf.3 on novel holdout data with Boruta selected spatial and 2nd order interaction echosounder features (p = 13). 'Floc' is known cases of flocculent matter determined by image capture. 'Other' include all substrates where flocculent matter was not detected during image analysis.

		Observed	
		Floc	Other
Predicted	Floc	2	2
	Other	10	52
	Total	12	54
	Sensitivity	0.16667	---
	Specificity	---	0.96296

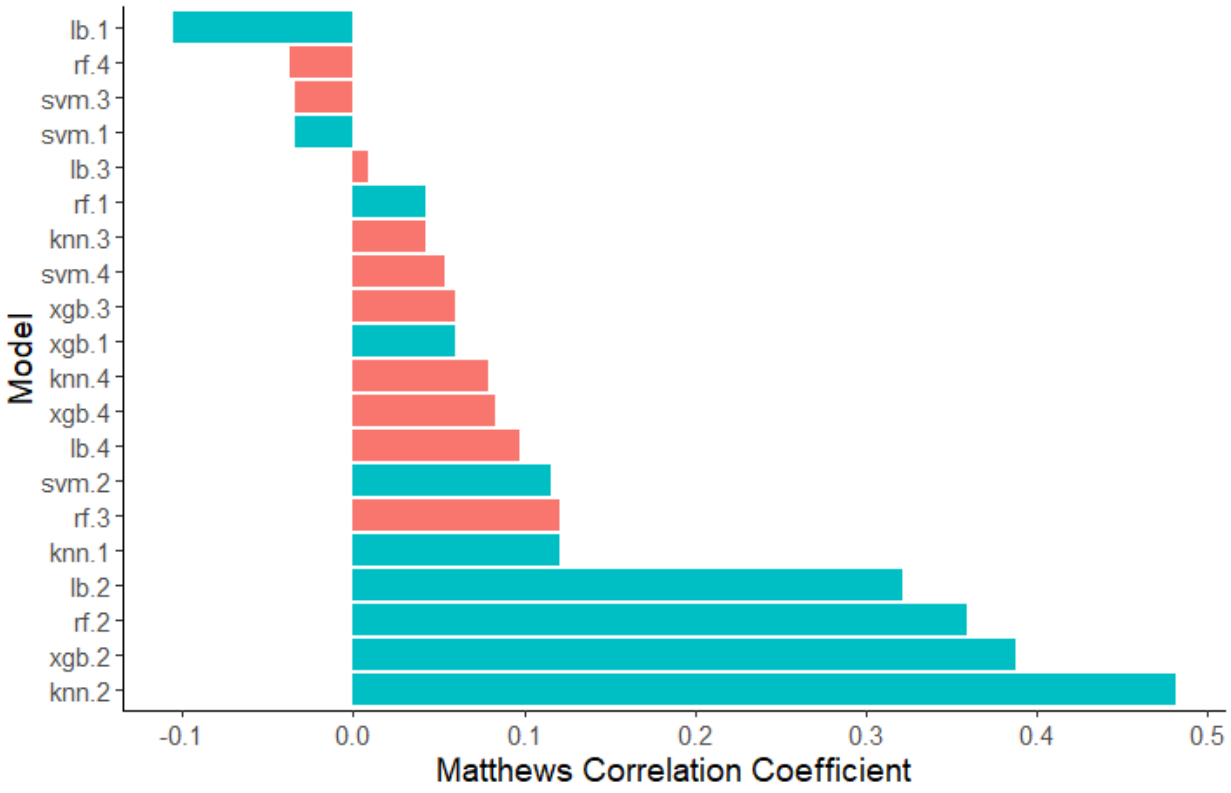

Figure 6. Summary of model performance on holdout data using different feature combinations (blue: single beam only features; red: 2nd order interactions filtered using the Boruta algorithm). Matthews correlation coefficient: 0 represents predictions no better than chance, 1 represents perfect agreement between predictions and known cases.

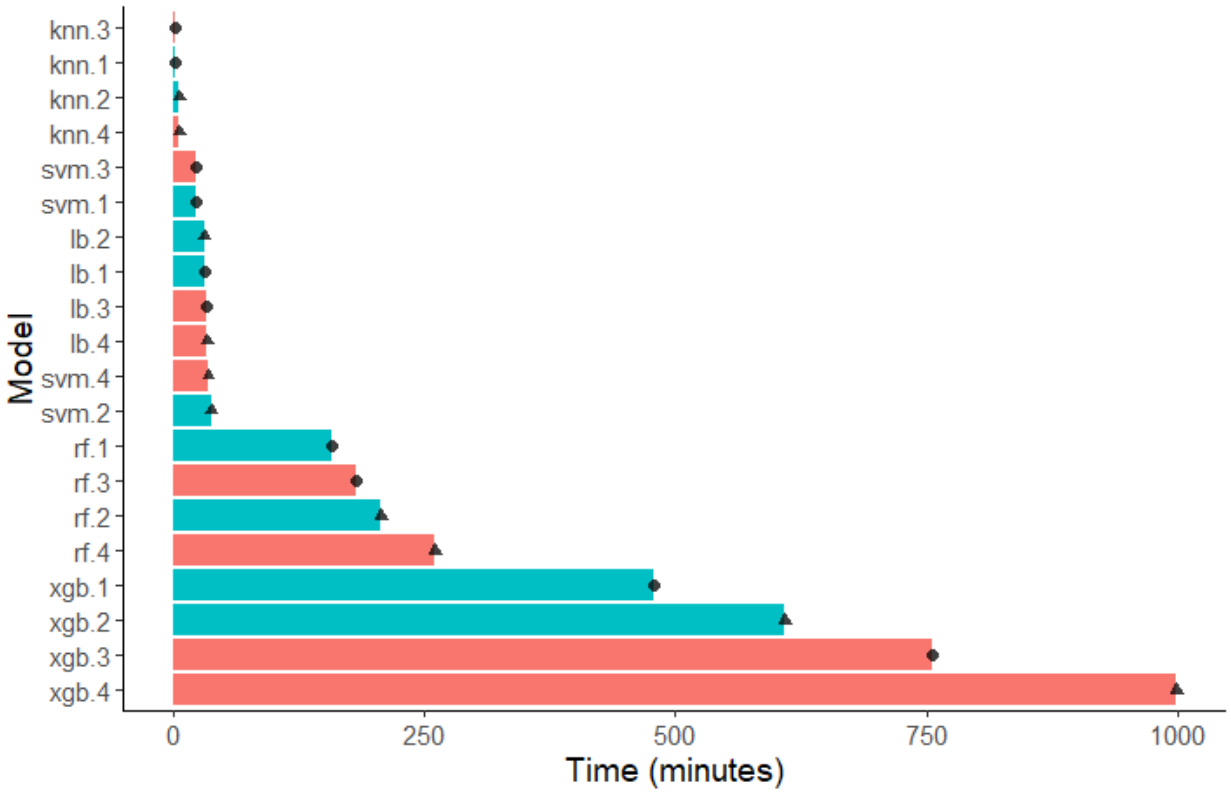

Figure 7. Computation time for model training (SMOTE: triangle = yes, circle = no; blue: single beam only, red: 2nd order interactions filtered using the Boruta algorithm)

2.3.2 Spatial interpolations

For the following spatial interpolations, site D had no flocculent matter detected as part of the groundtruthing survey. A, B and C had 31, 16 and 1 instances of flocculent matter presence based on groundtruthing.

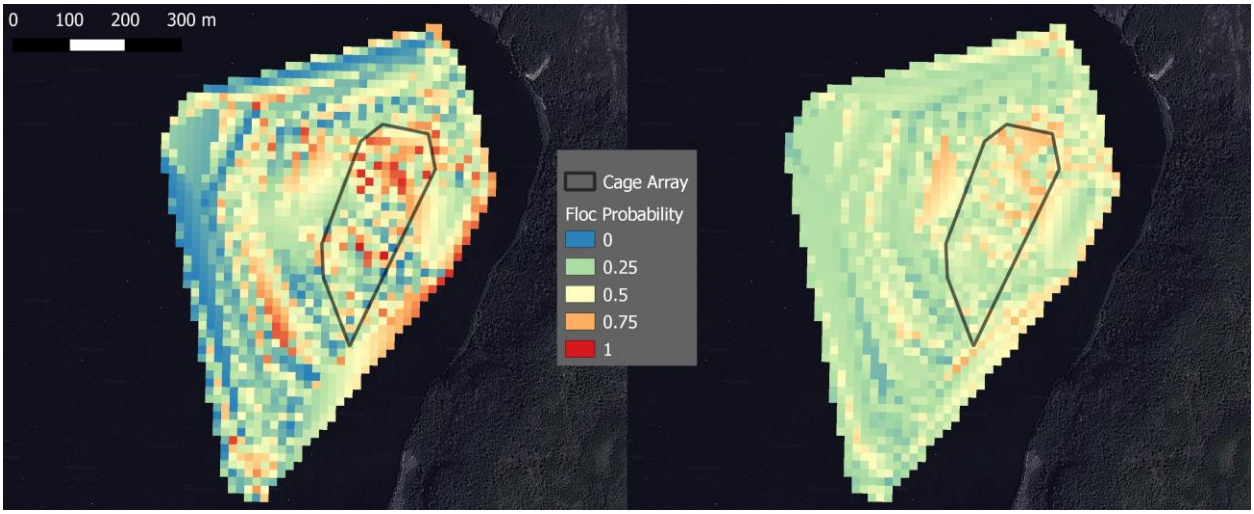

Figure 8. TIN interpolation of predicted probability of flocculent matter presence at Site A. Each map represents the best performing model for each feature set. Left: SMOTE k-nearest neighbor with single beam features only. Right: SMOTE random forest with Boruta selected features.

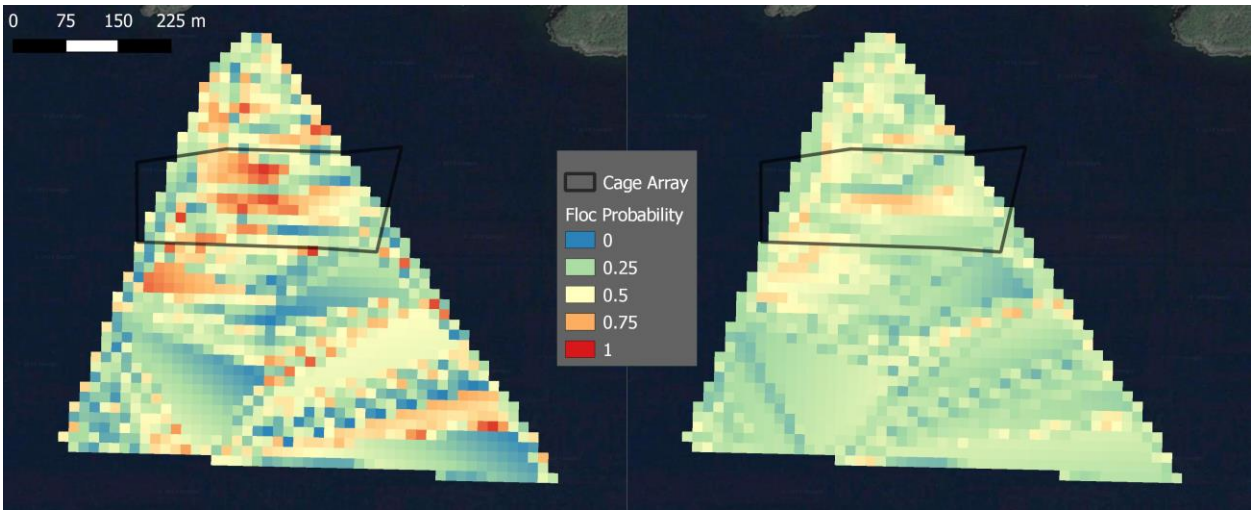

Figure 9. TIN interpolation of predicted probability of flocculent matter presence at site B. Each map represents the best performing model for each feature set. Left: SMOTE k-nearest neighbor with single beam features only. Right: SMOTE random forest with Boruta selected features.

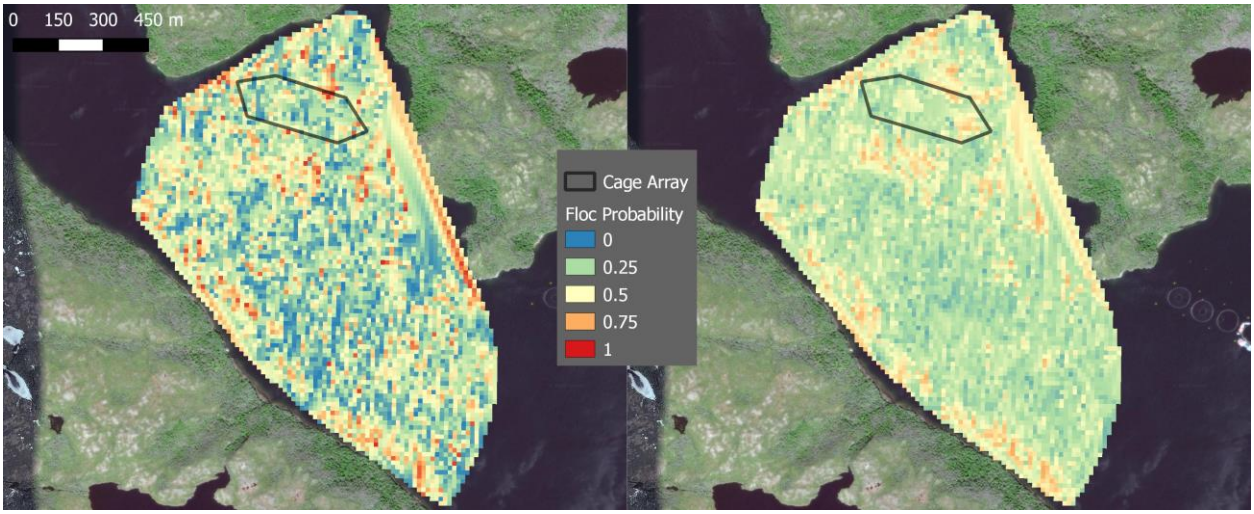

Figure 10. TIN interpolation of predicted probability of flocculent matter presence at Site C. Each map represents the best performing model for each feature set. Left: SMOTE k-nearest neighbor with single beam features only. Right: SMOTE random forest with Boruta selected features.

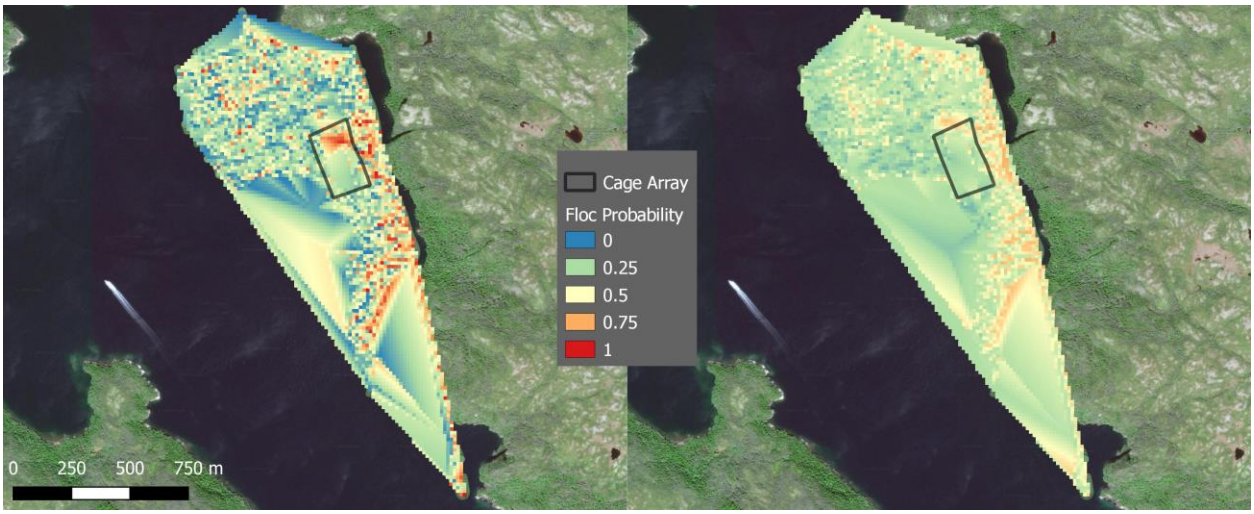

Figure 11. TIN interpolation of predicted probability of flocculent matter presence at Site D. Each map represents the best performing model for each feature set. Left: SMOTE k-nearest neighbor with single beam features only. Right: SMOTE random forest with Boruta selected features.

2.3.3 Relationships between incorrect predictions for single beam only models

Correct and incorrect predictions for model knn.2 (the best performing model using only single beam features) were treated as a binary response in a logistic regression to determine if differences between these groups could be attributed to features used in model creation. The results of this regression demonstrate that Feature 1 ($Pr(>|z|) = 0.0198$) was associated with significant decreases in correct classifications. No other features were found to be significant. A type III ANOVA was performed to account for the sequential nature of logistic regression and similarly concluded that Feature 1 ($Pr(>Chisq) = 0.0002$) significantly affected classifications. This result is expected as increased depth (especially as depth approaches the maximum echosounder limit) is associated with degraded signal quality. Feature 1 is a measure of fractal dimension of the first bottom echo and is therefore signal quality dependent.

2.4 Discussion

Based on the methods and equipment outlined in this paper, SBEs shows a limited ability to detect and discriminate flocculent matter from other substrates on Newfoundland's southern coast. The best performing model examining only single beam features detected flocculent matter in 67% of cases and correctly discriminated flocculent matter from other substrates in 78% of cases presented in novel holdout data, resulting in a moderate MCC of 0.482. This suggests that detection of flocculent matter using hydroacoustic data is possible but not sufficiently reliable to deploy as a regulatory or assessment tool with the equipment used in this study. Encouragingly, there is agreement between predictions and interpolated spatial plots, with a high probability of flocculent matter presence in top right of the pre-existing cage array at Site A and beneath the center of the cage array at Site B (Figure 8, Figure 9). This agreement

suggests that SBE data possesses a signal that could be exploited with equipment better suited for the difficult conditions presented at Newfoundland aquaculture sites.

Incorrect classifications are at least partly attributable to predictive models fitting data collected at the extremes of what can be recorded by the echosounder (100 m depth). Therefore, using a lower frequency echosounder with a greater maximum depth should improve model performance. Despite the maximum recording depth of the SBE used in this project being 100 m, we recommend that echosounder systems not exceed ½ the indicated maximum depth when collecting hydroacoustic data. Echosounder systems often use information from repeated bottom echoes to classify seafloor types and this information cannot be retrieved with depths exceeding 50 m (Bates et al., 2001; Foster et al., 2009). There are sources of error associated with the collection of hydroacoustic data which likely affected model performance. Acoustic returns can be affected by the presence of sessile benthic organisms, which reduce backscatter intensity and were observed during video analysis (Li et al., 2016). Alternatively, the use of image capture to classify substrates can be made unreliable if fine grained sediments obscure visual detection of harder substrates beneath (Siwabessy et al., 2013) which was likely with Newfoundland's complex and heterogenous seafloors. Acoustic returns between fine grained sediment and flocculent matter may possess similar properties. Due to restrictions in study design, we were unable to re-examine areas predicted to be flocculent matter or whether these areas were composed of fine sediments rather than aquaculture-linked deposits. The collection of additional groundtruthing to confirm predictions at these locations could either validate, or help refine, the modelling process.

The application of SMOTE resulted in overall increases in sensitivity and decreased specificity. This behavior has been observed in other cases where SMOTE has been used and is attributed to

the method in which SMOTE synthesizes observations. SMOTE uses k-nearest neighbors of the minority class to create synthetic observations but does not consider the positions of nearby majority examples (Stefanowksi & Wilk, 2008). Despite SMOTE undersampling the majority class (removal of majority class observations) to dampen this effect, overlap occurs because it does not consider the boundaries between majority and minority classes (Wang & Japkowicz, 2004; Stefanowksi & Wilk, 2008). Recent extensions of SMOTE which consider these boundaries or filter noisy data may help minimize decreases in model specificity (Han et al., 2005; Sáez et al., 2015). Another possibility is that the boundaries between classes in our example are obscured by poor readings collected near or at the maximal survey depth of the echosounder.

A future project examining hydroacoustic data as a method of detecting flocculent matter should consider the regular application of sampling transects providing adequate and regular seafloor coverage (i.e, more akin to the sampling done at sites C and D than at sites A and B). This would improve the quality of spatial interpolations and the likelihood that groundtruthing samples overlap with the single beam track. Additionally, the use of a lower frequency, narrower beamwidth single beam echosounder or the use of a multibeam echosounder will result in improvements in the signal-to-noise ratio and model performance. Large areas of surveyed sites neared or exceeded the operational depth of the echosounder used in this project, resulting in deteriorated signal quality. Additional groundtruthing should be conducted with careful consideration of the path taken by the survey vessel to ensure the maximal number of groundtruthing points are available as examples during analysis. Ideally, a site where cages have been removed could be observed with an AUV following a defined path to ensure all images captured of the seafloor were usable, thereby decreasing the error associated with combining

observations based on a 10 m distance threshold. By collecting a larger number of examples tied to echosounder values, a model's ability to generalize to locations where groundtruthing is not conducted should improve.

2.5 References

Anderson, M. R., Tlusty, M. F., & Pepper, V. A. (2005). Organic enrichment at cold water aquaculture sites—the case of coastal Newfoundland. In B. Hargrave (Ed.), *Environmental Effects of Marine Finfish Aquaculture* (pp. 99-113). Springer, Berlin.

Asche F, Roll KH, & Tveterås S (2008) Future trends in aquaculture: productivity growth and increased production. In M Holmer et al. (Eds), *Aquaculture in the Ecosystem*. Springer, Dordrecht

Bates, C. R., Whitehead, E. J., & Castle, B. (2001). Echo Plus measurements in Hopavagen Bay, Norway. *Sea Technology*, 42, 34-43.

Bergstra, J., & Bengio, Y. (2012). Random search for hyper-parameter optimization. *Journal of Machine Learning Research*, 13, 281-305.

Bishop, C. M. (2006). *Pattern Recognition and Machine Learning*. Springer, Berlin.

Borja, Á., Rodríguez, J.G., Black, K., Bodoy, A., Emblow, C., Fernandes, T.F., Forte, J., Karakassis, I., Muxika, I., Nickell, T.D., Papageorgiou, N., Pranovi, F., Sevastou, K., Tomassetti, P., & D. Angel. 2009. Assessing the suitability of a range of benthic indices in the evaluation of environmental impact of fin and shellfish aquaculture located in sites across Europe. *Aquaculture*, 293. 231-240.

Breiman, L. (2001). Random forests. *Machine learning*, 45, 5-32.

Breiman, L. (2017). *Classification and regression trees*. Routledge.

Boughorbel, S., Jarray, F., & El-Anbari, M. (2017). Optimal classifier for imbalanced data using Matthews Correlation Coefficient metric. *PLOS ONE*, 12, e0177678.

Broch O.J., Daae R.L., Ellingsen I.H., Nepstad, R., Bendiksen, E.Å., Reed, J.L., & Senneset, G. (2017) Spatiotemporal dispersal and deposition of fish farm wastes: a model study from central Norway. *Frontiers in Marine Science*, 4, 199

Carvalho, S., M. Barata, F. Pereira, M. B. Gaspar, L. Cancela da Fonseca, & P. Pousao-Ferreira. (2006). Distribution patterns of macrobenthic species in relation to organic enrichment within aquaculture earthen ponds. *Marine Pollution Bulletin*, 52. 1573–1584

Chandrashekar, G., & Sahin, F. (2014). A survey on feature selection methods. *Computers & Electrical Engineering*, 40, 16-28.

Chang, Y. W., Hsieh, C. J., Chang, K. W., Ringgaard, M., & Lin, C. J. (2010). Training and testing low-degree polynomial data mappings via linear SVM. *Journal of Machine Learning Research*, 11, 1471-1490.

Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). SMOTE: synthetic minority over-sampling technique. *Journal of Artificial Intelligence Research*, 16, 321-357.

Chen, T., & Guestrin, C. (2016). Xgboost: A scalable tree boosting system. In Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining (pp. 785-794). ACM.

Chopin, T. (2015). Marine Aquaculture in Canada: Well-Established Monocultures of Finfish and Shellfish and an Emerging Integrated Multi-Trophic Aquaculture (IMTA) Approach Including Seaweeds, Other Invertebrates, and Microbial Communities. *Fisheries*, 40, 28-31.

Chicco, D. (2017). Ten quick tips for machine learning in computational biology. *BioData Mining,* 10, 35.

Crawford, C. M., Mitchell, I. M., & Macleod, C. K. A. (2001). Video assessment of environmental impacts of salmon farms. *ICES Journal of Marine Science*, 58, 445-452.

Cortes, C., & Vapnik, V. (1995). Support-vector networks. Machine learning, 20, 273-297.

Costa-Pierce, B. A. (2010). Sustainable ecological aquaculture systems: the need for a new social contract for aquaculture development. *Marine Technology Society Journal*, 44, 88-112.

Cover, T., & Hart, P. (1967). Nearest neighbor pattern classification. *IEEE transactions on Information Theory,* 13, 21-27.

Donnet, S., Ratsimandresy, A.W., Goulet, P., Doody, C., Burke, S., & Cross, S. (2018) Coast of Bays Metrics: Geography Hydrology and Physical Oceanography of Aquaculture Area of the South Coast of Newfoundland. DFO. *Canadian Science Advisory Secretariat Research Document* 2017/076. x+109 p.

Fawcett, T. (2006). An introduction to ROC analysis. *Pattern Recognition Letters*, 27, 861-874.

Feng, K. Y., Cai, Y. D., & Chou, K. C. (2005). Boosting classifier for predicting protein domain structural class. *Biochemical and Biophysical Research Communications,* 334, 213-217.

Fisheries and Oceans Canada (DFO) (2016) *Statistics on Aquaculture production quantities and values, 2000-2017*. Retrieved from www.dfo-mpo.gc.ca/stats/aqua/aqua-prod-eng.htm

Fisheries and Oceans Canada (DFO) (2018*). Aquaculture Activities Regulation*. Retrieved from http://www.dfo-mpo.gc.ca/aquaculture/management-gestion/aar-raa- eng.htm

Fodelianakis, S., Papageorgiou, N., Karakassis, I., & Ladoukakis, E. D. (2015). Community structure changes in sediment bacterial communities along an organic enrichment gradient associated with fish farming. *Annals of Microbiology*, 65, 331-338.

Foster, G., Walker, B. K., & Riegl, B. M. (2009). Interpretation of single beam acoustic backscatter using lidar-derived topographic complexity and benthic habitat classifications in a coral reef environment. *Journal of Coastal Research*, 16-26.

Food and Agriculture Organization of the United Nations (FAO). (2016). *The State of World Fisheries and Aquaculture*. Retrieved from http://www.fao.org/3/a i5555e.pdf.

Friedman, J. H. (2001). Greedy function approximation: a gradient boosting machine. *Annals of Statistics*, 1189-1232.

Friedman, J., Hastie, T., & Tibshirani, R. (2000). Additive logistic regression: a statistical view of boosting (with discussion and a rejoinder by the authors). *The Annals of Statistics*, 28, 337-407.

Friedman, J., Hastie, T., & Tibshirani, R. (2001). *The Elements of Statistical Learning* (Vol. 1, No. 10). New York, NY, USA: Springer Series in Statistics.

Galloway, J. L., & Collins, W. T. (1998). Dual frequency acoustic classification of seafloor habitat using the QTC view. In *OCEANS'98 Conference Proceedings* (Vol. 3, pp. 1296-1300). IEEE.

Government of Newfoundland and Labrador (2016). Economic Impacts of the Newfoundland and Labrador Aquaculture Industry. Retrieved from http://www.fishaq.gov.nl.ca/publications/pdf/Aquaculture_Macro_FINAL.pdf.

Government of Newfoundland and Labrador (2017). *The Economic Review 2017.* Retrieved from https://www.economics.gov.nl.ca/pdf2017/theeconomicreview2017.pdf

Hamoutene D (2014) Sediment sulphides and redox potential associated with spatial coverage of *Beggiatoa* spp. at finfish aquaculture sites in Newfoundland, Canada. *ICES Journal of Marine Science*, 71, 1153–1157

Hamoutene, D., Salvo, F., Bungay, T., Mabrouk, G., Couturier, C., Ratsimandresy, A., & Dufour, S. C. (2015). Assessment of finfish aquaculture effect on Newfoundland epibenthic communities through video monitoring. *North American Journal of Aquaculture*, 77, 117-127.

Han, H., Wang, W. Y., & Mao, B. H. (2005). Borderline-SMOTE: a new over-sampling method in imbalanced data sets learning. In H. De-Shuang et al (Eds.*), International Conference on Intelligent Computing* (pp. 878-887). Springer, Berlin.

Hargrave, B. T., Holmer, M., & Newcombe, C. P. (2008). Towards a classification of organic enrichment in marine sediments based on biogeochemical indicators. *Marine Pollution Bulletin*, 56, 810-824.

Holmer, M., Wildish, D., & Hargrave, B. (2005). Organic enrichment from marine finfish aquaculture and effects on sediment biogeochemical processes. In B. Kronvang et al. (Eds), *Environmental effects of marine finfish aquaculture* (pp. 181-206). Springer, Berlin.

Hughes Clark, J. E., Wildish, D., & Duxfield, A. (2002). Acoustic imaging of salmonid mariculture sites. CHC 2002 Proceedings.

Hutin, E., Simard, Y., & Archambault, P. (2005). Acoustic detection of a scallop bed from a single beam echosounder in the St. Lawrence. *ICES Journal of Marine Science*, 62, 966-983.

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). *An Introduction to Statistical Learning (Vol. 112).* New York: Springer.

Jusup, M., Klanjšček, J., Petricioli, D., & Legović, T. (2009). Predicting aquaculture-derived benthic organic enrichment: model validation. *Ecological Modelling,* 220, 2407-2414.

Karabulut, E. M., & Ibrikci, T. (2014). Effective automated prediction of vertebral column pathologies based on logistic model tree with SMOTE preprocessing. *Journal of Medical Systems*, 38, 50.

Keller, J. M., Gray, M. R., & Givens, J. A. (1985). A fuzzy k-nearest neighbor algorithm. *IEEE Transactions on Systems, Man, and Cybernetics*, 580-585.

Komatsu, T., Igararashi, C., Tatsukawa, K. I., Nakaoka, M., Hiraishi, T., & Taira, A. (2002). Mapping of seagrass and seaweed beds using hydro-acoustic methods. *Fisheries Science,* 68, 580-583.

Kuhn, M. (2008). Caret package. *Journal of Statistical Software*, 28, 1-26.

Kursa, M. B., & Rudnicki, W. R. (2010). Feature selection with the Boruta package. *Journal of Statistical Software*, 36, 1-13.

Lee, W. S., & Lin, C. Y. (2018). Mapping of tropical marine benthic habitat: Hydroacoustic classification of coral reefs environment using single beam (RoxAnn™) system. *Continental Shelf Research,* 170, 1-10.

Li, J. (2013, December). Predicting the spatial distribution of seabed gravel content using random forest, spatial interpolation methods and their hybrid methods. In *The International Congress on Modelling and Simulation* (MODSIM) (pp. 1-6).

Li, J., Heap, A. D., Potter, A., & Daniell, J. J. (2011). Application of machine learning methods to spatial interpolation of environmental variables. *Environmental Modelling & Software*, 26, 1647-1659.

Li, J., Heap, A., Potter, A., & Daniell, J. J. (2011). Predicting Seabed Mud Content across the Australian Margin II: Performance of Machine Learning Methods and Their Combination with Ordinary Kriging and Inverse Distance Squared. *Geoscience Australia*, Record, 7.

Li, J., Tran, M., & Siwabessy, J. (2016). Selecting optimal random forest predictive models: A case study on predicting the spatial distribution of seabed hardness. *PLOS ONE*, 11, e0149089.

López, V., Fernández, A., & Herrera, F. (2014). On the importance of the validation technique for classification with imbalanced datasets: Addressing covariate shift when data is skewed. *Information Sciences*, 257, 1-13.

Mabrouk G, Bungay T, Drover D et al (2014) Use of remote video survey methodology in monitoring benthic impacts from finfish aquaculture on the south coast of Newfoundland (Canada). DFO Canadian Science Advisory Secretariat Document 39

Matthews, B. W. (1975). Comparison of the predicted and observed secondary structure of T4 phage lysozyme. *Biochimica et Biophysica Acta (BBA)-Protein Structure*, 405, 442-451.

Mazzola, A., Mirto, S., La Rosa, T., Fabiano, M., & Danovaro, R. (2000). Fish-farming effects on benthic community structure in coastal sediments: analysis of meiofaunal recovery. *ICES Journal of Marine Science*, 57, 1454-1461.

Mitas, L., & Mitasova, H. (1999). Spatial interpolation. In P. Longley et al. (Eds.), *Geographical Information Systems: Principles, Techniques, Management and Applications* (pp. 481-492). Wiley, Hoboken.

Moreno-Torres, J. G., Sáez, J. A., & Herrera, F. (2012). Study on the impact of partition-induced dataset shift on k-fold cross-validation. *IEEE Transactions on Neural Networks and Learning Systems*, 23, 1304-1312.

Mohri, M., Rostamizadeh, A., & Talwalkar, A. (2012). Foundations of machine learning. MIT press.

Montereale-Gavazzi, G., Roche, M., Lurton, X., Degrendele, K., Terseleer, N., & Van Lancker, V. (2018). Seafloor change detection using multibeam echosounder backscatter: case study on the Belgian part of the North Sea. *Marine Geophysical Research*, 39, 229-247.

Parnum, I., Siwabessy, J., Gavrilov, A., & Parsons, M. (2009, June). A comparison of single beam and multibeam sonar systems in seafloor habitat mapping. *In the proceedings of the 3rd international conference and exhibition of underwater acoustic measurements: Technologies & Results,* Nafplion, Greece (pp. 155-162).

Pochon, X., Wood, S. A., Keeley, N. B., Lejzerowicz, F., Esling, P., Drew, J., & Pawlowski, J. (2015). Accurate assessment of the impact of salmon farming on benthic sediment enrichment using foraminiferal metabarcoding. *Marine Pollution Bulletin*, 100, 370-382.

Pohle, G., Frost, B., & Findlay, R. (2001). Assessment of regional benthic impact of salmon mariculture within the l'Etang Inlet, Bay of Fundy. *ICES Journal of Marine Science*, 58, 417–426.

Preston, J., Inouchi, Y., & Shioya, F. (2006). Acoustic classification of submerged aquatic vegetation. In *Proceedings of the Eighth European Conference on Underwater Acoustics*, ECUA (p. 317e322).

Reid, D., Scalabrin, C., Petitgas, P., Masse, J., Aukland, R., Carrera, P., & Georgakarakos, S. (2000). Standard protocols for the analysis of school based data from echo sounder surveys. *Fisheries Research*, 47, 125-136.

Ren, P., Yao, S., Li, J., Valdes-Sosa, P. A., & Kendrick, K. M. (2015). Improved prediction of preterm delivery using empirical mode decomposition analysis of uterine electromyography signals. *PLOS ONE*, 10, e0132116.

R.L. Naylor, R.J. Goldburg, J.H. Primavera, N. Kautsky, M.C.M. Beveridge, J. Clay, C. Folke, J. Lubchence, H., & Mooney, M. (2000) TroellEffect of aquaculture on world fish supplies *Nature*, 405, 1017-1024

Quinlan, J. R. (2014). C4. 5: Programs for Machine Learning. Elsevier.

Quintino, V., Freitas, R., Mamede, R., Ricardo, F., Rodrigues, A. M., Mota, J., Pérez-Ruzafa, Á., & Marcos, C. (2009). Remote sensing of underwater vegetation using single beam acoustics. *ICES Journal of Marine Science*, 67, 594-605.

Sáez, J. A., Luengo, J., Stefanowski, J., & Herrera, F. (2015). SMOTE–IPF: Addressing the noisy and borderline examples problem in imbalanced classification by a re-sampling method with filtering. *Information Sciences*, 291, 184-203.

Salvo, F., Wiklund, H., Dufour, S. C., Hamoutene, D., Pohle, G., & Worsaae, K. (2014). A new annelid species from whalebones in Greenland and aquaculture sites in Newfoundland: *Ophryotrocha cyclops*, sp. nov(*Eunicida: Dorvilleidae*). *Zootaxa*, 3887, 555-568.

Salvo, F., Hamoutene, D., & Dufour, S. C. (2015). Trophic analyses of opportunistic polychaetes (Ophryotrocha cyclops) at salmonid aquaculture sites. *Journal of the Marine Biological Association of the United Kingdom*, 95, 713-722.

Salvo, F., Dufour, S. C., & Hamoutene, D. (2017). Temperature thresholds of opportunistic annelids used as benthic indicators of aquaculture impact in Newfoundland (Canada). *Ecological indicators*, 79, 103-105.

Salvo, F., Oldford, V., Bungay, T., Boone, C., & Hamoutene, D. (2018). Guide for video monitoring of hard bottom benthic communities of the south coast of Newfoundland for aquaculture impact assessments. Canadian data report of fisheries and aquatic sciences. Fs 97013/1284E-PDF: ix + 41 p.

Siwabessy, P. J. W., Daniell, J., Li, J., Huang, Z., Heap, A. D., Nichol, S., Anderson, T.J., & Tran, M. (2013). Methodologies for seabed substrate characterisation using multibeam bathymetry, backscatter and video data: A case study from the carbonate banks of the Timor Sea, Northern Australia. *Geoscience Australia*.

Snellen, M., Siemes, K., & Simons, D. G. (2011). Model-based sediment classification using single beam echosounder signals. *The Journal of the Acoustical Society of America*, 129, 2878-2888.

Stefanowski, J., & Wilk, S. (2007). Improving rule based classifiers induced by MODLEM by selective pre-processing of imbalanced data. In *Proceedings of the RSKD Workshop at ECML/PKDD*, Warsaw (pp. 54-65).

Tulldahl, H. M., Philipson, P., Kautsky, H., & Wikström, S. A. (2013, June). Sea floor classification with satellite data and airborne lidar bathymetry. In *Ocean Sensing and Monitoring V* (Vol. 8724, p. 87240B). International Society for Optics and Photonics.

Wildish, D. J., Hughes-Clarke, J. E., Pohle, G. W., Hargrave, B. T., & Mayer, L. M. (2004). Acoustic detection of organic enrichment in sediments at a salmon farm is confirmed by independent groundtruthing methods. *Marine Ecology Progress Series*, 267, 99-105.

Wildish DJ, Pohle GW (2005) Benthic macrofaunal changes resulting from finfish mariculture. In: B. Hargrave (Ed*.), Environmental Effects of Marine Finfish Aquaculture*. Springer, Berlin

Wilding, T. A., Cromey, C. J., Nickell, T. D., & Hughes, D. J. (2012). Salmon farm impacts on muddy-sediment megabenthic assemblages on the west coast of Scotland. *Aquaculture Environment Interactions*, 2, 145-156.

Wolpert, D. H., & Macready, W. G. (1997). No free lunch theorems for optimization. *IEEE Transactions on Evolutionary Computation*, 1, 67-82.

Yang, C.S., Kao, S.P., Lee, F.B., & Hung, P.S. (2004) Twleve different interpolation methods: A case study of Surfer 8.0 In *Proceedings of the XXth ISPRS Congress*, 38, 778-785.

Chapter 3. Machine learning mediated benthic aquaculture impact assessment using oligonucleotide frequencies.

Abstract

Aquaculture is a rapidly expanding industry that now constitutes one of the primary sources of all consumed seafood. Intensive aquaculture production is associated with organic enrichment, which occurs as organic material settles on the seafloor, depleting oxygen and disrupting ecological processes. Bacteria have been shown to be sensitive bioindicators of organic enrichment, and supervised classifiers using features derived from 16s rRNA gene sequences have shown potential to become useful in environmental monitoring frameworks. Current taxonomy-based approaches, however, are time intensive and built upon emergent features which cannot easily be condensed into a monitoring pipeline. Here, we use a taxonomy-free approach to examine isolated 16s rRNA gene sequences derived from flocculent matter underneath and in proximity to hard bottom salmon aquaculture sites in Newfoundland, Canada. Tetranucleotide frequencies (k = 4) were tabulated from sample sequences and included as features in a machine learning pipeline using the random forest algorithm to predict four levels of benthic disturbance; resulting classifications were compared to those obtained using a published taxonomy-based approach. Our results show that k-mer count features can effectively be used to create highly accurate predictions of benthic disturbance and can resolve intermediate changes in seafloor condition. In addition, we present a robust assessment of model performance which accounts for the effect of randomness in model creation. This work outlines a flexible framework for environmental assessments at aquaculture sites that is both alignment and reference free.

3.1 Introduction

Aquaculture is a global industry producing over 80 million tonnes of food fish annually (FAO 2018). Over the last three decades the industry has seen continued growth in production and now contributes up to 46% of the global output of capture fisheries and aquaculture fisheries combined (FAO 2018).

However, concerns exist about the sustainability of aquaculture operations, in part due to the potential for negative environmental modification of associated ecosystems (Keeley et al., 2014; Salvo et al., 2017; Verhoeven et al., 2018). Effluent and particulate matter from aquaculture operations released into the environment can drive significant benthic community changes, the detrimental effects of which have been widely studied and show that extended aquaculture activities typically lead to changes in macrofaunal succession, decline in species diversity, and in some cases, complete elimination of native infauna (Keeley et al., 2014; Stoeck et al., 2018).

Tracking these detrimental changes is often performed through environmental monitoring and impact assessment programs, in which typical approaches include the characterization of macrofaunal biodiversity, as well as the detection or loss of specific indicator species associated with both ecosystem health and disturbance (Keeley et al., 2014; Salvo et al., 2017; Hamoutene et al., 2018). While effective, these methods are comparatively labour intensive as imaging data or environmental samples need to be collected, and taxonomic expertise and labour are required to obtain, analyse and interpret results (Maurer 2000; Cordier et al., 2019).

More recently, the utilization of high-throughput sequencing to characterize microbial communities has been explored as a more streamlined and automatable method for detecting ecosystem change (He et al., 2019; Cordier et al., 2019) Often, such methods involve the

amplification and sequencing of a marker gene (for bacteria, typically a portion of the 16S rRNA gene), combining closely related sequences into operational taxonomic units (OTU), which can then subsequently be used for elucidating the taxonomic composition of a community and gather information on the relative abundance of occurring OTUs (Pollock et al., 2018).

Microbial communities are sensitive to environmental stimuli (Logue et al., 2015), and previous work has highlighted the potential of using shifts in their taxonomic and OTU composition, as well detecting the presence of specific biomarker taxa in microbial communities to infer aquaculture-based impacts and organic enrichment at fish farms (Verhoeven et al., 2016, 2018; Stoeck et al., 2018). Nevertheless, several limitations can make the use of sequence and taxonomic based approaches suboptimal: taxonomic classification is inherently limited to classifying sequences for microorganisms that are identical or highly similar to those present within the reference databases used, which can leave a large proportion of sequences unclassified or classified at a less informative taxonomic level (Youssef et al., 2015), and thus unusable for biomarker studies. In addition, typical amplicon sequencing experiments produce high dimensional and sparse OTU datasets representing the complete genotypic diversity present in each investigated sample, from which extracting specific or co-occurring features significantly related to ecosystem status can be challenging (Gloor et al., 2017).

Such challenges can in part be addressed by combining marker-gene analysis with supervised machine learning (SML) approaches. SML algorithms generate predictive models based on user-supplied training datasets, from which specific features (or combination of features) correlating to the known classification are autonomously detected. Once established, the predictive model can subsequently be used to predict a classification for future, unknown, samples.

Within the context of biomonitoring, the integration of SML has enabled new approaches in analyzing amplicon data, including the possibility of employing a taxonomy- and reference database-free approach, using OTU sequences directly as inherent features of investigated environments. Recent work has shown that not only are OTU-SML based approaches capable of accurately predicting environmental biotic index values, they also outperform the traditional, taxonomy based assessment of these indices (Cordier et al., 2018). However, grouping sequences into OTUs has several undesirable properties, including sensitivity to bioinformatic pipeline and associated settings causing variations in OTU composition, the possibility of combining closely related sequences into phylogenetically incoherent OTUs, as well as the inherent inability to compare OTUs from different datasets, as the boundaries and membership of OTUs are dependent on, and invalid outside, the dataset in which they are defined (Callahan et al., 2017).

Instead, the distribution of oligonucleotides of specific length (k-mers), calculated from biological sequences, can be used as input features for performing machine learning (Asgari et al., 2018). Oligonucleotide distributions are a well-defined representation of 16S rRNA amplicon sequence data, in which sequence similarities are naturally incorporated, and are robust to bioinformatic pipeline and parameter variations, making them a particularly well suited feature set for downstream machine learning (Asgari et al., 2018). Indeed, recent studies have shown that k-mer representations of 16S rRNA gene sequencing experiments contain sufficient information for SML to accurately predict the phenotypical and environmental characteristics of biological samples, in a variety of applications (Asgari et al., 2018). As such, the usage of oligonucleotide distributions, coupled with SML, can potentially be a valuable tool in assessing changes to specific environmental niches in response to external stimuli, such as anthropogenic impacts.

Previously, using 16S rRNA gene sequencing, we reported that salmon aquaculture operations create significant benthic disturbances that in turn drive large scale specific shifts in benthic bacterial populations (Verhoeven et al., 2016, 2018). Here, we reanalyze 16S rRNA gene sequencing data from our previous study and investigate the potential for utilizing oligonucleotide distribution representations, specifically tetranucleotide frequencies (TNF, k=4), in combination with SML, as a possible automated method for predicting benthic disturbance levels.

3.2 Material and methods

3.2.1 Data description

This study examined a previously investigated microbiome dataset (NCBI BioProject PRJNA503189), containing Illumina based sequencing data of the V6–V8 16S rRNA gene region, performed on 108 flocculent matter samples collected below and near salmon aquaculture operations in Newfoundland, Canada (Verhoeven et al., 2018). Samples were previously assigned an environmental impact interpretation and categorized as recently disturbed (N=13), low (N=34), intermediate (N=19) or high (N =42) impact, based on bacterial biodiversity and the percentage of total organic carbon in relation to distance and production cycle (Verhoeven et al., 2018).

3.2.2 TNF calculation

TNF frequencies were calculated per sample by using a sliding window (k=4) across all sequences for each sample, summing TNF occurrences in a matrix. TNF occurrence count data was then subsequently normalized using the centered log ratio (clr) transform available in the *codaSeq* R package (Gloor & Reid 2016).

3.2.3 Supervised machine learning workflow

Model creation and statistical analysis were performed in R (v3.5.2) using the RStudio v1.1.463 IDE (R Core Team 2015). The caret package (v6.0-81) was used to partition data, perform cross-validation, hyperparameter optimization and model fitting (Kuhn 2008). Stratified random sampling was performed with *caret::createDataPartition* to maintain class ratios between training and test sets. Seventy-five percent of observations (N = 83) were included for model training with 25% (N = 25) withheld to evaluate model performance on unseen data. Predictive models were trained with ranger (v0.11.1), a multithreaded implementation of the random forest algorithm (Wright & Ziegler 2017) All visualizations were created with ggplot2 (Wickham 2009).

The *caret::trainControl* function was used to specify resampling and hyperparameter search methods. We used repeated, stratified 10-fold cross validation (CV) to search the hyperparameter space for tuples minimizing classification error. Ten folds were selected to reduce variance and to ensure that at least one of each class was present in each partition. One hundred repetitions were performed to account for covariate drift during division of training examples into folds and to ensure performance estimates had stabilized (Moreno-Torres et al., 2012). Hyperparameter tuning was done via grid search over CV folds with the best performing hyperparameter tuple fit to the entirety of training data. Tuned hyperparameters included: i) *mtry:* number of features randomly selected as candidates for each split, ii) *splitrule:* the split quality evaluation function and iii) *num.trees:* the number of trees in a forest. The number of trees was set to 2001, with an odd number specified to ensure no ties could occur during generation of class predictions. All 256 TNF combinations were included as features with values corresponding to their frequencies after clr transformation. *Caret::confusionMatrix.train* and *caret::confusionMatrix* were used to create confusion matrices and summary statistics for results from CV and test set predictions.

To account for the effect of randomness on model performance, a for-loop generating 2000 random seeds was created, with each iteration resulting in unique train/test splits, predictions and patterns of tree growth. Hyperparameters were held constant at values identified via grid search. These 2000 models were assigned to two groups with a Wilcoxon rank sum test (*base::wilcox.test*) applied to determine whether values were derived from the same distribution, signifying that results were not affected by inherent randomness in the model building process.

3.3 Results

3.3.1 Average accuracy of resampling folds

Hyperparameters resulting in the lowest classification error were fit to the entirety of training data by evaluating differing hyperparameter tuples on folds in our repeated cross validation procedure. Predictions made during CV were aggregated to provide initial estimates of model performance and stability.

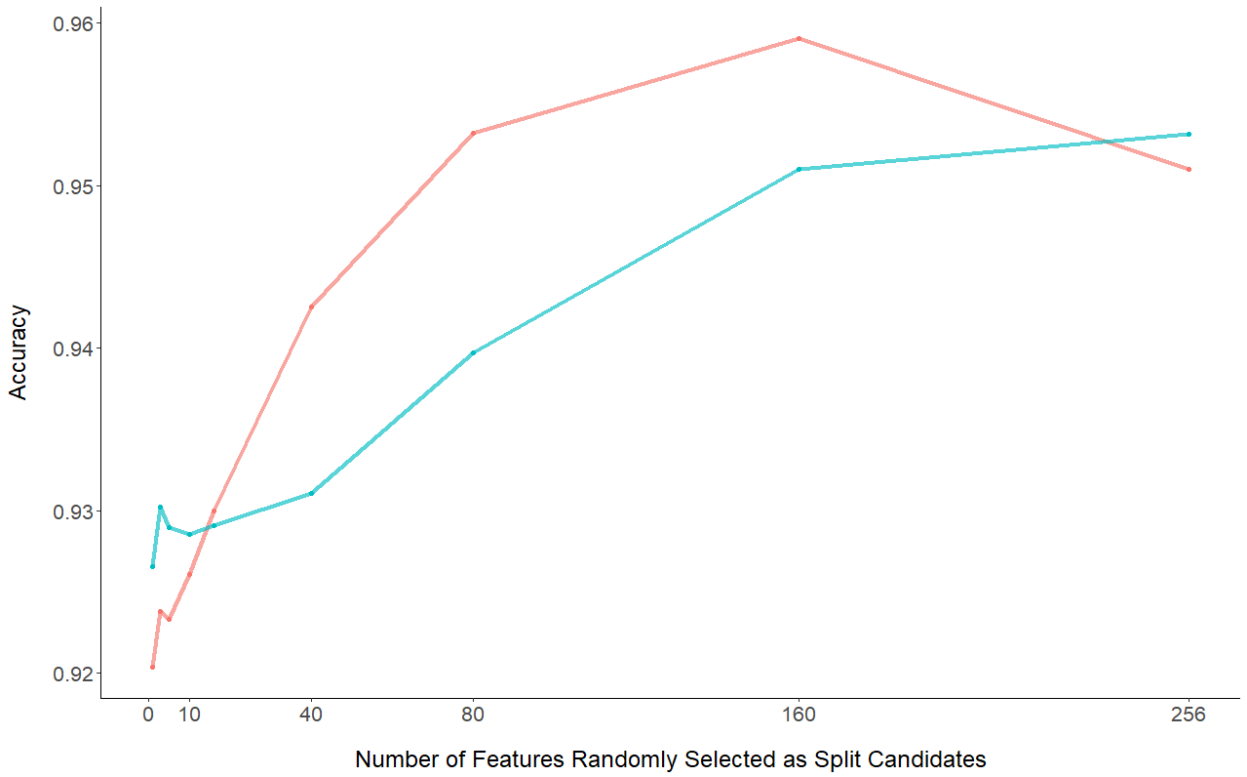

Figure 12. Performance of predictive models with varying hyperparameters. Shown are the accuracy of aggregated validation scores created through repeated 10-fold cross validation on the vertical axis, while the number of randomly selected features as split candidates (*mtry*) are indicated on the horizontal axis. Red and blue lines correspond to gini and extratrees (*splitrule)*, respectively, which determine how the algorithm creates decision tree splits.

Predictions generated on resampling folds resulted in an average accuracy of 0.959, with the lowest performance reported at 0.920. Accuracy increased with *mtry* size before peaking at 160 features. Setting the *splitrule* hyperparameter to gini impurity resulted in higher accuracy predictions when compared to extratrees. Accuracies for models with varying hyperparameters were between the range of 0.920 - 0.959.

Low impact samples were most consistently predicted with 0.997 of cases accurately reported. Conversely, intermediate disturbance level was the least accurate category (0.860) with incorrect classifications being labelled as low and recently impacted for 7% and 6% of cases, respectively. Similarly, 3% of high impact predictions were misclassified as intermediate impact. Overall, only 1.7% of observations were misclassified by >1 level of impact.

Table 4. Confusion matrix of aggregated counts from 10-fold, 100 repetition cross validation created during hyperparameter search (N=83). Column and row values correspond to known and predicted cases of seafloor disturbance, respectively, with four levels of seafloor disturbance ranging from low to high.

		Actual impact			
		Low	Recent	Intermediate	High
Predicted impact	Low	2591	0	116	0
	Recent	0	979	94	1
	Intermediate	9	0	1290	99
	High	0	21	0	3100

3.3.2 Model evaluation on withheld test data

Predicted labels on withheld test data showed a high level of agreement with known cases for all disturbance levels (Table 5). Model predictions significantly outperform (p-value: 1.126e-10) the No Information Rate which represents a naïve prediction of all observations belonging to the majority class.

Table 5. Confusion matrix demonstrating model performance on withheld test data (N=25) when predicting levels of seafloor disturbance ranging from low to high. Column and row values correspond to known and predicted cases of seafloor disturbance, respectively, with four levels of seafloor disturbance ranging from low to high.

		Actual impact			
		Low	Recent	Intermediate	High
Predicted impact	Low	8	0	0	0
	Recent	0	3	0	0
	Intermediate	0	0	4	0
	High	0	0	0	10

In addition, random seed iteration testing with constant hyperparameters indicated that the random seed state did not significantly impact accuracy scores, with mean and median model accuracies of 0.9719 and 0.96, being detected respectively. In addition, most models (0.9985) fell within the

95% CI created with *caret:confusionMatrix* which performs an exact binomial test to determine the probability of success in a Bernouli experiment. A Wilcoxon rank sum test further indicates all random instances of train/test splits and tree growth are derived from the same distribution (W: 494290, p-value: 0.6288), indicating that predictive performance is not attributable to the effect of randomness in data partitioning and model creation.

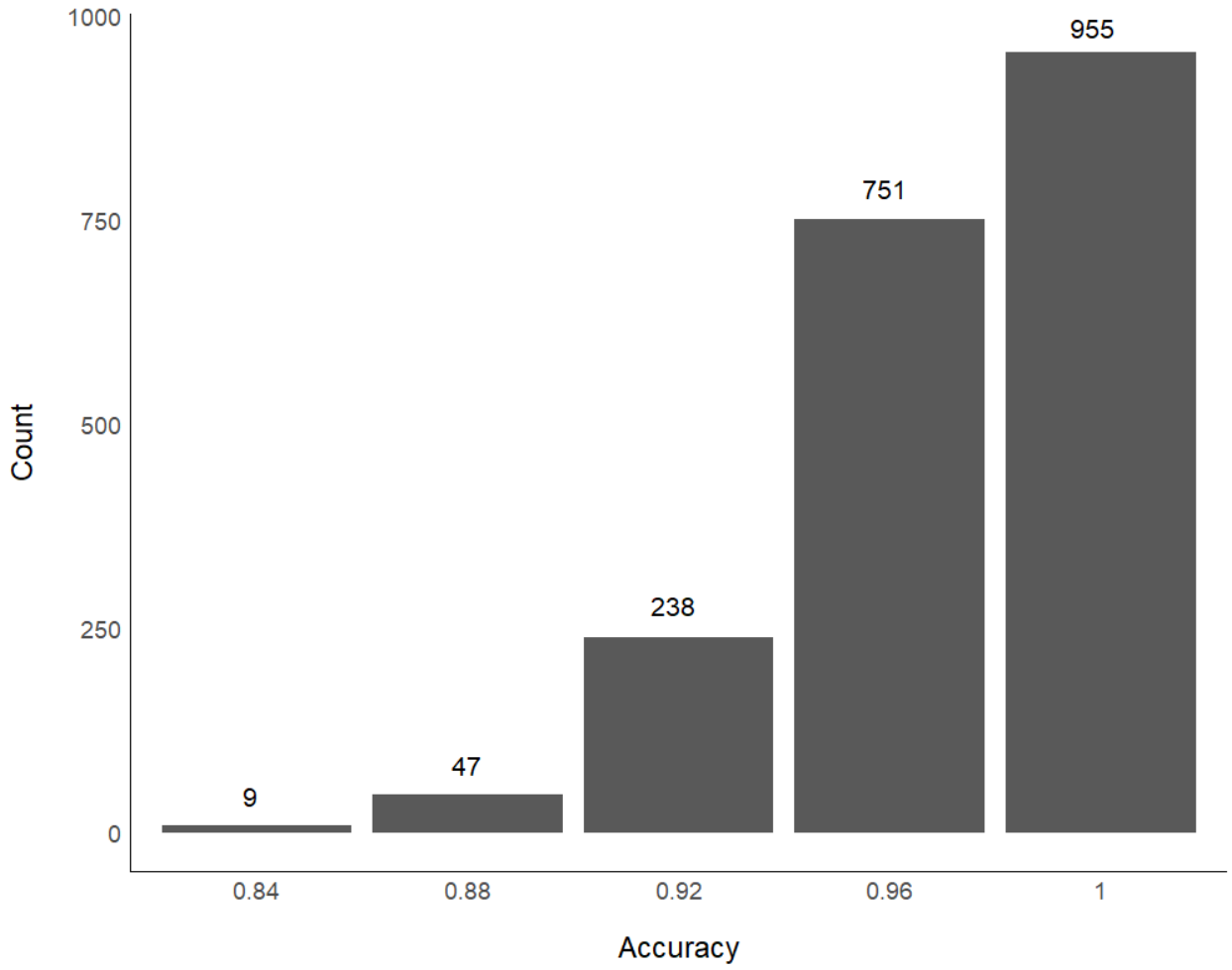

Figure 13. Frequency of accuracy results for 2000 models with unique seed states to assess the effect of randomness on model performance. For each iteration, new train/test splits are created, with models fit using identical hyperparameters selected via grid search. This process controls for variables like hyperparameter and algorithm selection while evaluating the effect of randomness associated with stratified random sampling and patterns of tree growth.

3.4 Discussion

Classification of benthic disturbances near aquaculture sites has received increased attention, but practical application and the potential for real-time assessment have yet to be presented. Here, we examined the use of k-mer count features (k = 4) in a model tasked with classifying levels of benthic disturbance at aquaculture sites and demonstrate that highly accurate predictions of seafloor condition can be generated using a defined feature set.

Traditionally, biotic indices which examine macroinvertebrate richness and diversity have been used to assess ecological quality and disturbance level (Borja & Dauer 2008; Rygg & Norling 2013). Recently, bacterial eDNA metabarcoding has shown promise by associating specific bacterial community compositions with environmental disturbances, demonstrating their potential as highly responsive bioindicators (Lejzerowicz et al., 2015; Stoeck et al., 2018). Machine learning has seen increased use with features derived from 16s rRNA sequencing as model inputs such as OTUs which effectively predict biotic indices (Cordier et al., 2017), outperforming taxonomy-based assessments and providing faster evaluation of seafloor condition (Cordier et al., 2018). K-mer count features have been used to accurately classify distinct ecological environments (Asgari et al., 2018), but their ability to resolve within-environmental change has not been previously addressed.

In this study, we demonstrate that TNFs can effectively distinguish levels of benthic disturbance with performance maintained across validation folds (Table 4), holdout test data (Table 5) and 2000 iterations of model creation with random train/test splits (Figure 13). While longer k-mer count features have been found to improve classification performance (Alsop & Raymond 2013; Vervier et al., 2016; Asgari et al., 2018), the discriminatory power of tetranucleotides is well documented (Teeling et al., 2004; Yoon et al., 2017) and their use in the current context balances

performance with computation time by limiting features which increase quartically with k-mer length. The use of TNF features in a supervised classifier circumvents taxonomic assumptions associated with seafloor condition and simplifies data processing by restricting the feature set to 256 tetranucleotide combinations (Asgari et al., 2018). This is a desirable quality in developing monitoring pipelines as it standardizes predictive model inputs. TNF-based classifications do not require sequence alignments and reference databases to identify bacterial groups or the construction of OTUs which can vary depending on settings used in diverse bioinformatics pipelines and may not reflect genuine taxonomic relationships. Furthermore, OTU construction and taxonomy-based approaches have emergent and location dependent features which are difficult or impossible to standardize, with comparisons between sample sites not possible with OTUs constructed from different datasets (Callahan et al., 2017).

While the utility of machine learning approaches, like those used in this paper, is indisputable, concerns regarding reproducibility of machine learning algorithms and the reporting of model performance have been raised (Drummond 2009; Henderson et al., 2017; Colas et al., 2018). Setting seed states allows random events, such as partitioning data into training and test sets to be replicated and compared, but replication may not be sufficient to arrive at genuine performance estimates (Drummond 2009). In several cases, the best or average of n-best performing seeds is selected for publication (Henderson et al., 2017; Colas et al., 2018). This behavior of seed optimization is problematic as it allows investigators to report seeds resulting in good performance without disclosing trials with poorer outcomes or those which do not improve upon currently existing benchmarks. Including results from numerous seed states demonstrates that our model is stable over random iterations accounting for differences in train/test splits and patterns of tree growth (Figure 2). Model stability over 2000 train/test splits accounts for variance associated with

the small (but representative) sample retained for model evaluation (N = 25). When computationally feasible, we recommend these statistics be reported.

In conclusion, k-mer features such as TNF are a valuable addition to the benthic assessment toolkit, reducing computation costs associated with sequence alignment and reference database comparison while outperforming OTU and taxonomic features when predicting environment types (Asgari et al., 2018). Future studies should examine larger collections over wider geographic areas as to better characterize the robustness of seafloor condition boundaries and assess the generalizability of predictions over larger spatial scales. Additionally, the establishment of an open source database of sequenced samples near aquaculture sites and the inclusion of different 16s rRNA hypervariable regions could provide increased flexibility to seafloor condition classifications and allow investigators to detect high resolution changes with a variety of eDNA sequencing pipelines.

References

Alsop E.B., & Raymond J (2013). Resolving Prokaryotic Taxonomy without rRNA: Longer Oligonucleotide Word Lengths Improve Genome and Metagenome Taxonomic Classification. *PLOS ONE*, 8, e67337.

Asgari E, Garakani K, McHardy A.C., & Mofrad MRK (2018) MicroPheno: predicting environments and host phenotypes from 16S rRNA gene sequencing using a k-mer based representation of shallow sub-samples. *Bioinformatics*, 34, 32–42.

Borja A & Dauer D.M. (2008). Assessing the environmental quality status in estuarine and coastal systems: Comparing methodologies and indices. *Ecological Indicators*, 8, 331–337.

Callahan B.J., McMurdie P.J., & Holmes S.P. (2017). Exact sequence variants should replace operational taxonomic units in marker-gene data analysis. *The ISME Journal*, 11, 2639–2643.

Colas C, Sigaud O., & Oudeyer P.Y. (2018). How Many Random Seeds? Statistical Power Analysis in Deep Reinforcement Learning Experiments. *arXiv*:180608295.

Cordier T, Esling P, Lejzerowicz F, Visco J, Ouadahi A, Martins C, Cedhagen T & Pawlowski J (2017). Predicting the Ecological Quality Status of Marine Environments from eDNA Metabarcoding Data Using Supervised Machine Learning. *Environmental Science and Technology*, 51, 9118–9126.

Cordier T., Forster D., Dufresne Y., Martins C.I.M., Stoeck T., & Pawlowski J. (2018). Supervised machine learning outperforms taxonomy-based environmental DNA metabarcoding applied to biomonitoring. *Molecular Ecology Resources*, 18, 1381-1391.

Cordier T., Lanzén A., Apothéloz-Perret-Gentil L., Stoeck T., & Pawlowski J. (2019). Embracing Environmental Genomics and Machine Learning for Routine Biomonitoring. *Trends in Microbiology*, 27, 387–397.

Drummond C. (2009). Replicability is not Reproducibility:Nor is it Good Science. In *Proceedings of the Evaluation Methods for Machine Learning Workshop*, Montreal, Canada

FAO (2018) *The State of World Fisheries and Aquaculture 2018 - Meeting the sustainable development goals*. Rome. FAO.

Gloor G.B., Macklaim J.M., Pawlowsky-Glahn V., & Egozcue J.J., (2017). Microbiome Datasets Are Compositional: And This Is Not Optional. *Frontiers in Microbiology*, 8, 2224.

Gloor G.B., & Reid G. (2016). Compositional analysis: a valid approach to analyze microbiome high-throughput sequencing data. *Canadian Journal of Microbiology*, 62, 692–703.

Hamoutene D., Salvo F., Cross S., Dufour S.C., & Donnet S. (2018). Linking the presence of visual indicators of aquaculture deposition to changes in epibenthic richness at finfish sites installed over hard bottom substrates. *Environmental Monitoring and Assessment*, 190, 750.

He X., Sutherland T.F., Pawlowski J., & Abbott C.L. (2019). Responses of foraminifera communities to aquaculture-derived organic enrichment as revealed by environmental DNA metabarcoding. *Molecular Ecology*, 28, 1138–1153.

Henderson P., Islam R., Bachman P., Pineau J., Precup D., & Meger D. (2017). Deep Reinforcement Learning that Matters. *arXiv*:170906560.

Keeley N.B., Macleod C.K., Hopkins G.A., & Forrest B.M. (2014). Spatial and temporal dynamics in macrobenthos during recovery from salmon farm induced organic enrichment: When is recovery complete? *Marine Pollution Bulletin*, 80, 250–262.

Kuhn M. (2008). Building Predictive Models in R Using the caret Package. *Journal of Statistical Software*, 28, 1–26.

Lejzerowicz F., Esling P., Pillet L., Wilding T.A., Black K.D., & Pawlowski J. (2015). High-throughput sequencing and morphology perform equally well for benthic monitoring of marine ecosystems. *Scientific Reports*, 5, 13932.

Logue J.B., Findlay S.E.G., & Comte J. (2015). Microbial Responses to Environmental Changes. *Frontiers in Microbiology*, 6, 1364.

Maurer D. (2000). The Dark Side of Taxonomic Sufficiency (TS). *Marine Pollution Bulletin*, 40, 98–101.

Moreno-Torres J.G., Saez J.A., & Herrera F. (2012) Study on the Impact of Partition-Induced Dataset Shift on k-Fold Cross-Validation. *IEEE Transactions on Neural Networks and Learning Systems*, 23, 1304–1312.

Pollock J., Glendinning L., Wisedchanwet T., & Watson M. (2018). The Madness of Microbiome: Attempting To Find Consensus "Best Practice" for 16S Microbiome Studies. *Applied Environmental Microbiology*, 84, e02627-17.

R Core Team (2015). R: A Language and Environment for Statistical Computing. Available at: http://www.R-project.org/.

Rygg B., & Norling K. (2013). Norwegian Sensitivity Index (NSI) for marine macroinvertebrates, and an update of Indicator Species Index (ISI).

Salvo F., Mersereau J., Hamoutene D., Belley R., & Dufour S.C. (2017). Spatial and temporal changes in epibenthic communities at deep, hard bottom aquaculture sites in Newfoundland. *Ecological Indicators*, 76, 207–218.

Stoeck T., Frühe L., Forster D., Cordier T., Martins C.I.M.,, & Pawlowski J. (2018). Environmental DNA metabarcoding of benthic bacterial communities indicates the benthic footprint of salmon aquaculture. *Marine Pollution Bulletin*, 127, 139–149.

Teeling H., Meyerdierks A., Bauer M., Amann R., & Glöckner F.O. (2004). Application of tetranucleotide frequencies for the assignment of genomic fragments. *Environmental Microbiology*, 6, 938–947.

Verhoeven J.T.P., Salvo F., Hamoutene D., & Dufour S.C. (2016). Bacterial community composition of flocculent matter under a salmonid aquaculture site in Newfoundland, Canada. *Aquaculture Environment Interactions*, 8, 637–646.

Verhoeven J.T.P., Salvo F., Knight R., Hamoutene D., & Dufour S. (2018). Temporal bacterial surveillance of salmon aquaculture sites indicates a long lasting benthic impact with minimal recovery. *Frontiers in Microbiology*, 9, 3054.

Vervier K., Mahé P., Tournoud M., Veyrieras J.B., & Vert J.P. (2016). Large-scale machine learning for metagenomics sequence classification. *Bioinformatics*, 32, 1023–1032.

Wickham H. (2009). *Ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag, New York.

Wright M.N., & Ziegler A. (2017). Ranger: A Fast Implementation of Random Forests for High Dimensional Data in C++ and R. *Journal of Statistical Software*, 77, 1–17.

Yoon S-H., Ha S-M., Kwon S., Lim J., Kim Y., Seo H., & Chun J. (2017). Introducing EzBioCloud: a taxonomically united database of 16S rRNA gene sequences and whole-

genome assemblies. *International Journal of Systematic and Evolutionary Microbiology*, 67, 1613–1617.

Youssef N.H., Couger M.B., McCully A.L., Criado A.E.G., & Elshahed M.S. (2015). Assessing the global phylum level diversity within the bacterial domain: A review. *Journal of Advanced Research*, 6, 269–282.

Chapter 4. Summary and General Conclusions

4.1 Development of a machine learning mediated benthic monitoring pipeline

As the aquaculture industry expands, so does its environmental impact. It is critically important that monitoring techniques be developed that can effectively scale with the industry's expansion to ensure that marine resources are managed effectively and sustainably. Of equal importance is the ability to not only resolve changes in seafloor condition, but to monitor these changes over time, using techniques which facilitate large scale and accurate analysis of collected data. The development of more sensitive protocols will help ensure that aquaculture sites are given enough time to recover from production cycles, as inadequate fallowing periods result in a deterioration in seafloor health and a rapid return to pre-fallow conditions (Keeley et al., 2015).

Present methods that rely on discrete measurements captured by images and sediment samples are costly in terms of time and labour as point measurements are insufficient to accurately delineate the spatial bounds of organic matter deposition. Areas are too large and depositional patterns too unintuitive for true spatial distributions to be quantified using devices which need to repeatedly be lowered and recovered from the bottom of the ocean.

Like previous studies examining MBE (Wildish et al., 2004) and side-scan imagery (Hughes et al., 2002), we demonstrate that SBE hydroacoustic data can detect the presence of flocculent matter derived from aquaculture production. While machine learning algorithms can create modestly performing models derived from SBE data, their performance is contingent on the use of appropriate equipment during the data collection phase. Using the method outlined in chapter 2, a lower frequency echosounder will likely result in higher signal quality and improved model performance. Ideally, the development of this technique will facilitate assessments independent of groundtruth collection. While this outcome requires further refinement, we prescribe steps to

expedite the process. In particular, the collection of hydroacoustic data with site appropriate equipment and groundtruth methods should be mindful of the predictive modelling process. Merging groundtruthing information with hydroacoustic data could be made more effective by using a weighted neighbor analysis which accounts for distances of multiple observations within the merging threshold. This process would reduce error inherent in the model building process by decreasing single observation variance.

The SBE approach introduced in chapter two highlights inherent limitations of high frequency echosounders to detect a benthic aquaculture footprint in deep water and hard bottom substrates. The principal limitations of this technique involve depth restrictions and the acoustic properties of high frequency sounds which largely reflect off hard surfaces (Preston et al., 2006; Quintino et al., 2009). Alternatively, our SBE approach is more appropriate for locations with homogenous, soft bottom seafloors <50 m in depth to ensure that acoustic energy is returned to the transducer at optimal angles, improving signal quality (Snellen et al., 2011). The collection of hydroacoustic data in shallower depths would allow information associated with 1st and 2nd echoes to be reliably collected, which are features commonly used in substrate classification (Bates et al., 2001; Foster et al., 2009). With higher quality acoustic data, the inclusion of different features in the model building process may improve performance. Examples of potential features include distances of observations from the cage array centroid and bathymetric measurements of slope, curvature and aspect.

Future attempts should explore aquaculture sites which have had cages removed. In this way, hazards which prohibit the use of ROVs and AUVs can be avoided, and continuous imagery of the seafloor can be collected which follows survey vessel paths while collecting hydroacoustic data, further reducing error associating with merging these two sources of data. Once predictive

models have learned from sufficient examples to effectively discriminate flocculent matter from other substrates at test sites, hydroacoustic data alone could be used to predict the presence of flocculent matter at novel locations. This development would result in assessments that capture the maximal spatial extents of flocculent matter deposition, and identify high probability areas of collecting sediment samples, reducing instances of failed sampling attempts. Additionally, this technique could be useful to establish baseline site characteristics for later comparison with data collected after the onset of aquaculture production.

In chapter 3, we demonstrate that bacterial oligonucleotides obtained from eDNA are highly effective features which can discriminate intermediate levels of benthic impact. To our knowledge, this is the first application of oligonucleotide frequency to determine within-environmental change, which builds on previous work which examined their use to distinguish distinct environments (Asgari et al., 2018). The use of oligonucleotide frequencies has several advantages compared to other machine learning approaches which use eDNA derived sequence data as features. Most importantly, oligonucleotide frequencies circumvent the need to resolve taxonomic relationships of bacterial sequences or the construction of OTUs which reduces computation costs and model complexity. Compared to the aforementioned techniques, our oligonucleotide approach simplifies and controls the modelling process by condensing the data into a uniform feature set. While OTUs and taxonomies are diverse depending on sample location, oligonucleotide frequencies are restricted by the number of k-mer word combinations nucleotides can form and are therefore applicable regardless of sampling location. Further development of this technique could facilitate on-site assessments of seafloor conditions, circumventing lengthy turnarounds and costs associated with outsourcing expertise for sediment analysis while reducing the risk of sample degradation. In Newfoundland, hard bottom seafloors

present difficulties when acquiring sediment samples. Water samples collected at aquaculture sites recorded increased abundances of bacterial communities associated with flocculent matter (Verhoeven et al., 2018). Potentially, our oligonucleotide approach could be adapted to bacterial eDNA sequenced from water samples rather than sediment. The ability of water sampling as a benthic monitoring tool requires further exploration but could circumvent difficulties present with current techniques.

Together, the two techniques introduced in this thesis offer an exciting prospect for the future of benthic monitoring and environmental assessments in relation to aquaculture production. Echosounder data and machine learning can help map the true extents of seafloor impact, while machine learning applied to eDNA can determine levels of benthic disturbance and its change over distance and time. These technologies and their applications have direct benefit for both regulators and members of industry, providing more accurate assessments of benthic condition while reducing the labor associated with collecting seafloor data.

4.2 References

Asgari E., Garakani K., McHardy A.C., & Mofrad M.R.K. (2018). MicroPheno: predicting environments and host phenotypes from 16S rRNA gene sequencing using a k-mer based representation of shallow sub-samples. *Bioinformatics*, 34, 32–42.

Bates, C. R., Whitehead, E. J., & Castle, B. (2001). Echo Plus measurements in Hopavagen Bay, Norway. *Sea Technology*, 42, 34-43.

Foster, G., Walker, B. K., & Riegl, B. M. (2009). Interpretation of single beam acoustic backscatter using lidar-derived topographic complexity and benthic habitat classifications in a coral reef environment. *Journal of Coastal Research*, 1, 16-26.

Hughes Clark, J. E., Wildish, D., & Duxfield, A. (2002). Acoustic imaging of salmonid mariculture sites. CHC 2002 Proceedings.

Preston, J., Inouchi, Y., & Shioya, F. (2006). Acoustic classification of submerged aquatic vegetation. In *Proceedings of the Eighth European Conference on Underwater Acoustics*, ECUA (p. 317e322).

Snellen, M., Siemes, K., & Simons, D. G. (2011). Model-based sediment classification using single beam echosounder signals. *The Journal of the Acoustical Society of America*, 129, 2878-2888.

Keeley, N. B., Forrest, B. M., & Macleod, C. K. (2015). Benthic recovery and re-impact responses from salmon farm enrichment: implications for farm management. *Aquaculture*, 435, 412-423.

Quintino, V., Freitas, R., Mamede, R., Ricardo, F., Rodrigues, A. M., Mota, J., Pérez-Ruzafa, Á., & Marcos, C. (2009). Remote sensing of underwater vegetation using single beam acoustics. *ICES Journal of Marine Science*, 67, 594-605.

Verhoeven, J. T. P., Salvo, F., Knight, R., Hamoutene, D., & Dufour, S. (2018). Temporal bacterial surveillance of salmon aquaculture sites indicates a long-lasting benthic impact with minimal recovery. *Frontiers in Microbiology*, 9, 3054.

Wildish, D. J., Hughes-Clarke, J. E., Pohle, G. W., Hargrave, B. T., & Mayer, L. M. (2004). Acoustic detection of organic enrichment in sediments at a salmon farm is confirmed by independent groundtruthing methods. *Marine Ecology Progress Series*, 267, 99-105.