School of Engineering and Computer Science
Te Kura Mātai Pūkaha, Pūrorohiko

PO Box 600
Wellington
New Zealand

Tel: +64 4 463 5341
Internet: office@ecs.vuw.ac.nz

Rapid determination of bulk composition and quality of marine biomass in Mass Spectrometry

Jesse Wood

Supervisors: Bach Hoai Nguyen, Bing Xue, Mengjie Zhang, Daniel Killeen

Submitted in partial fulfilment of the requirements for Doctorate of Philosophy - Artificial Intelligence.

Abstract

This document gives some ideas about how to write a project proposal, and provides a template for a proposal. You should discuss your proposal with your supervisor.

Contents

Chapter 1

Introduction

> Go maybe even one step further and would say, the next phase must comprise much more the application of those Technologies, in areas where it is not yet applied, and I'm not thinking only of developing countries. I'm thinking of Education, I'm thinking of Agriculture, if you look at just those two areas, they are very old-fashioned, so this is a large let's say opportunity for those Technologies to penetrate much better, and serve the overall economy [1].
>
> *Klaus Schwab*
> *Founder - WEF*

The introduction provides a **Problem Statement**, **Motviations**, **Limitations**, **Research Goals**, **Summary**, and **Organisation of the Proposal**. Each of those sections will be explored in greater detail in the remainder of this chapter.

1.1 Problem Statement

Problem statement: This research aims aides the Cyber-marine Research Programme's goal of maximizing waste utilization in fish processing, and maximized value for harvested and aquacultured marine biomass.

This proposal is about fish analysis - rapid determination of bulk composition and quality of marine biomass in Mass Spectrometry. Specifically, this research aim to identify the type of fish and assess its suitability for use in fish products. It is undertaken in collaboration with Plant & Food Research [2] and Callaghan Innovation [3]. This research serves as a proof-of-concept as part of a larger joint endeavour, the Cyber-marine research programme [4], which aims to achieve 100% utilisation and maximised value for all harvested wild and aquacultured seafood.

This section introduces the scope for this research. This research is application-based, it intends to fill a gap in the fish processing industry, by using machine learning to analyze rapid mass spectrometry. To understand that gap, this sections explains the global fishing

industry, New Zealand's unique place in the fishing industry, the potential for automation in fish processing, the importance of AI applications, and the counter-intuitive requirements for industry adoption. Specifically, this section covers the global fishing industry, fish processing in New Zealand, the potential for automation, application-based AI and industry adoption.

1.1.1 Global Fishing Industry

This research focuses on improving waste utilization in the global fishing industry. According to [5], approximately 100 million tonnes of wild fish are captured each year, and only about 40% of these fish are processed into edible parts. The remaining portions are often processed into fish oil and fish meal, or discarded as non-fillet material. In addition, many fisheries are in decline and global fishing has not significantly increased in the past 30 years, making waste utilisation an important focus globally. The fishing industry must maximize the utilization and value of every kilogram of marine biomass to preserve our fish stocks and ensure there are plenty of fish in the sea for future generations to reel in.

The many steps in the supply chain from ocean to plate, are prone to human error and criminal activity. Consider the 2013 European Horse Meat Scandal. Adulteration watered down high-value beef mince products with low-value horse meat, and sold them to an unaware public, as a criminal enterprise to increase profits. The beef with adulteration applies to the global fishing industry. According to [6] a meta-analysis comprised of 51 studies of the global fishing industry, there was an average mislabelling rate of 30%. Consumers of fish products want to be confident they know what are eating, fish processing plants must ensure the labels on seafood products are accurate. Tools for quality assurance that can determine the composition and quality of fish products are needed.

1.1.2 Fish Processing in New Zealand

The New Zealand fishing industry prides itself on sustainability. New Zealand fisheries are well-regulated with strict quotas for over 100 marine species [7]. The NZ fishing industry does not have many 'high volume' fisheries, e.g. Hoki our largest fishery, as approximately 110,010 tonnes of quota each year [8]. On a global scale, this is minuscule, Norway alone have an aquaculture production of salmon of 4,000,000 tonnes a year [9]. This makes it difficult for fish processing, due to the variability in the catches, different boatloads of fish require different processing to maximize their value. The MBIE CyberMarine programme [4] seeks to develop a flexible factory, that can rapidly determine the composition of incoming fish biomass, and then choose an optimal processing route for this largely NZ-specific problem.

1.1.3 Potential for Automation

We aim to employ machine learning techniques to detect spoilage indicators, Quality Control, and contamination (ideally) on fresh marine biomass. Tools for quality control in fish processing are needed. Marine biomass is highly prone to spoilage, and spoiled products cannot be sold. Spoilage can include enzymatic spoilage, where the proteases and lipases inside the fish begin to digest animals, microbial digestion, or due to oxidation in the air. The lipids in marine biomass make them especially prone to oxidation in the air because they are highly unsaturated. Marine biomass must be handled extremely carefully after it is caught to prevent this oxidation. Cyber-Marine is interested in deploying machine learning techniques to measure the level of oxidation in marine biomass. This can be used as a marker for quality control in fish processing. There are numerous other Quality Control parameters for

marine products, especially so for marine oils, this work seeks machine learning techniques that can accurately profile these QC parameters also. Marine biomass can be contaminated with several things, for example, plastics and mineral oil - which is carcinogenic (it kills). This work seeks to develop tools that can identify contamination in marine biomass. Techniques that work on fresh (uncooked) marine biomass are needed, as cooking the fish can destroy valuable proteins, collagen and active enzymes. Cooking is also energy-intensive and time-consuming, it adds time and cost to fish processing, so processing fresh marine biomass is preferred.

1.1.4 Application-based AI

> Go do AI for some vertical [10].
>
> *Sam Altman*
> *Co-founder CEO - OpenAI*

Artificial Intelligence needs to penetrate markets outside of technology and academia - it is time for AI-applications. In the quotes given as preamble to this chapter, Altman and Scwhab, outline the need for AI to be applied to industry. A vertical refers to a specific industry or field of expertise, Scwhab refines this assessment by outlining the need for AI-powered innovation in the agriculture, a closely adjacent and related field to harvested and aquacultured marine biomass - the industry of this research. AI applications need to apply a Goal-oriented design [11], and address the needs of the domain experts, i.e. the chemists, to achieve their goals. Goal-oriented application-based AI will aid the chemist in their job, not replace them. In order to provide benefit to these highly specialized and trained practitioners, chemists need to understand how these systems work, and trust their predictions. Building trust in Aritificial Intelligence for industry adoption is the focus of the next subsection.

1.1.5 Industry

Callaghan innovation hosted an industry workshop for the Cyber-Marine project [4]. The work in [12] was presented, as well the research several chemists like [13]. The three most important takeaways from that workshop were:

- **Adoption** - for the adoption of technology, for example, AI models, models that can be understood and trusted by domain experts are needed.
- **Explainable AI** - XAI is almost more important than accuracy, for the adoption of technology by domain experts in academia and industry.
- **Economic incentive** - for adoption in the industry, there needs to be an economic incentive, accuracy is not enough! There need to be profits.

Given the overwhelming presence of agile methodology in the tech industry, and technology stunts like Dalle-2, ChatGPT or Microsoft Bing. It has become overwhelming clear, the need for hands-on demonstration of working products. The research can be state-of-art, but without clear science communication, and demonstration of its relevance to the appropriate stakeholders, those research papers will never be converted into real-world applications. The knowledge gap, or more accurately knowledge canyon, between industry and academia, needs to bridged to fully utilize AI technology.

Explainable AI, is important, to move way from pre-conceived academic notions of interpretability [14], and move towards tools that can be understood by their users [15, 16]. Domain-specific AI-powered tool that aid practitioners where they need it most. Chemists will not be replaced by AI tools, rather replaced by another chemist using these AI tools.

> Any sufficiently advanced technology is indistinguishable from magic [17].
>
> *Arthur C. Clarke*

Automation of fish processing reduced laborious manual labour, and expensive domain expertise, and speed up production lines. To meet the requirements of a factory setting, models are needed that can be deployed and understood in real-time. This is challenging, reduces the scope of machine learning techniques, eliminates black-box methods, and focuses this work on explainable AI, whose models can be reasoned with by domain experts from chemistry without prior machine learning knowledge. These domain experts, chemists, need to build trust in the predictions of the model, understand the nuts and bolts, and be able to verify/troubleshoot the model in real-time. This gives the constraints of accurate, efficient and interpretable models.

1.2 Motivations

This section addresses the motivations of this research. This proposal seeks to address real-world applications of AIML in fish processing, for the analysis of rapid mass spectrometry. These applications apply to fish identification, contamination detection, and traceability. The motivations behind each of these research goals, and a brief introduction, will be addressed in the remainder of this section.

1.2.1 Identification

This subsection describes existing works in profiling biomass, and how this research can be extended with AIML techniques to automate fish processing in New Zealand. To apply AIML techniques for identification of marine biomass using rapid mass spectrometry, the models need to address three key areas. Specifically, these models must cope with three issues, (1) *Heterogeneous marine biomass*, (2) *Low sample complexity*, (3) *Concept drift*.

Identification provides relevant information to profile a sample of marine biomass. These profiles include, but are not limited to, the species of the fish, and the body part from which the sample was taken. The Cyber-Marine flex-factory [4] aims to maximize waste utilization of marine biomass. Therefore, identifying characteristics of marine biomass waste, such as their species and body part, is useful. This knowledge informs decisions on how best to reduce, reuse and re-cycle that waste, to maximize the value of that marine biomass.

Existing works into identification of biomass, let alone marine biomass, using rapid spectrometry are limited to [18, 19]. Due to Rapid Evaporative Ionisation Mass Spectrometry (REIMS) [20] being a recent technological development in chemistry, and the diffusion of innovation [21], access to the REIMS mass spectrometer, and subsequent research, and real-world applications of said technology, is sparse. As of the time of writing, the tools are cost-prohibitive for widespread adoption and use in industry. However, this work, as part of the

greater Cyber-Marine research project [4] serves as a proof-of-concept, for the adoption of REIMS for rapid analysis of marine biomass in the factory of the future - the flex-factory. This research aims to demonstrate the use case for rapid mass spectrometry in real-world applications of fish processing. Rapid spectrometry has been shown effective in detecting adulteration in biomass, specifically beef mince that is contaminated with horse meat [18]. Adulteration is the (often criminal) process of debasing the quality of food products, by intentionally mixing them with products of lower value, to maximize profits, and dishonestly sell them labeled as ONLY the higher value product [20]. The study, [18], showed that REIMS can detect adulteration of beef samples with cross-species contamination at levels as low as 1%, for certain horse-meat offal. Rapid spectrometry has demonstrated a use-case in marine biomass, when identifying species of marine biomass for the real-world application of fish fraud detection. Previous works demonstrate that REIMS can be used to combat fraud and adulteration in food processing. This research aims to apply this method of analysis for determining the bulk composition and quality of marine biomass.

Heterogeneous marine biomass - Firstly, AIML techniques need to handle New Zealand's . To apply rapid mass spectrometry methods to fish processing in New Zealand, this research aims to tackle the unique market of New Zealand's seafood industry. Unlike other countries, take for example Canada or the United States, New Zealand has a high-variability in marine biomass. In layman's terms, when a catch comes in from a fishing vessel, there is a diverse range of species, in that catch. The catches coming from trawling vessels in Canada or the United states are consist mostly of one species - a homogeneous composition of marine biomass. However, the catches coming in from New Zealand vessels, consist of a diverse range of species - a heterogeneous composition of marine biomass. This translates to a multi-class problem with many classes in machine learning.

Low-sample complexity - Secondly, AIML techniques must address low-sample complexity of New Zealand's marine biomass. A factor unique to New Zealand's seafood industry, and due to our much smaller fishing fleet and population, is a low sample size. Large trawling vessels in international waters, or the United States or Canada, have a large volume of homogeneous marine biomass, to collect and analyze with chemistry methods. Due to New Zealand's smaller size, and isolated geographical location, there is a much smaller volume of fish to create datasets from for analysis via chemistry techniques. As demonstrated in previous works [12], fish analysis for New Zealand marine biomass, is performed on high-dimensional data with low sample complexity. Mutli-class problems with low sample complexity shelf the stock-standard toolkit of deep learning methods. With low sample complexity, DL methods risk overfitting by memorizing the training data, and not generalizing well on unseen data. The low-sample complexity requires an algorithm that is sample efficient, DL methods are often not sample efficient at all, as they (often) require thousands of samples to achieve reasonable performance at a given task. Furthermore, they don't produce interpretable models, that can be understood easily, by domain experts in chemistry and fish processing, nor can their results be verified and troubleshot in their real-world application, the Cyber-Marine flex-factory [4], for which they would be deployed.

Concept drift - Thirdly, AIML techniques require robustness to concept drift. A problem not unique to New Zealand marine biomass, and widely applicable to other marine biomass - is seasonal variation. The chemical composition of marine biomass changes dramatically, in periodic and reoccurring patterns, related to the behavior of those fish. Take for example, Hoki shown in fig. 1.3. When spawning, the female Hoki extracts (almost) all her lipids to give as nutrients to her eggs when spawning [8]. This dramatically changes the chemi-

cal composition, specifically the lipid profile, of the spawning adult female Hoki. A naive machine learning model, that is not robust to seasonal variation, may misclassify this adult female Hoki after spawning, not recognizing the mother as a Hoki, absent of her lipids. The phenomena of seasonal variation draws parallels to concept drift from data mining [22, 23]. Concept drift describes the shift in the distribution of data over time, in (possibly) reoccurring or periodic nature. To apply AIML techniques for fish processing in New Zealand, robust models that can handle concept drift such as seasonal variation, are needed, for real-world applications to deliver (any) commercial value. A robust model impervious to concept drift would be seasonal invariant.

1.2.2 Contamination Detection

This subsection describes existing works in contamination detection, and how this research can be extended with AIML techniques to automate fish processing in New Zealand. Contamination detection can be implemented in three stages, (1) *detection* - binary classification, (2) *analysis* - multi-label classification, (2) *quantification* - multi-output regression. Each stage extends the previous, to add another level of difficulty to the existing problem. The three stages can be grouped under the umbrella term qualitative contamination analysis.

Contamination Detection can detect contaminants such as cross-species - where two species of fish are mixed together in one sample, or Mineral Oil - where oil from the fishing vessels engine or factory machinery has spoilt the marine biomass. This method can identify potential hazards and/or quality control issues in fish processing. It can be used to ensure the Cyber-Marine flex-factory [4] is running optimally, and verify it can produce food grade products that are safe for human consumption [24], and whose labels are accurate to what should be inside that product [6].

Existing works each provide a piece of the puzzle, however, none address this research problem in its totality. The end goal of this research is to create a novel and effective method for contamination detection of marine biomass in the Cyber-Marine flex-factory [4], via rapid spectrometry. REIMS have proven an effective tool for identifying spoilage in livestock and seafood products [19, 19]. In response to scandals like the 2013 horse-meat scandal, [19] showed rapid spectrometry can identify adulteration of beef products with horse-meat offal, in concentrations as small as 1% (for certain offal). This work uses supervised learning techniques of PCA-LDA [25, 26] with thresholding to detect outliers. This tackles the problem of cross-species contamination detection, an area addressed in this proposal. Previous work of the same author, [18], uses the same technique for fish fraud detection, to find fish products that have been mislabelled. This is a binary classification task of fish species prediction, that employs the same supervised learning techniques of PCA-LDA, on REIMS data for fish. More advanced Artificial Intelligence Machine Learning techniques, such as Generative Adversarial Networks (GAN)s in the work [27], have been applied to anomaly detection in factory settings. These also rely on thresholding techniques to identify outliers, which indicate factory equipment that has likely malfunctioned.

Detection - The first task is a simple binary classification - given a sample, a positive class is contaminated, and a negative class is not. Take for example cross-species contamination, a sample with a mix of Hoki and Mackerel would be positive. A sample with only Hoki would be negative.

Analysis - The second task extends this to multi-label classification, each instance may be-

long to multiple classes. For example, a contaminated sample may contain, 30% Hoki and 70% Mackerel, this sample label would be [Hoki, Mackeral]. In contrast to the binary classification problem, the model would also have to be able to distinguish between [Hoki], [Mackerel], [Hoki, Mackeral]. For binary classification, both [Hoki], [Mackerel] would simply be negative, not contaminated. But for multi-label classification, both the individual species and their combination, are combined to give the label annotation.

Quantification - The third task extends the previous two tasks, to associate a percentage of contamination associated with each contaminant. Not only does it perform the previous two tasks, contamination detection, and contamination analysis, it then provides quantification to those contaminants that it has identified. Recall the example from earlier, with 30% Hoki and 70% Mackerel. Quantification would identify two classes present, and their relative contribution to the same, i.e. [Hoki - 30%, Mackeral - 70%]. This is a multi-output multi-label regression. Given a relative percentage of composition has to sum to 100%. The annotated label is similar to the softmax [28] function on an output layer of the neural network, which normalizes the output layer of that neural network to sum to 1, to fit a probability distribution.

1.2.3 Traceability

This subsection describes existing works in traceability detection, their limitations, and how those limitations can be extended with AIML techniques to automate fish processing in New Zealand. Traceability can be implemented in two different ways, one simpler and another more complex, (1) *detection* - a pair-wise comparison, and (2) *instance recognition*. Similar to before, each method extends the previous, adding complexity to the problem. These two sub-tasks are grouped under the umbrella term traceability, as they are concerned with determining, or tracing, the common origin of a sample.

Traceability deals with isolating a contaminated sample, to track the origin, and identify potential causes for that contamination. The previous subsection addressed contamination, once a contaminated sample is found, a factory seeks to find ALL other marine biomass that originated from the same origin, that same sample, to isolate and discard the potentially hazardous samples, that are likely not safe for human consumption [24]. To do so, traceability must identify the unique characters of an individual sample, not just the species or body part, but unique characters that distinguish ONE particular fish, from the others. In computer vision, this is referred to as instance recognition [29]. Existing works related to traceability can be found in computer vision, those include instance segmentation [30], instance identification [31], and contrastive learning [32, 33].

Detection - Siamese neural networks, proposed in 1993 by LeCun [32], and prominent today in deep learning for object detection and segmentation [34], and ransomware classification [33]. Siamese neural networks are a type of contrastive learning. Contrastive learning is a type of unsupervised learning where the goal is to learn a similarity metric between two inputs, by contrasting them with other inputs. Siamese networks [32] were originally developed for the task of signature verification. Given two signatures, an authentic signature known to belong to an individual, and the "query signature" whose veracity is being tested, determine if the query and reference signature were written by the same person. The model would predict if a signature is genuine or forged. The *detection* task in traceability is a simplification, a pair-wise comparison between two rapid spectrometry samples, to see if they originate from the individual fish. Although signature verification and sample attri-

bution are from different domains, the task is identical. Given two inputs, predict if they are the same. Siamese networks consist of two identification neural networks, sharing the same weights and architecture. Given a pair of signatures, a reference, and a query, one network takes the reference, the other network takes the query. The output of both networks is combined using a distance metric, to produce a similarity score. In the paper [32], the Euclidean distance was used to compute the distance between the two outputs. The score would indicate the similarity between the two signatures, the closer the Euclidean distance, the more likely the query was genuine. The greater the distance, the more likely the query was forged. A similar method concept to the thresholding method for detecting outliers, was previously mentioned in [18, 19]. However, contrastive learning is useful for few-shot learning for three reasons: (1) it allows efficient use of small amounts of labeled data, (2) it can lever labeled and unlabeled data to learn robust and discriminative representations of the data, improving the model's ability to generalize to new classes with only a few labeled examples, (3) by learning to contrast similar and dissimilar examples, the model can develop a rich understanding of the underlying structure of the data, which can further improve its ability to generalize to new classes with few labeled examples.

In the REIMS dataset, the sample complexity is very low. For the instance recognition task, the sample complexity is lower. As individual fish may only have a few (or single) labeled examples per instance. Therefore, instance recognition can be considered a few-shot, or in extreme cases a one-shot, learning problem. Traditional machine learning models often require large amounts of labeled data to achieve high performance, whereas few-shot learning specializes in training models to learn quickly from only a few examples. Due to low example complexity, and few-shot nature of the instance recognition ask, this research aims to amortize the training data, to allow for few-shot or even one-shot inference on unseen data. This amortization can be achieved through self-supervised contrastive learning, such as siamese networks, or transfer learning, where we share information between related tasks, to improve performance on new related tasks.

Instance recognition - The machine learning term for recognizing individuals that may belong to the same class is "instance recognition" [29] or "individual recognition". Instance recognition would involve recognizing each individual fish in the samples and assigning a unique identifier or label to each of them. This would allow the model to differentiate between individual fish even if they belong to the same species. Instance recognition is a type of object recognition task that goes beyond simply recognizing object classes and aims to identify each individual instance of an object class. It is commonly used in various fields such as wildlife monitoring, security surveillance, and biometrics. To avoid confusion, the term instance recognition from computer vision, is not to be confused with instance identification [31], or instance segmentation [30]. However, in the domain of fish processing, and chemical analysis via rapid mass spectrometry, the term sample attribution is fitting for the real-world application. Thus, for a Chemist and AI Researcher, the terms sample attribution and instance recognition can be considered equivalent and used interchangeably. Instance identification [31] and Instance identification [30] are both computer vision techniques used to identify objects and their instances in images. Although they are related, they have distinct differences in their objectives and output. Instance identification [30] is a computer vision task that involves identifying whether a particular object is present in an image and, if so, which class it belongs to. In this task, the goal is to identify all instances of a particular object class, but not to distinguish between different instances of the same class. The output of instance identification is a bounding box that encloses the object and a label that identifies its class. Instance segmentation [31], on the other hand, is a task that aims to identify and locate each individual instance of a class of objects in an image and assign a unique label to

each instance. This means that each pixel in the image is classified according to which object instance it belongs to. The output of instance segmentation is a mask that outlines each instance of the object class, making it possible to differentiate between multiple instances of the same object class in the same image.

This research deals with rapid spectrometry samples, not pixel images from computer vision, although those samples can be thought of as snapshots that capture the chemical composition of a fish. Given the REIMS is not pixel images, it is not concerned with drawing bounding boxes or segmenting the mass spectrograph into different classes. There may be an overlap in mass-to-charge ratio intensities where both species share a common chemical structure that contributes to the same peaks, so segmentation, or bounding boxes, are not appropriate nor applicable for rapid mass spectrometry. The research objective (2c) multi-label classification task for contamination detection is similar to instance identification - given a sample that may contain one or more species of fish (if contaminated), identify which species are present in the object, without distinguishing between different individual fish of the same species. The optional research objective (3c) multi-instance sample attribution, is similar to instance identification - given a sample contaminated with two fish of the same species, the model would learn to uniquely identify the two and differentiate between them both, despite both belonging to the same species.

Existing work for instance recognition task can be found in computer vision [29, 35, 36].

In [35], the authors propose HotSpotter a model to recognize instances based on their unique spots. This is a species invariant model, that differentiates between dissimilar species, e.g. zebras, giraffes, leopards, and lionfish. Fish and mammals are dissimilar but share spots. [35] was trained on a database with 1,000 images with five different species of animals, approximately 200 images per class.

While images are far from rapid mass spectrometry data, this research aims to perform a similar task, by providing a species-invariant model that differentiates between dissimilar species of fish, e.g. whitefish and oily fish, based on their unique chemical compositions.

The work of [36] proposed a multi-modal instance recognition that employees dense feature extraction on multi-modal features. This model is benchmarked on the Willow dataset from [37], which contains 37 views of 37 different objects (37 x 35 = 1295 images total) to be detected in a variety of tabletop scenes. Similar to [35], the classes are dissimilar objects, that must be uniquely identified from multiple observations of that same object. This work [36] was trained on a dataset of similar size to [35], with 1,295 and 1,000 images, respectively. Datasets with sufficicent sample complexity, i.e. more than 1,000 instances, are naturally suited towards deep learning methods that require large volumes of data. Traceability must perform the same task. To uniquely identify marine biomass from multiple observations. In [36, 35] the observations are images - a computer vision task. However, in this research the observations are Rapid Evaporative Ionisation Mass Spectrometry (REIMS) measurements - a chemical analysis task. Although, different fields a mass spectrograph and pixel image are similar, they share local connectivity in their multi-dimensional representation, where values close to each other are related, and their proximity to each other is information in itself.

In [29], they take instance recognition one step further than [36], with single-view instance recognition. They employ a general-to-specific training procedure, that pretrains the neural network on problems, of increasing granularity. The network is pretrained on a large multi-view dataset, then fine-tuned on a smaller single-view dataset. The neural network takes a feature embedding representation learnt from a general task, that can be transferred and applied to a more specific task. The largest multi-view dataset has 100 images of 124 objects (100 x 124 = 124,000 images total). The smallest single-view dataset has 1 view for

300 objects. In contrast to other works [36, 35], whose datasets contain more than 1,000 instances, this is a one-shot classification task with very few training instances. The REIMS dataset has 306 samples of marine biomass. Training on generalized tasks that are related, but where greater volumes of data are available, can improve the few or single-shot classification performance on datasets with low sample complexity (few training instances).

1.3 Limitations

> AI systems with human-competitive intelligence can pose profound risks to society and humanity, as shown by extensive research [38] and acknowledged by top AI labs [39] As stated in the widely-endorsed Asilomar AI Principles, Advanced AI could represent a profound change in the history of life on Earth, and should be planned for and managed with commensurate care and resources. Unfortunately, this level of planning and management is not happening, even though recent months have seen AI labs locked in an out-of-control race to develop and deploy ever more powerful digital minds that no one – not even their creators – can understand, predict, or reliably control [40]
>
> Max Tegmark
> Physicist, Cosmologist and AI Researcher.

This section addresses the limitations of the existing research, this proposal suggests ways to extend the existing works for real-world applications of Artificial Intelligence in fish processing. This section addresses identification, contamination detection, and traceability - the same areas in the previous section on motivations. However, here we address the limitations of existing work in each area. The remainder of this section addresses each limitation in further detail.

1.3.1 Identification

> Everything should be made as simple as possible, but not simpler [41].
>
> *Albert Einstein (attributed)*
> *Physicist, Nobel Prize (1921)*

The limitations of existing work in biomass analysis using rapid spectrometry is in there

application. [19] performs contamination detection by identifying outliers. However, it is not capable of analysis or quantification of the contamination it detects. [18] performs multi-class classification on a dataset with five species of white fish. It predicts one species, therefore is a single label prediction multi-class classification task.

The research is also only limited to white fish species. The model cannot handle cross-species contamination, where a sample contains biomass from one (or more) species of fish, or other types of marine biomass, such as oily fish, salmonid, shellfish, flatfish, tropical fish, freshwater, ... etc. Existing work [18] is geared towards homogeneous marine biomass typical of incoming trawling vessels in international waters, or countries like Canada, United States, or Norway [9].

In rapid spectrometry, there exists few dataset for New Zealand's unique marine biomass. NZ has a highly variable biomass. Therefore existing AI models, are not well suited for the niche NZ seafood industry [18]. This research uses datasets of chemical analysis of marine biomass, for a niche seafood market. There is limited (next to no) existing datasets for this specific task, to draw from to increase the sample complexity. In fact, that was the motivation for Cyber-Marine project [4] to produce these rapid spectrometry datasets in the first place. Due to availability and cost of the REIMS measurement, the domain expertise required to operate it, and the niche of its application to the seafood industry, the dataset contains few samples. Each sample is high-dimensional due to the fine-grain resolution of spectrometry methods. High-dimensional low-sample complexity data is typical of fish oil analysis using chemistry techniques [12]. There exists one paper that performs REIMS for marine biomass, however, this was not New Zealand.

For real-world AI in safety critical applications, e.g. human-grade seafood products, the biggest limitation is industry adoption. Industry is cautious to adopt new methods and techniques, as they introduce risk and uncertainty. Black-box AI techniques add to that uncertainty, by producing accurate models that work most of the time, but when they fail, their is no explanation or cause to diagnose. The explainability crisis nearly ground research on Large Language Model (LLM)s to halt. There was an open letter and petition to pause giant AI experiments for 6 months [40], with notable signatories in that letter [40], such as Yoshua Bengio [42], Stuart Russell [43], Elon Musk (SpaceX, Tesla) and Steven Wozniak (Microsoft co-founder). More recently, Geoffrey Hinton, the "Godfather of AI" [44] left Google AI, with fears the generative artificial intelligence arms race will cause real-world harm. The claims of real-world harm and danger surrounding LLMs derive from the black-box nature of these models. Humans cannot understand their process, from input to output, as the weight matrices of a +32 Billion parameter deep neural network. Recall the Arthur C. Clarke and note that a technology we don't understand is equivalent to magic. If industry are cautious about new technologies that are well understood, the black magic of neural networks [45], is going to be hard sell. Existing works, [18, 19] perform dimensionality reduction that obfuscates the meaning of its features. The results are accurate, but humans, more specifically the domain experts in bio-chemistry and fish processing, cannot understand and interpret the model.

1.3.2 Contamination Detection

The existing work in fish fraud detection [18], is limited as it is for multi-class classification only, it does not provide a model for cross-species contamination, where a single sample may contain fish from two (or more) species. [19] provides adulteration detection with thresholding techniques to identify outliers. However, no qualitative profiling of those outliers is given. Their technique does not say what the adulterant is, and to what concentration it is present. This is due to the nature of the thresholding technique for outlier detection.

The GANs used in [27], is advantageous over the PCA-LDA methods proposed in [18,

19], as they require less manual parameter tuning and domain expertise in the application, however, they produce black-box models, which can't be trusted or understood when deployed in fish processing.

Similar to GANs [27], the PCA-LDA [25, 26] used in [18, 19] produce feature embedding that are not interpretable either. Principal Component Analysis (PCA) is a dimensionality reduction technique that projects features from a high-dimensional space, into a lower-dimensional space. By projecting along the top k eigenvectors of the covariance matrix [42] The principal components are linear combinations of the original features, and their interpretation in relation to the original features is not straightforward. The PCA dimensionality reduction technique seeks to preserve the variance of the data, but the original semantic meaning of the features is lost.

Limitations of GANs [27] [CITATION NEEDED FOR EACH]:

1. Requires large amounts of data: GANs require large amounts of training data to learn meaningful representations. When the dataset is small, the generator may not learn a good representation of the data distribution, leading to poor-quality samples.

2. Mode collapse: In some cases, the generator can learn to produce only a small set of representative samples, which results in mode collapse. In other words, the generator fails to capture the full diversity of the training data.

3. Training instability: GANs are notoriously difficult to train because they involve training two networks, a generator and a discriminator, in a two-player minimax game. This can lead to problems such as mode collapse, where the generator produces a limited range of outputs, and oscillations in the loss function, which can make training unstable and unpredictable.

4. Limited diversity: GANs can generate high-quality samples that are similar to the training data, but they may not be very diverse. The generator may learn to generate only a few modes of the data distribution, resulting in a limited range of output samples.

5. Sensitive to hyperparameters: GANs are sensitive to the choice of hyperparameters such as learning rate, batch size, and architecture. Choosing the right hyperparameters can be time-consuming and require extensive experimentation.

6. Difficulty in evaluating performance: Evaluating the performance of a GAN is not straightforward. Metrics such as Inception Score and Frechet Inception Distance (FID) are commonly used, but they may not always capture the quality and diversity of the generated samples.

1.3.3 Traceability

Traceability, or instance recognition, is the most difficult task proposed in this research. It is a few-shot, or in the extreme one-shot learning task on a dataset with low sample complexity. Therefore, with limited training instances, this work aims to maximize the knowledge that can be extracted from each instance - the amortization of data. Another way achieve amortization is through transfer learning. To learn on a more general task, and transfer that knowledge a more specific task through fine-tuning, as seen in [29] for single-view instance recognition.

Alternatively to pre-training, siamese networks [32, 34, 33] is another effective technique for few-shot contrastive learning. Contrastive learning is a useful technique for low sample complexity datasets, were sample efficiency is critical. However, this method is limited by

1. Limited to fixed-length inputs: Siamese networks typically require fixed-length input vectors, which can limit their ability to handle inputs of varying lengths or sizes.
    - TODO [x] Is REIMS a fixed-length representation, i.e. a set number of mass-to-charge ratio measurements?
2. May require large amounts of training data: Siamese networks can require a large amount of training data to learn effective similarity metrics, especially in high-dimensional spaces where the number of potential comparisons can be very large.
    - With 1024 features, relatively high-dimensional dataset, the Siamese network may require a high sample complexity to learn to fit the training data.
3. Limited to pairwise comparisons: Siamese networks are designed to compare pairs of inputs and are not well-suited for multi-class classification tasks, where there are more than two classes to choose from.
    - Not well suited towards the multi-instance recognition task proposed in the optional research objective 3c.
4. Sensitive to hyperparameters: The performance of a siamese network can be sensitive to hyperparameters such as the number of layers, the size of the hidden layers, and the choice of activation functions, which can require extensive tuning.
5. Can be computationally expensive: Siamese networks typically require the computation of pairwise similarity scores for all pairs of inputs, which can be computationally expensive and scale poorly with the number of inputs.

Existing work on traceability is limited to the related task of instance recognition from computer vision. Works [35, 36, 29] show applications of instance recognition for dissmilar classes, [29] extends this for one-shot instance recognition where sample complexity is low.

Both [35] [36] require large sample complexity - a high volume of training data. Many deep learning and traditional machine learning methods require many training instances to achieve high quality performance. The REIMS dataset in this research is low sample complexity, with only 306 training instances. These deep learning and traditional machine learning techniques will not work out-of-the-box on a dataset with low sample complexity. In the extreme cases, there may likely not be sufficient data for these models to fit data. It is more likely that these models will overfit the training data, and fail to generalize on unseen data. However, it is possible to using pre-training (or transfer learning) to allow for few or one-shot learning on a dataset with low sample complexity, as shown in [29].

The limitations of [35], are being a relatively dated paper, 2012 paper [46] that proposed Local Naive Bayes Nearest Neighbours (LNBNN), an extension of Naive Bayes Nearest Neighbours (NBNN) [47], where "only the classes represented in the local neighborhood of a descriptor contribute significantly and reliably to their posterior probability estimates".

The authors admit LNBNN, did not beat state-of-the-art methods such as feature pyramid networks [48], which rely on local soft assignment and max pooling operators. Convolutions and max-pooling are utilized in Convolutional Neural Networks (CNN)s [49], a powerful model for computer vision-related tasks. Which with advancements in hardware, and the lifting of the AI winter, are efficient to train at scale using GPUs. Since then, a a plethor of Convolutional Neural Networks (CNN)-based architectures dominate computer-vision tasks, such as LeNet [49, 50, 51, 52], AlexNet [53], VGG-16 [54], GoogLeNet [55], ResNet [56].

1.4 Research Goals

This proposal is application-oriented, it aims to implement a real-time (online) fish contamination detection and identification algorithm(s). This is a supervised machine learning task operating on Rapid Evaporative Ionisation Mass Spectrometry (REIMS) [57] fish oil data. Types of contamination include cross-species and mineral oil. Specifically, this proposal outlines the need for algorithms to perform the following tasks summarized in fig. 1.1.

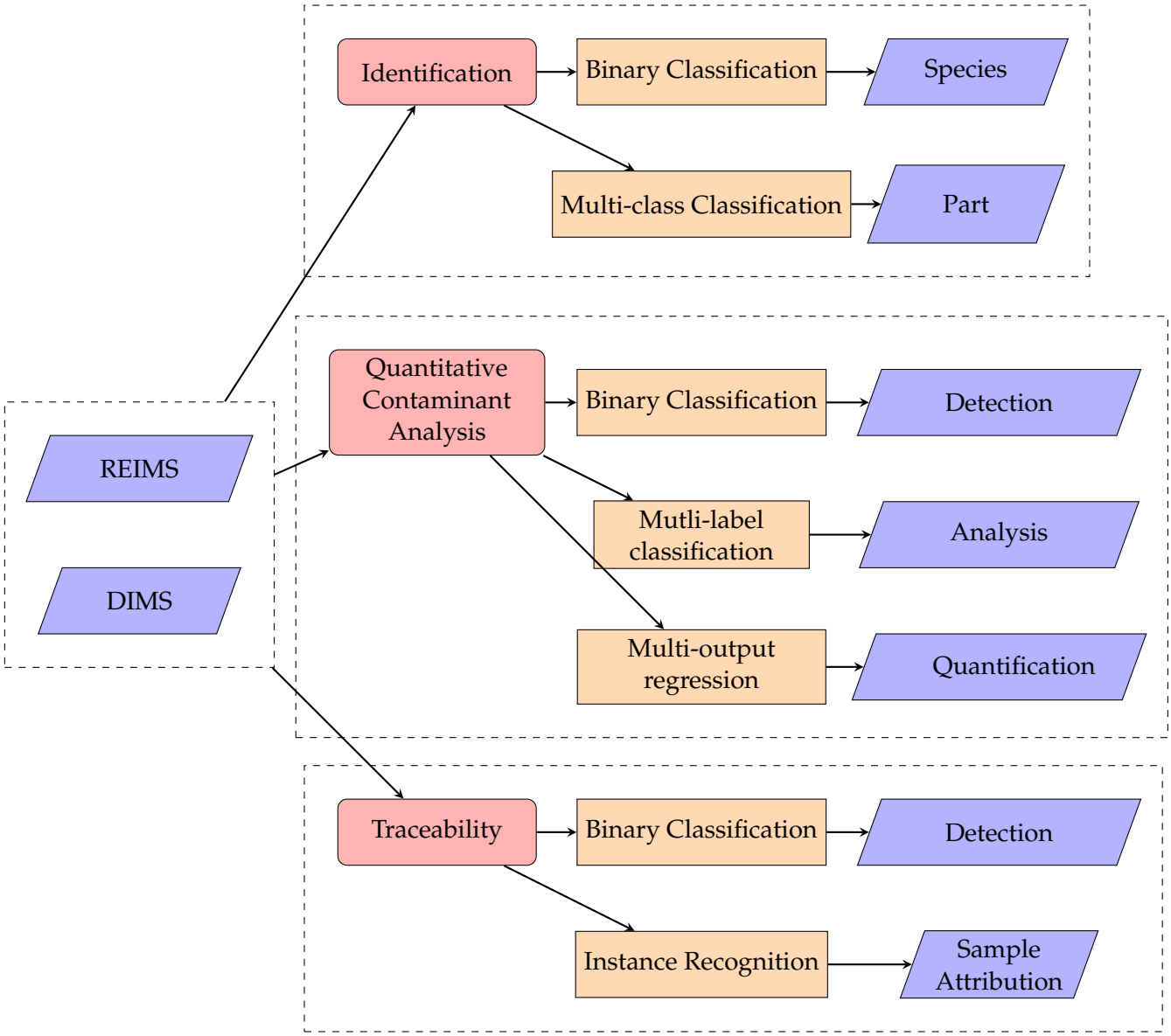


Figure 1.1: Research Goals

Starting from the left, the chart shows two mass spectrometry datasets, REIMS and DIMS. These datasets are used for 3 tasks, each delimited by a dotted box. Each task can be broken down into machine-learning techniques, and downstream applications in the fish processing domain. The tasks, and their retrospective sub-tasks, are given in ascending order of assumed difficulty, top-to-bottom from easy to hard.

This thesis proposes three research objectives, which are equivalent to tasks, each task is composed of sub-tasks. For the remainder of this section, **tasks** are given in **bold**, and *subs-tasks* in *italics*. An itemized summary of those tasks and sub-tasks, is given here:

1. **Identification**
   (a) *Binary Classification*
   (b) *Instance Detection*
2. **Qualitative Contaminant Anaylsis**
   (a) *Binary Classification*
   (b) *Multi-label classification*
   (c) *Multi-output regression*
3. **Traceability**
   (a) *Detection*
   (b) *Sample Attribution*
   (c) *Multi-instance Sample Attribution* (Optional)

A brief explanation of the research objectives given above and diagram - including the tasks, and sub-tasks, inputs and outputs - is provided here:

inputs. The inputs to the pipeline are labeled "REIMS" and "DIMS," and they are enclosed in a dashed box. The arrows represent the flow of data through the pipeline, with each arrow pointing from the output of one node to the input of another.

tasks. 1. **Identification** - The first task has two subtasks: *Binary Classification* and *Multi-class Classification*. The Binary Classification subtask has an output labeled "Species," while the Multi-class Classification subtask has an output labeled "Part."

2. **Qualitative Contaminant Anaylsis** - The second task has three subtasks: *Binary Classification*, *Multi-label Classification*, and *Multi-output Regression*. The Binary Classification subtask has an output labeled "Detection," the Multi-label Classification subtask has an output labeled "Analysis," and the Multi-output Regression subtask has an output labeled "Quantification."

3. **Traceability** - The third task has two subtasks: *Binary Classification* and *Instance Recognition*. The Binary Classification subtask has an output labeled "Detection," while the Instance Recognition subtask has an output labeled "Sample Attribution."

The remainder of this section elaborates on concepts given both above and in fig. 1.1. Each task is clearly defined with respect to chemistry, fish processing and machine learning. This section is excellent reference material for downstream applications of machine learning in fish processing presented in chapter 3.

1.4.1 Identification

Identification is the process of identifying characteristics of a sample. In particular, given a fish tissue sample, identification has two sub-tasks, (1) predict the *species* of the fish, a binary classification task, and (2) predict the *part* of the fish, a multi-class classification task

Species - the first sub-task is concerned with predicting the species of the fish. For the mass spectrometry datasets provided, there are two species of fish, Hoki and Mackerel shown in fig. 1.3. This sets up a binary classification problem, to predict the species of fish. From the figure, a human could easily make the distinction between species by eye. The eager AI researcher may construct this problem as a computer vision task. However, in a factory setting, once fish is gutted, filleted, minced or otherwise processed, the samples become a homogeneous blend of marine biomass. No longer a trivial computer vision task, more complex methods of rapid Mass-Spectrometry are used to determine the characteristics of that marine biomass.

Figure 1.2: Hoki *Macruronus novaezelandiae*

Figure 1.3: Mackerel *Trachurus symmetricus*

Part - the second sub-task is to predict the part (or tissue) where the measurement was taken. For these mass spectrometry datasets, there are six different classes of fish parts, e.g. fillet, liver, skin, guts, frame heads. This is trickier than the species identification, because we have more than two classes. This constructs a multi-class classification problem, where a model predicts the fish part given many classes. Previous work [12] on chemistry datasets has also demonstrated that classifying fish parts proves more difficult than species. This is perhaps because their are more differences between to difference species, than there are similarities between the same part, but from different species. To put it in computer vision terms, there is more intraclass variation in part, than species.

This work seeks to resolve the tackle both these limitations - to offer models that (1) require little domain expertise or manual hyper-parameter tuning, and (2) produce models that are explainable and can be understood, trusted, and verified by those operating them in the flex-factory.

This work seeks to address the interpretability limitations of existing works, by produc-

ing accurate models that can be understood by domain experts in the application domain of biochemistry and fish processing. The feature embedding of the models will be "mass-spectrometrically" [sic] meaningful - semantically meaningful in their application domain. Models that preserve interpretability, and explainable AI, can be verified and troubleshot easily, and provide new insights and knowledge to those domain experts using them. With explainable AI tools and domain experts using them, the AI doesn't aim to automate or replace their job, it helps aid their understanding to enhance their ability. AI tools that can be understood by the domain experts, can be trusted and relied on, which is critical for industry adoption.

This work extends [18] by generating a model that can differentiate between white fish (e.g. Hoki) and oily fish (e.g. Mackerel). Different types of fish have different chemical compositions. These different chemical compositions each require an accurate representation. The AI model must distinguish between different types of fish with dissimilar chemical compositions. Not relatively homogeneous chemical compositions of similar whitefish, as in [18]. The New Zealand seafood industry offers a unique problem, the heterogeneous marine biomass of incoming marine biomass. This work seeks to address this niche, by creating models that are robust to dissimilar, multi-class and variable marine biomass. The models will handle many species of marine biomass, that are dissimilar (heterogeneous). Other types of fish, such as salmonid, shellfish and freshwater are excluded, as salmonid and shellfish typically belong to aquaculture [5]. fresh-water fish are caught in lakes and rivers, not the salt-water trawling vessels that are the scope of the Cyber-Marine flex-factory [4]. These types of fish can be excluded from the scope of the research, as they come from different sources entirely, and don't require sorting via rapid mass spectrometry analysis.

Challenges to address:

1. High dimensional data
2. Low example complexity
3. Variance
    - Extreme variance in REIMS data
    - Extreme variance in Hoki lipid profiles
    - In extreme cases, there is more variance in individual Hoki, than between Hoki and Mackerel
4. Seasonal variation $\approx$ (reoccurring) concept drift

1.4.2 Quantitative Contaminant Analysis

Qualitative Contaminant Anaylsis (QCA) is concerned with spoiled fish products. Identifying spoilage in fish products is needed for Quality Assurance in the flex-factory. QCA involves defining contaminants at three levels of granularity. These sub-tasks are given in ascending order of difficulty (top-to-bottom from easy to hard), (1) *Detection*, (2) *Analysis*, and (3) *Quantification*.

Detection - the first sub-task identifies if a samples are contaminated. Contamination detection constructs a binary classification task, that predicts true of false for if a sample is contaminated. Detection is unaware of which contaminants are present, and be thought of as a red flag, that warrants further investigation. In a factory setting, the detection task may

provide higher recall than other models. An accurate contamination detection model can be used to identify areas of concern where future investigation may be warranted.

Analysis - the second sub-task identifies which contaminants are present. Contamination analysis tasks detection on step further. Not only does it identifying samples that are contaminated, analysis says which contaminants are present in that sample. For the Mass-Spectrometry datasets given, there are two forms of contamination, mineral oil and cross-species. Either an individual model can be trained for each task, or an overarching model which analyzes both. Take for example analysis for cross-species contamination,

Quantification - the third and hardest sub-task is quantification. This performs the first two sub-tasks implicitly, with the added difficulty of providing a percentage ratio for those contamination. Contamination quantification constructs a multi-output regression problem, that predicts the contaminants present, and their respective percentage that contaminant contributes to the composition of the sample. Take for example quantification for cross-species contamination. Consider a contaminated sample with a mixture of both Hoki and Mackerel fish species. Quantification would tell you what percentage of that contaminated sample is Hoki, and what percentage is Mackerel. Alternatively, for Mineral Oil contamination, quantification would predict what percentage of the sample is Mineral Oil, and the rest fish. This task is similar to a softmax activation [28] as the final layer of a neural network. For a mutli-class problem with each class indexed in a vector, the value at each index would correspond to the probability of that class, and the vector represents are probability distribution (which sums to 1). The quantification problem assumes the composition of known classes makes up the entire sample. For example, for cross-species contamination, if quantification predicts 33% Hoki, the remaining $(100\% - 33\%) = 67\%$ would be Mackerel.

This work seeks to address the limitations of the existing work [18], by extending the task to handle cross-species contamination - a multi-label mutli-class classification. That is to (1) detect the presence of contamination, i.e. one (or more) species present, (2) analyze (or identify) present species, (3) quantify their relative percentage, in a sample.

Challenges to address:

1. High dimensional data
2. Low sample complexity
3. Class imbalance
4. Extreme variance in REIMS data
5. Extreme variance in Hoki lipid profiles
    - In extreme cases, there is more variance in individual Hoki, than between Hoki and Mackerel
6. Robust classifiers

1.4.3 Traceability

Traceability is concerned with tracking measurements back to the sample the originate from. This can be broken down into two sub-tasks, (1) *Detection* and (2) *Sample attribution*.

Detection - The first sub-task is to detect if two fish are the same. This is a simple heuristic that can be used to trace contamination to other samples, once a contaminated sample has been identified. In a fish processing factory, samples close to detected source of contamination can be tested to see if they originate from the sample. Previous sections compare this task to signature verification from Siamese neural networks [32], where a reference signature is compared to a query, to see if they match, i.e. a genuine or forged signature. The task is similar to Next Sentence Prediction (NSP) from bert [58], which is given two sentences, and predicts if those two sentences were adjacent. Instead, here we have two samples, and detect if they come from the same fish. This constructs a pair-wise binary classification problem, given two samples, predict if they came from the same fish. In layman's terms "same fish detection". But in more technical terms, same sample detection, to account for the possibility super-set of marine biomass, that includes more than just fish. This is similar to contamination detection from task 2 sub-task 1.

Sample attribution - The second sub-task is an extension of detection. Sample attribution is not just interested in if two samples are from the same fish. It is concerned in keeping track of those individual fish too. Sample attribution would take a batch of fish samples, and could identify and isolate the individual fish present in that batch. This is a similar concept to semantic instance segmentation from computer vision. Where both are not only interested in distinguishing between classes, but look to uniquely identify individual instances. Sample attribution is likely the most difficult task proposed in this research, as it requires a very sensitive model to learn individual fish from very-few shot learning. It must account for seasonal variation and distinguish between different species of fish, where each species has different variance. Initial findings from PFR [2] showed there was more variation in individual Hoki samples taken from the same fish, than there was from different Mackerel.

(Optional) *Multi-sample attribution* - A more complex form of this problem is multi-instance recognition, where a sample can contain multiple unique instances, for which a classifier model must identify each unique individual present. This task could be formed with the cross-species or Mineral Oil contamination data, from the previous research objective. This research objective is marked as (Optional), should time allow, this is an interesting direction to explore. Alternatively, this could make an interesting field for future research.

This work seeks to address the explainable limitations of existing work [18, 19] by focusing on transparent, explainable and semantically meaningful models [15, 14, 16]. For industry adoption and application of real-world AI, models are needed that enhance the existing knowledge of domain experts, and assist them in their role. Rather than existing works [18, 19], which get the same answers (most of the time), but refuse to show their working in a way that can be understood by humans. Transparency and explainablity are required to be trusted and used in real-world application.

Challenges to address:

- High dimensional data
- Low sample complexity
- Species invariant
- Few-shot (even one-shot) learning
- Instance recognition - very hard!

1.5 Summary

Limitations	No profiles for NZ biomass	Need tools to detect contamination	Hard to isolate samp
Motivations	Profiling NZ marine biomass	Detect contamination	Isolate individual san
Research Goals	Identification	Quantitative Contaminant Analysis	Traceability

Table 1.1: Limitations, motivations and research goals.

In table 1.1, the information presented in the introduction chapter is summarized in tabular form. The table gives the limitations, motivations, and research goals, each grouped respectively. This table provides coherence, linking problems in the real-world application domain, to the research goals given in this proposal. This is closely related to the fig. 1.1, which explores those research goals in further detail.

The final section of the introduction provides an overview of the organization for the remainder of the proposal.

1.6 Organisation of the Proposal

This proposal is divided into four main chapters, (1) **Introduction** (this section), (2) **Literature Review**, (3) **Preliminary Work**, and (4) **Contributions and Project Plan**. Each chapter and section provides a brief description of its contents (like this one here) for clarity.

Readers can acquaint themselves with each chapter and its contents, and read to their level of expertise and interest. A brief summary of the chapter titles given above is provided here for clarity. The first chapter, the *Introduction*, gives the scope of the problem and the solutions proposed in this work. The second chapter, *Literature Review*, outlines existing work in the field and its limitations. The third chapter, *Preliminary Work*, covers automated fish oil analysis and exploratory data analysis. The final chapter, *Contributions and Project Plan* provides an outline for the thesis and its execution.

Please see the table of contents for a more detailed breakdown of the contents of this proposal. This document is structured with the suggestions in [59], and with inspiration from the layout of the very usable textbook and guide to user experience [11].

Chapter 2

Literature Review

This chapter focuses on outlining the existing work in this field. This includes work in the disciplines of chemistry, fish processing and machine learning. This thesis is application-driven, so it focuses on the intersection of those disciplines, and how knowledge can be transformed into innovation, to transform fish processing with artificial intelligence. This chapter, outlines marine biomass, chemistry, machine learning, and their limitations. Finally, the chapter concludes with a summary, which positions this thesis, as a potential step to address these limitations. Specifically, the sections of this chapter are:

1. **Marine Biomass**
2. **Mass Spectrometry**
3. **Machine Learning**
4. **Evolutionary Computation**

The following sections will be examined in more depth in the remainder of this chapter.

2.1 Marine Biomass

This covers marine biomass - a fancy word for fish (see glossary for disambiguation) - that is used to describe the incoming raw biological materials that enter the flex-factory. It is important to note the variability of this biomass, fish wastage is likely to contain a mix of fish species, body parts, and (potentially) contaminants. Even within a particular given species of fish, the measurements given by chemistry techniques are susceptible to seasonal variation in the composition of those fish. This section covers the variability of incoming marine biomass, contamination/adulteration, and seasonal variation in marine biomass.

Marine biomass has seasonal variation - the chemical composition as measured by mass spectrometry changes dramatically between seasons. The seasons, caused by Earth's 23$\circ$ tilt [60], cause a reoccurring change in the temperature, sunlight and nutrient availability. This has a significant impact on diets of fish, in the types and quantities of food they consume. Migration and reproductive behaviour also alter fish chemical composition on regular intervals.

Take for example Hoki a common New Zealand whitefish. In the process of spawning, where fish produce offspring, the females lay eggs and males fertilize. When the Hoki produce their eggs, the female extract many of their own lipids, and put them into their eggs. The spawned female is spent after this process, and her chemical composition has changed dramatically [8], with a noticeable lack of lipids.

An AIML model for species prediction of Hoki would need to account for this. Robust models would be able to identify all Hoki species, regardless of seasonal variation, what is called seasonal invariant. A more complex model for instance recognition, would perform tasks two fold, identify the species as Hoki, and use the seasonal variation as a potential marker for an individual. Seasonal variation is closely related to conceptual drift from data stream mining [61, 22]. Concept drift occurs when the underlying distribution of the data changes significantly, e.g. the spawning Hoki lipid profile. Reoccurring concept drift is where those distribution shifts occur on a regular and predictable pattern. Drift detection algorithms [62, 23] can be used to detect reoccuring conceptual drift, and identify seasonal variation in marine biomass. A flexible system could detect seasonal variation in marine biomass, and then decide which model is best.

2.2 Mass Spectrometry

This work focuses on two state-of-the-art chemistry techniques,

1. **Rapid Evaporative Ionisation Mass Spectrometry (REIMS)** [57]
2. **Direct Infusion Mass Spectrometry (DIMS)**

These are two of the most powerful analytical tools for Mass-Spectrometry. These tools are very expensive, but as prices decrease they may be affordable for deployment in a marine biomass processing facility. REIMS [57] has shown promise in beef processing, where it was able to detect horse meat contamination in beef [19]. Most impressively, horse meat contamination was detected at ¡INSERT STATISTICS FROM PAPER HERE¿ very low levels. This demonstrates the REIMS technique is incredibly sensitive to contamination. REIMS has been applied to fish fraud detection to identify fish species and identify catch methods for fish products. The method was so accurate it was able to identify incorrectly labelled instances in the training data. However, it has not been applied to Adulteration detection and identification in marine biomass. This work applied machine learning algoritms to REIMS data for the tasks of fish species and part identification, cross-species / mineral oil contamination, identify QC parameters, and individual identification. The research shall compare the results from REIMS to DIMS - the direct infusion of lipid extracts from the marine biomass samples. DIMS is much slower than REIMS, but provides high-resolution measurements as a qualitative benchmark.

Many alternative state-of-the-art chemistry techniques could be considered for the task. The alternative chemistry techniques that could be considered were:

- **Light-based** - One approach is to use analytical techniques based on light e.g. UV or fluorescence spectrophotometry, or vibrational spectroscopy (infrared, near-infrared or Raman spectroscopies). These techniques have been applied in combination with Genetic Programming to nutrient assessment in horticultural products [63, 64].
- **DNA Sequencing** - is limited due to extremely low sample size, and very high-dimensional data, e.g. the average human genome contains 3 billion base pairs and 30,000 genes. The dimensionality, and consequently the computation required to process it, rules out genomics data for real-time fish contamination detection. DNA identification methods were examined in a meta-analysis which revealed an average mislabelling rate of 30% in seafood processing [6]. DNA methods are limited, as they only differentiate between species, and are not useful for determining different body parts from the same species, or non-organic matter (e.g. engine oil) [18].

- **Gas-Chromatography Mass-Spectrometry** - Previous work [12] demonstrated that Gas-Chromatography Mass-Spectrometry (GC-MS) can identify fish species with high accuracy. However, GC-MS techniques significant time and domain expertise is required to prepare and analyze samples. This is not applicable for real-time fish contamination detection.

2.3 Machine Learning

This subsection will address the existing literature on fish analysis for REIMS data. This section introduces each paper, then identifies the limitations, and how this proposal intends to address those.

In [18], REIMS data modeled with PCA-LDA was able to detect species and catch method. Cross-species contamination is a more complex variation of this problem. In [18], each sample belonged to one species, however, for this problem, each sample can belong to multiple classes, e.g. a mix-species contaminated sample contains a mixture of two species. [19] performed detection and identification beef adulteration. It can identify samples that are adultered with offal, and specify which offal was present.

- Instance recognition [35]
  - HotSpotter — Patterned species instance recognition
  - Non-species specific, dimissilar classes: dataset → Grevy's and plains zebras, giraffes, leopards, and lionfish
  - Two methods for extracting and matching keypoints or "hotspots"
    1. (1) The first tests each new query image sequentially against each database image, generating a score for each database image in isolation, and ranking the results.
    2. (2) The second, building on recent techniques for instance recognition, matches the query image against the database using a fast nearest neighbor search
  - Method for (2): It uses a competitive scoring mechanism derived from the Local Naive Bayes Nearest Neighbor algorithm recently proposed for category recognition.
  - Results: We demonstrate results on databases of more than 1000 images, producing more accurate matches than published methods and matching each query image in just a few seconds.
- Deep Learning
  1. LeNet, developed and presented in [49, 50, 51, 52]: This is one of the earliest CNN models and was introduced by Yann LeCun et al. in 1998. It consists of seven layers and was primarily used for handwritten digit recognition.
  2. AlexNet [53]: This is a groundbreaking CNN model introduced by Alex Krizhevsky et al. in 2012. It won the ImageNet Large Scale Visual Recognition Challenge (ILSVRC) by a large margin, and its success spurred the development of deeper CNN architectures. AlexNet consists of eight layers, including five convolutional layers and three fully connected layers.
  3. VGG-16 and VGG-19 [54]: These are deep CNN models introduced by Karen Simonyan and Andrew Zisserman in 2014. They consist of 16 and 19 layers, respectively, and have a uniform architecture that only uses 3x3 convolutional filters.

VGG-16 and VGG-19 achieved excellent performance on the ImageNet dataset and helped establish the importance of deeper CNN architectures.

4. GoogLeNet [55] (Inception v1): This is a deep CNN model introduced by Christian Szegedy et al. in 2014. It consists of 22 layers and uses a novel architecture called the Inception module, which allows for efficient use of computing resources by using multiple filter sizes at each layer. GoogLeNet was the winner of the ILSVRC 2014 challenge.
5. ResNet [56]: This is a family of deep CNN models introduced by Kaiming He et al. in 2015. The original ResNet model (ResNet-50) consists of 50 layers, and the deeper versions (ResNet-101, ResNet-152) have up to 152 layers. ResNet models use a residual learning framework that allows for training of very deep networks by adding shortcut connections between layers. ResNet achieved state-of-the-art performance on several computer vision tasks, including ImageNet classification and COCO object detection.

2.4 Evolutionary Computation

> Biologists, too, use models to express what they think is going on inside organisms and in ecosystems. But I want to say something altogether more radical. An animal is a model. Any organism is a model of the world in which it lives. One way to understand this is to imagine a zoologist presented with the body of an animal she has never seen before. If allowed to examine and dissect the body in sufficient detail, a good zoologist should be able to reconstruct almost everything about the world in which the animal lived. To be more precise, she would be reconstructing the worlds in which the animal's ancestors lived [65]
>
> *Richard Dawkins*
> *Evolutionary Biologist*

Evolutionary Computation (EC) borrows concepts from biology. Specifically, population-based evolutionary search strategies that utilize Darwin's principle of survival of the fittest that he proposed in his work [66], originally published in 1859. More recently, evolutionary biologist Richard Dawkin's expanded that idea, in 1976 he proposed memes, cultural propagation of ideas, in his seminal work The Selfish Gene [67]. In later work from 1996, Dawkins proposed the evolved imagination, where every organism is a microcosm of it's environment [65]. This is the bread and butter of EC, where each individual is a candidate solution, an approximation of domain-specific task being solved, a model of the world. Dawkins argued that by examining an individual organism, one could deduce the characteristics of its

environment. Dawkins gives examples to support his argument presented in the epigraph to this section,

> "By reading the animal's feet and its eyes and other sense organs, the zoologist should be able to tell how it found its food. By reading its stripes or flashes, its horns, antlers, or crests, she should be able to tell something about its social and sex life." [65]

This draws parallels to computer science, take an Artificial Intelligence Researcher presented with an accurate and explainable AI model representation. If they have sufficient domain expertise in the application, and understanding of the model, a good AI researcher should be able to reconstruct knowledge about the application domain, and potentially produce novel insights. More recently in deep learning, prominent AI Researchers, Schmidhuber [68] and LeCun [69] have argued strong AI require an explicit world model [70].

However, unlike those deep learning approaches, EC offers AI models with explainable representations. In biology, the terms genotype and phenotype, refer to the genetic make-up (or DNA), and the expression of those genes, respectively. Take for example a child, with a single recessive gene for ginger hair - the genotype, with a brown hair colour - the phenotype. EC borrows these concepts, where genotype refers to the representation of the model, e.g. a tree, vector, neural net, and the phenotype refers to its evaluation that representation, e.g. a classification label, a regression output, a one-hot encoded vector. In previous work [12], the EC technique of Particle Swarm Optimisation (PSO) [71] was used for feature selection in fish species and part identification. In the following chapter on preliminary work, for that same task EC techniques of Single-Tree Genetic Programming (ST-GP) [72] and Multi-Tree Genetic Programming (MT-GP) [73, 74] are used for feature construction and classification.

2.5 Limitations

This proposal seeks to address the limitations of the existing literature that will be resolved in the thesis. In particular, those limitations are:

1. **Domain knowledge**
2. **No state-of-the-art techniques**
3. **No transfer learning/pre-training/synthetic data**
4. **Online learning**
5. **No taxonomy (lost in translation)**

The remainder of this section addresses each of those limitations in more detail.

2.5.1 Domain Knowledge

The thresholds to determine outliers are determined manually by domain experts. Their expertise in chemistry is needed to choose hyperparameters for every model - time. Significant markers are analysed and identified post hoc, relying on domain expertise in chemistry and human intuition.

Hyperparamters are parameters whose whose value is used to control the learning process. Take for example a K-Nearest Neighbours (KL) [75]. The KL model has hyperparameter k, this controls the tradeof between bias and variance. k determines the number of nearest

neighbours the model will consult to make a prediction. When k is low, the model has low bias and high variance, a low-k model is very sensitive to outliers and noise. Conversely, when k is high, the model has high bias, and low variance, a high-k model is robust to noise and outliers, but susceptible to underfitting - where it fails to capture complex patterns in the data.

For more complex models, a typical neural network has hyperparameters that correspond to the architecture and behaviour of that network, e.g. learning rate, number of hidden layers, neurons per layer, activation function, batch size, epochs, dropout, regularization, optimizer. Ultimately, these hyperparameters are nuisance variables, that must be decided upon before evaluation, with what usually amounts to combination brute-force search, human-crafted rules of thumb, and esoteric deep learning domain expertise. Criticism of is often levelled at "deep learning theory" (or the lack thereof), with comparisons Arcane rituals or black magic, so much so that [45] coined the term "grad student descent" - this describes the non-theory driven manual brute-force exploration of the hyper-parameter space by postgraduates.

Previous work in the REIMS literature suffers from this same critique. Hyperparamaters such as the number of principal componets, the Relative Standard Deviation (RSD) threshold for outliers, the mass range for Mass-Spectrometry in [18, 19] seem to be chosen rather arbitrarily by humans. An automated model which pro-grammatically searche the hyperparameter space for ideal configurations for these variables. Or models could be chosen that don't need those hyperparamters at all! This research aims to automate exploration of the hyper-parameter space through intelligent heuristics, as opposed to handcrafted rules-of-thumb discovered via trial-and-error. This reduces the need for domain expertise in chemistry to design models and avoids falling into the same pitfalls of previous work.

2.5.2 State-of-the-art ML

Mature statistical techniques are used for dimensionality reduction and classification, not state-of-the-art machine learning.

- Basic dimensionality reduction techniques (e.g. Principal Component Analysis (PCA) [25]) were used.
  - PCA [25] Project data along the principal components, the axis of maximum variance in descending order.
  - The first principal component is the axis of maximum variance, the second principal component is orthogonal to the first and has the second largest variance, and so on.
  - This method does not take into consideration feature interactions, interactions with the class labels, and feature redundancy/relevance.
  - Future work should consider T-distributed stochastic neighbor embedding (t-SNE) [76], Uniform Manifold Approximation and Projection for Dimension Reduction (UMAP) [77]
    - t-SNE [76]
      1. it creates a probability distribution of the similarity between points in the high-dimensional space.
      2. it defines a similar probability distribution over points in the low dimensional space.
      3. Then minimizes the Kullback-Leibler (KL) divergence [78] between the two distributions.

- Basic supervised statistical models (e.g. LDA, OPLS-DA) was used for classification. Future work should consider CNNs [79, 80], GANs [27], Diffusion [81, 82]
  - Denoising Diffusion Probabilistic Models (DDPM) [81], the original diffusion paper, behind diffusion-based image generation models.
  - Denoising Diffusion Implicit Models (DDIM) [82], a generalized DDPM that is faster and deterministic.
  - Genetic Programming for classification [83], feature construction [73, 74], feature selection

2.5.3 Transfer Learning

There is a large body of existing Mass-Spectrometry data. Knowledge from these datasets is not incorporated.

- Potential for transfer learning (incorporate previously existing data) to improve performance for few-shot classification tasks.
  - Due to manual labour, cost of machinery, domain expertise and high-resolution datasets, REIMS datasets have low sample complexity and high dimensionality.
- Semi-supervised / unsupervised
  - Unsupervised learning techniques have utilized unlabelled data from the same distribution to improve classification accuracy. The REIMS dataset contains Quality Control (QC) samples. These don't belong to any class (?) and are used to calibrate/tune the machine, unlabelled instances drawn from the same distribution. Zemma et al. [84] incorporated unlabelled instances to draw more accurate support vectors and improve the classification accuracy for breast cancer diagnosis with SVM.
  - METLIN metabolites database, and LIPID MAPS can provide annotated labels for spectra [19].
  - This looks like that (R-CNN) [85], give annotated labels for lipids used to make a classification/regression decision (significant markers $\approx$ important features).
  - The key behind utilizing semi-supervised / unsupervised techniques, is that they require no/little human supervision, and can improve accuracy of our models. These adjacent tasks provide a free performance boost, by utilizing information from unlabelled training data from the same domain, e.g. mass spectrometry.
- Pre-training
  - Transformer revolution, Bert [58]
  - Attention is all you need! [86]
  - While, deep learning models are not appropriate for low sample complexity and explainable AI.
  - Key lesson, when human supversied label annotated data is expensive, as is the case with chemistry datasets.
  - Take advantage of semi-supervised learning, let the models train themselves, pre-training [87].

  - Task-specific conditional learning [88], Tasks are often specified in text, a task-specific language model, conditioned to perform certain tasks. Learn arbitrary tasks, simply by being trained on a very large corpus.
  - Language models are few-shot learners [89], in-context learning, GPT-3 can be taught a new task in the input text passed to trained model.
  - Attention [86], and pre-training [58], allows for semantically meaningful features, trained on related tasks, that can the be applied to down-stream applications.
  - [88] and [89], teach us that task-specific conditioned learning, and in-context few shot learning is possible from a large corpus of knowledge. Perhaps, these principles from NLP can be used for transfer learning on large datasets of mass spectrometry data also. Whether it be the raw mass spectrometry data,
  - or fine-tuning language models with human supervision from chemists on task-specific mass spectrometry literature.
  - OpenAI take this idea one step further, with reinforcement learning with human feedback (RLHF) [90], for aligning language models with human intent.
  - Pre-training on exist mass spectrometry datasets, for example the METLIN metabolists database and LIPID MAPS, could create a semantically meaningful input vector, for the machine learning models to then utilize.

- Synthetic datasets
  - Synthetic datasets are often useful for exploring the limitations of a model in a controlled environment.
  - A recent paper from Uber regarding MRMR for market segmentation [91], uses Synthetic dataset to test effectiveness of feature selection algorithms, in a simluated customer data problem.
  - In the original paper for χ^2 feature selection algorithm [92], a synthetic dataset is used to simulate various levels of noise in the data, to test the algorithms robustness to said noise.
  - In the original Releif-F paper [75], synthetic datasets are used to model relationships of increasing high-order of polynomial complexity. The synthetic datasets can be used to control the strength of the noise, and the complexity of the signal.
  - In this research, the datasets have low sample complexity, due to time-consuming and laborious task of producing chemistry datasets.
  - Synthetic datasets can be used to explore robustness of models, test edge cases that are not present in real-world measurements thus far, and artificially inflate the sample complexity, to provide more training data.

2.5.4 Online Learning

- (see glossary for disambiguation)
- Real world examples: Tesla FSD
- The long tail of AI
- Fish processing:
  - Sample complexity will increase over time, as more samples are analyzed by the factory.

– Important not to have static models, that are rigid, and not robust to conceptual drift and out of distribution data anomalies.

2.5.5 Taxonomy

A clear taxonomy of equivalent terms across domains is needed. The terminology used to describe their methodology with chemistry/statistics jargon. A clear explanation of the equivalent terms between chemistry/statistics/Machine Learning terminology would open the field to further multi-disciplinary input from ML researchers. The glossary in this proposal the start of building that bridge between these disciplines.

- Jargon - the chemistry people say variable, the AI people say feature. The terminology can be used inter-changeably, but there are important differences.
- AI people use the term feature with domain agnosticism, AI researchers don't care / or understand the exact meaning of the feature with respect to the domain. In fact, AI researchers would rather not have to, good to build models that don't require domain expertise at all, or at least very little.
- Chemistry people use the term variable. This refers to the domain and the task at hand. If they are interested in lipids, a variable is a lipid of interest. When a chemist says variable it is inherently linked to domain-specific knowledge and means a very specific thing.
- Identification
- Profile
- Detection [27]
- Significant markers [18, 19]
- Outliers [18, 19]
- Relative Standard Deviation threshold [18, 19]
- Quality Control (QC) [18, 19]

2.6 Summary

This section provides a summary of the limitations of the existing work presented in the literature review, and how this thesis intends to fill those gaps. In particular, the research will focus on domain knowledge, state-of-the-art, transfer learning, and taxonomy.

- **Domain knowledge** - The thresholds to determine outliers are determined manually by domain experts. Their expertise in chemistry is needed to choose hyperparameters for every model - time. Significant markers are analysed and identified post hoc, relying on domain expertise in chemistry and human intuition. Manual hyper-parameter tuning (e.g. # principal components, RSD threshold for outliers, mass range) can be automatically selected, or replaced by models that don't need them at all!
- **state-of-the-art** - Mature statistical techniques are used for dimensionality reduction and classification, not state-of-the-art machine learning. Basic supervised statistical models (e.g. LDA, OPLS-DA) was used for classification. Future work should consider CNNs [79, 80], GANs [27], Diffusion [81, 82]

- **Transfer learning** - There is a large body of existing Mass-Spectrometry data. Knowledge from these datasets is not incorporated.Potential for transfer learning (incorporate previously existing data) to improve performance for few-shot classification tasks.
- **Online learning** - Many AIML models completely collapse when presented with new data, whether that be out-of-distribution anomalies [27], or conceptual drift where the underlying probability distribution changes over time - for example seasonal variation in composition of Hoki [8]. A flex-factory needs robust models, that can be updated with new information, and an online learning scenario, where edge cases are fed back as training data, to make them more robust.
- **Taxonomy** - The terminology used to describe their methodology with chemistry/statistics jargon. A clear explanation of the equivalent terms between chemistry/statistics/Machine Learning terminology would open the field to further multi-disciplinary input from ML researchers. The glossary in this proposal the start of building that bridge between these disciplines.

Chapter 3

Preliminary Work

This research builds on an existing body of research, this includes existing works presented in the previous literature review section and my own preliminary work. In this chapter, the focus is the preliminary work - work done before the proposal seminar. This section discusses classification and feature selection techniques that were applied to other fish chemistry datasets; these include support vector machines, Particle Swarm Optimisation, Convolutional Neural Networks, and Genetic Programming. The end chapter ends with an exploratory data analysis on a new fish chemistry dataset, Rapid Evaporative Ionisation Mass Spectrometry (REIMS), and discusses how the preliminary work can and cannot, be applied to the new dataset. Specifically, the sections of this chapter are:

1. **Automated Fish Classification on GC-MS data**
2. **Genetic Programming for GC-MS data**
3. **REIMS Exploratory Data Analysis**

In the remaining portion of this chapter, a detailed exploration of each of these sections will be conducted.

3.1 Automated Fish Classification on GC-MS data

The preliminary work starts by introducing previous research [12], this is important to understand the following preliminary work and future research directions. This work was undertaken outside the scope of this PhD but lays the groundwork for my preliminary work. In particular, this work provides a detailed explanation of the Gas-Chromatography Mass-Spectrometry (GC-MS) dataset. It includes an evaluation of classification and feature selection methods for fish species and part identification. This proposal also looks to find machine learning techniques for fish species and part identification, but now instead on state-of-the-art Mass-Spectrometry techniques. Should you be interested in Gas-Chromatography Mass-Spectrometry (GC-MS), species and part identification, I would recommend this paper, [12], as supplementary reading material, to avoid repetition, I will not repeat the contents of that paper here.

3.2 Genetic Programming for GC-MS data

This section describes preliminary work using Genetic Programming (GP) on Gas-Chromatography Mass-Spectrometry (GC-MS) data. The preliminary work on evolutionary computation provides insight into useful techniques for fish analysis on chemical datasets. These techniques

could be applied to the REIMS dataset. Specifically, this section covers the theory, the datasets, the experimental setup, and the results.

3.2.1 Theory

In the Genetic Programming (GP) subsection of the preliminary work, experiments benchmark three GP methods, to my previous work, [12], that was addressed in the last subsection. In particular, the three GP methods proposed in this work are:

1. Single-Tree Genetic Programming (ST-GP)
2. Multi-Tree Genetic Programming (MT-GP)
3. Multiple Class-independent Feature Construction Method (MCIFC)

The first method, ST-GP, is a standard Genetic Programming (GP). MT-GP is an extension of that which returns a list of single-tree GP. Algorithm 1 shows the pseudo-code of the Multi-Tree Genetic Programming (MT-GP). The multi-tree representation has m trees, with elitism ratio e.

Algorithm 1 GP-based multiple feature construction

```
Input : train_set, m;
Output : Best set of m trees;
Initilize a population of GP invidiuals. Each individual is an array of m trees;
best_inds ← the best e individuals;
while Maimum generation is not reached do
    for i = 1 to Population Size do
        transf_train ← Calculate constructed features of individual i on train_set;
        fitness ← Apply fitness function on transf_train;
        Update best_inds the best e individuals from elitism and offspring combined;
    end for
    Select parent individuals using tournament selection for breeding;
    Create new individuals from selected parents using crossover or mutation;
    Place new individuals into population for next generation;
end while
Return best individual in best_inds;
```

Representation

Multiple Class-independent Feature Construction Method (MCIFC) [74]. is a Multi-tree GP that constructs a smaller number of high-level features, proportional to the number of classes, from the original features. This method is based on the intuition that problems with more classes are likely to be more complex, and thus require more features to capture said complexity. The number of constructed features m, determined by $m = r \times c$, where r is the construction ratio (set to 2), and c is the number of classes. MCIFC constructs 8 features for the 4-class fish species problem and 12 features for the 6-class fish species problem.

Crossover and Mutation

MCIFC limits both the crossover and mutation operators to only one of the constructed features described in Algorithm 2. This approach favours exploitation over exploration,

making small random changes to constructed features with monotonically increasing fitness due to elitism.

Algorithm 2 MCIFC Crossover and Mutation.

```
prob ← randomly generated probability;
doMutation ← (prob < mutationRate);
if doMutation then
    p ← Randomly select an individual using tournament selection;
    f ← Randomly select a feature/tree from m trees of individual p;
    s ← Randomly select a subtree in f;
    Replace s with newly generated subtree;
    Return one new individual;
else
    p1, p2 ← Randomly select 2 individuals using tournament selection;
    f1, f2 ← Randomly select a features/trees from m trees of p1 and p2, respectively;
    Swap s1 and s2;
    Return two new individuals;
end if
```

Fitness

MCIFC takes the balanced classification accuracy of an SVM classifier as the fitness function. The SVM classifier is known to be effective for fish oil data [12]. Balanced accuracy avoids results bias towards the majority class, which is relevant for the fish species dataset, with the majority class 44% of samples belonging to fish species blue cod. The balanced accuracy is given by

$$\text{Balanced Accuracy} = \frac{1}{c}\sum_{i=1}^{c}\frac{TP_i}{TP_i + FN_i} \tag{3.1}$$

Where TP_i is the number of true positives for class i, and FN_i is the number of false negatives for class i, c is the number of classes.

3.2.2 Datasets

The gas chromatogram is the artefact of the Gas Chromatography method. The x-axis represents the time required to separate the individual fatty acids (or a packet), and the y-axis represents peak intensity (or the packet intensity), which is proportional to the concentration of each fatty acid. Chemists integrate the area under each peak to measure how much of each fatty acid is present, and use this information to understand the best use of the oil. This process can be slow, labour-intensive and expensive.

Figure 3.1 gives a gas chromatogram - the artefact of the gas chromatography - for tissue taken from the skin of a Snapper. Please see [12], for an example gas chromatogram and a more thorough description of the measurement technique.

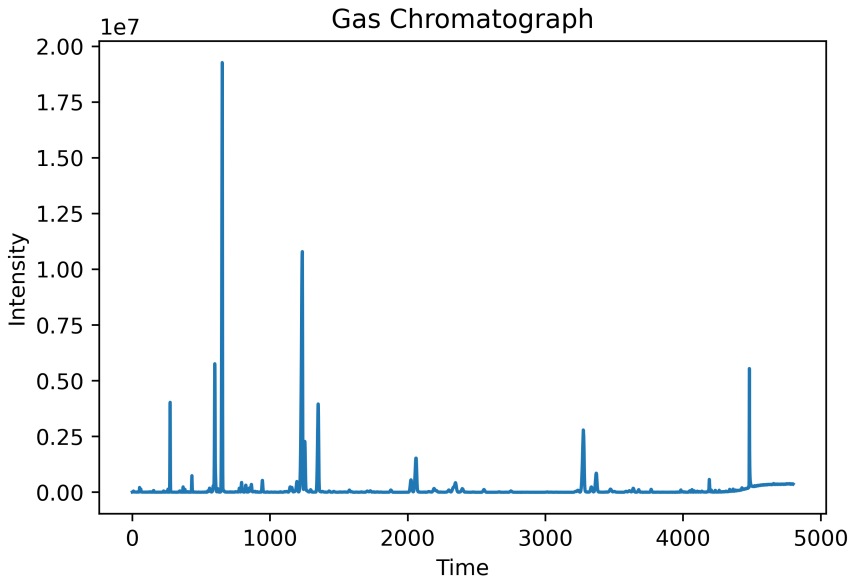


Figure 3.1: Gas Chromatogram of Fatty Acid Methyl Esters from Snapper Skin.

Table 3.1: Gas Chromatography Datasets.

Dataset	Features	Instances	Classes	Class Distribution
Fish Parts	4800	153	4	44% 17% 20% 19%
Body Parts	4800	153	6	15% 22% 14% 22% 14% 13%

Table 3.1 shows the datasets used in the experiments and their respective characteristics. Due to the high dimensionality of gas chromatography data, this paper employs a GP-based FC approach. The dataset is suited towards dimensionality reduction, as previous work [12] demonstrated FS can improve classification accuracy. The small number of instances is due to the expensive and time-consuming nature of performing Gas Chromatography on fish tissue. The data is pre-processed to fix the instrumental drift by imputing missing timestamps with zero filling. Features are normalized in the range $[0, 1]$ based on the training set.

3.2.3 Experimental Setup

Table 3.2 describes the parameter settings of all GP-based methods used in the experiments. The function set has standard arithmetic operators $+, -, \times$, a protected division operator that prevents division by zero returning 0 instead, and the unary *neg* operator reverses the sign. The feature set, and randomly generated constant $r \in [-1, 1]$, are used in the terminal set. A population of 100 individuals is used for all experiments, with 300 generations. The construction ratio r used to determine the number of features constructed is experimentally chosen as 2.

3.2.4 Results

Table 3.3 compares the classification results from [12], to the ST-GP, MT-GP, and MCIFC methods proposed in this preliminary work. The experiments use the same evaluation settings proposed in the original paper. The balanced classification average over stratified

Table 3.2: Paramter settings.

Function Set	$+, -, *$
Teriminal Set	$x_1, x_2, ..., x_n, r \in [-1, 1]$
Maximum Tree Depth	8
Population size	4800 (= #features)
Initial Population	Ramped Half and Half
Generations	300
Crossover	0.8
Mutation	0.2
Elitism	0.1
Selection	Tournament
Tournament Size	3
Construction ratio	2

Table 3.3: Results

Dataset	Method	Train	Test
Species	KNN [93]	83.57	74.88
	RF [94]	100.0	85.65
	DT [95]	100.0	76.98
	NB [96]	79.54	75.27
	SVM [97]	100.0	98.33
	MT-GP	97.52	72.61
	MCIFC	**100.0**	**99.64**
Parts	KNN	68.95	43.61
	RF	100.00	72.60
	DT	100.00	60.14
	NB	65.54	48.61
	SVM	100.00	79.86
	84.30	**84.30**	**86.80**
	MCIFC	**97.81**	84.30

cross-validation ($k = 10$) averaged over 30 independent runs. Balanced accuracy is used to counteract the class imbalance in the fish species dataset. The GC-MS dataset is expensive to time-consuming, leading to a low sample size, which motivates the use of cross-validation. The table gives an average over 30 runs to ensure results are statistically significant due to the stochastic nature of population-based Genetic Programming.

- MCIFC performs best on the test set for fish species identification.
- MCIFC overfits to the training set, and fails to generalize well on the test set, for fish part identification.
- MT-GP performs best for the test set for fish part identification.
- MT-GP overfits to the training set, and fails to generalize well on the test set, for fish species identification.
- When compared to FS methods from [12]:
  - for fish species identification.

    * MCIFC exceeds performance of all FS methods, [92, 71, 98, 99], with SVM [71]

- for fish part identification.

    * MCIFC is better than χ^2 [92] and the full dataset.
    * MCIFC offers same performance as PSO [71]
    * MCIFC is worse than ReliefF [98] and MRMR [99]
    * MT-GP offers competitive performance to MRMR [99], 86.80 % compared to 86.94 %, respectively.

3.3 REIMS Exploratory Data Analysis

This section reports Exploratory Data Analysis (EDA) on the new Rapid Evaporative Ionisation Mass Spectrometry (REIMS) dataset. First, it breaks down the theory. It explains the label annotations and breaks down relevant terminology, and, introduces species identification tasks. Second, the mass spectrum - the artefact produced by the REIMS dataset. Then, the results of preliminary classification models, and the implications of those results, in concert with domain expertise. Finally, ablation studies verify conjectures made by domain experts that serve as possible explanations for the results. The remainder of this section addresses each point with its own subsection.

3.3.1 Theory

This subsection on theory covers the relevant domain expertise on fish, chemistry and machine learning. First, the label annotations for the REIMS dataset are explained. Second, the species identification task is introduced, briefly enough to understand the proceeding experiments, but elaborated on further in the following chapter.

Annotated Labels

Figure 3.2 shows the annotated labels for the Rapid Evaporative Ionisation Mass Spectrometry (REIMS) dataset. This bar chart gives an effective view of the full dataset. This dataset is separated into five sub-datasets to address five sub-tasks: species, part, cross-species, mineral oil, and individual.

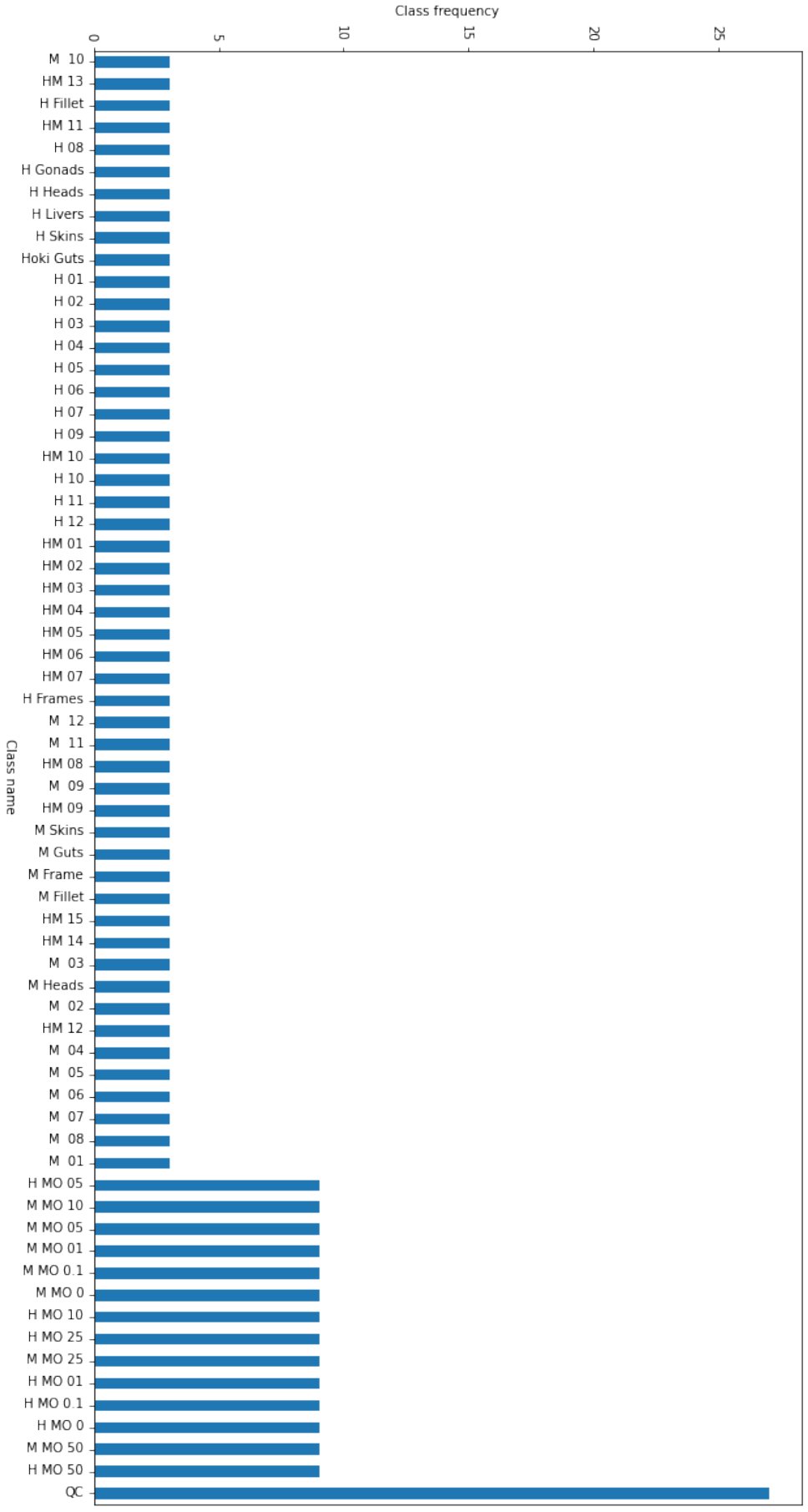


Figure 3.2: Class Distribution

The annotated labels encode information about what each instance is. For example, for the species identification task, the ”H” and ”M” letters correspond to the species of fish, and their combination represents a cross-species contaminated sample:

- H $\rightarrow$ Hoki - a species of fish.
- M $\rightarrow$ Mackeral - a different species of fish.
- HM $\rightarrow$ Hoki-Mackeral - a contaminated sample contains both species.

Proceeding with the species tag, there is either a number - the individual, tissue - the body part the sample belongs to, or Mineral Oil (MO).

Part - The part (or tissue) refers to which tissue of the fish body the sample was taken from. The fish parts considered in this research include fillet, frames, gonads, head, liver & skin.

Mineral Oil (MO) - The former are self-explanatory, but for the latter - MO, these annotations contain a decimal afterwards. Take, for example, ”M MO 0.1”, this represents a Mackerel species, contaminated with Mineral Oil, at a contamination rate of 0.1%. The Mineral Oil contamination rates $\in$ [0.1%, 1%, 5%, 10%, 25%, 50%]. Samples are contaminated at different rates because chemists are interested in the sensitivity of the contamination detection system. As the contamination rate decreases, it is expected the contamination detection task becomes more difficult.

Quality Control (QC) - or check samples, these are all identical, if the technique was working properly they should be tightly clustered, due to measurement noise they are not. The QC samples are a 50-50 mixture of the Hoki and Mackerel, they aim to be an average of the two fish. These are used as a baseline to calibrate and assess the quality of the measurements overall. Should these show high variance in a predictive model, this indicates it is not well suited to the REIMS dataset.

Relative Standard Deviation (RSD) threshold - The QC samples serve as additional data drawn from the same distribution, that can measure the quality of a model. Each predictive model should perform its sub-task well, and (additionally) show low variance for predicting this QC samples. Additionally, the QC samples serve an additional purpose, they identify spurious data points, in particular, when noise exceeds a threshold for identical QC samples. In Mass-Spectrometry, chemists often set an arbitrary 30% Relative Standard Deviation (RSD) threshold for noise. If a particular data point varies in the QC samples by more than 30% RSD, that measurement is removed from consideration for ALL samples in the dataset.

Species Identification

Species identification is a classification task, to identify the species of the sample, that belongs to a single class. In this preliminary work, the species identification task is to classify an instance as either Hoki or Mackerel, see fish in fig 1.3. Please see subsection 4.3 Species Identification for more information on this contribution. This subsection presents early results for the species identification task, addressing the limitations discussed in section 2.5 State-of-the-art ML.

3.3.2 Datasets

A mass spectrum measures mass charge versus intensity, where the **charge ratio** or m/z ratio is on the x-axis, where m is the **mass** - the amount of matter in an object, z is the **charge** of the ion. The mass charge ratio m/z is useful, as it allows us to differentiate between molecules of

the same mass, but different charges, or the same charge but different masses. The **intensity** on the y-axis refers to the relative abundance of ions in a mass spectrum, the intensity peak in a mass spectrum represents the number of ions with a particular mass-to-charge ratio that are detected by the mass spectrometer.

Figure 3.3 gives the mass spectrum, the artifact of the Mass-Spectrometry, for the first instance of the REIMS datasets. This mass spectrum was taken from a Hoki Fillet, that is the fish species of Hoki, and the body part Fillet.

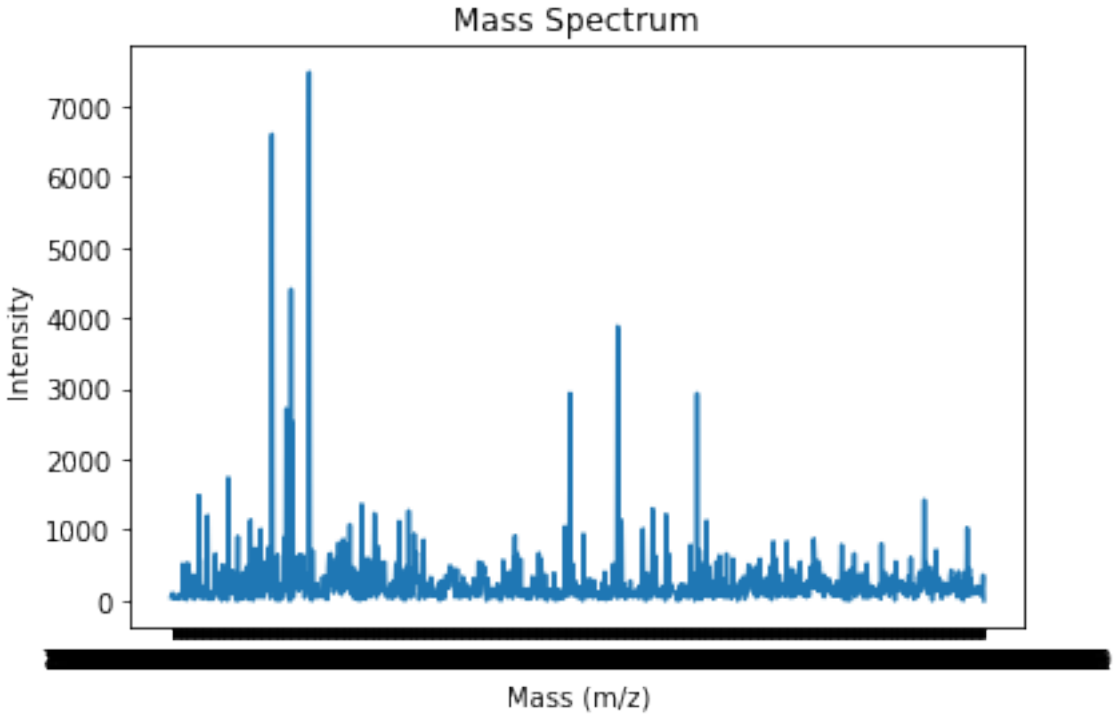


Figure 3.3: Mass Spectrum for a Hoki Fillet

Figure 3.4 gives the mass spectrums for the entire REIMS dataset. This gives an intuition for the range and variability across these measurements. The colours differentiate between the different annotated labels which are given in figure 3.2.

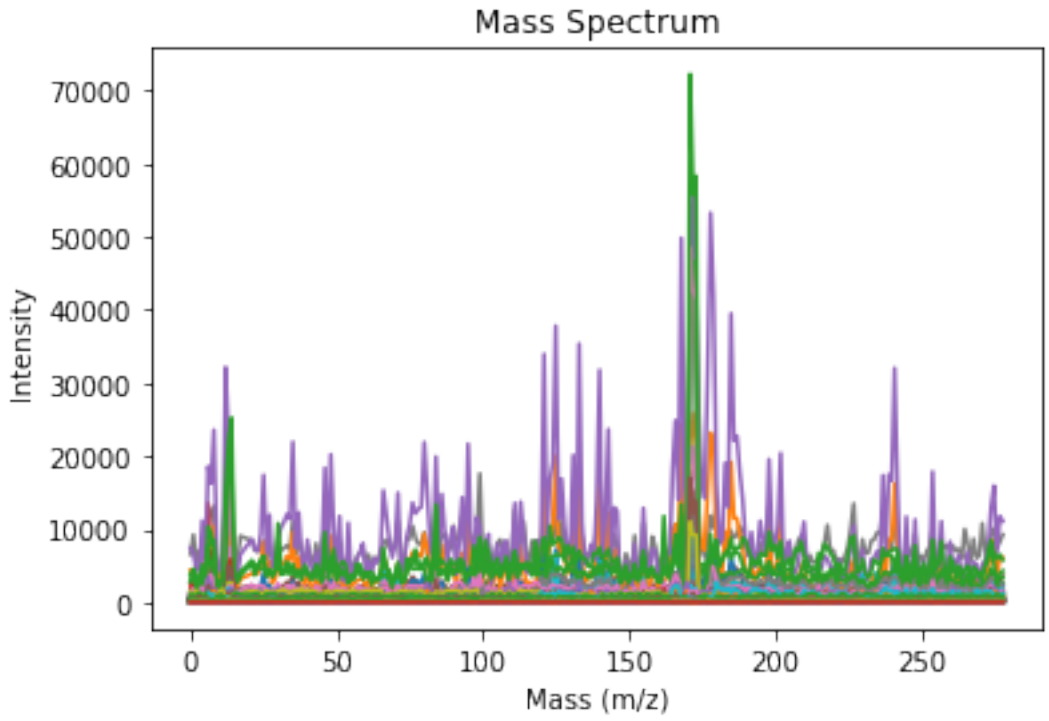


Figure 3.4: Mass Spectrums for Entire REIMS dataset

3.3.3 Results

Table 3.4 gives the results for preliminary experiments, exploring the performance of different dimensionality reduction techniques and classification algorithms on the REIMS dataset. In these preliminary experiments, the classification task is species identification. The dimensionality reduction techniques create $n = 20$ features. The table gives the mean and standard deviation classification accuracy on the test set over 10-fold cross-validation. The best-performing reduction method and classification, and respective classification accuracy, are in bold.

Method	SVC [97]	KNN [93]	DT [96]	RF [94]	XGBoost [100]	**LDA** [101]
PCA [25]	0.88 ± 0.17	0.85 ± 0.13	0.83 ± 0.15	0.87 ± 0.13	0.88 ± 0.14	$\mathbf{0.92 \pm 0.13}$
t-SNE [76]	0.70 ± 0.11	0.68 ± 0.11	0.55 ± 0.09	0.68 ± 0.07	0.69 ± 0.10	0.65 ± 0.11
UMAP [77]	0.84 ± 0.13	0.86 ± 0.14	0.81 ± 0.11	0.87 ± 0.12	0.88 ± 0.13	0.87 ± 0.14

Table 3.4: Dimensionality Reduction / Classification Methods for Species Identification

The table shows PCA-LDA [25, 101] (**in bold**) has a mean classification accuracy of 92% with a standard deviation of 10.3%. For reference, Principal Component Analysis - Linear Discriminant Analysis (PCA-LDA) is the primary technique used in existing literature, [18, 19] for REIMS datasets in the classification of raw biomass. The staple technique used in existing literature outperforms more recent feature reduction methods and a variety of classification methods. These initial experiments show, that despite neither PCA nor LDA being state-of-the-art when used in combination, on REIMS dataset, they perform incredibly well. The strengths of each of these techniques should be investigated, to find similar techniques that can provide competitive results.

Insights:

- PCA [25] Project data along the principal components, the axis of maximum variance in descending order.

- The first principal component is the axis of maximum variance, the second principal component is orthogonal to the first and has the second largest variance, and so on.

- The chemists at Plant and Food Research New Zealand Ltd. (PFR) said the first two principal components for REIMS seem to only capture noise. It is the third, fourth and later principal components that capture meaningful signals in the data.

- Perhaps, the reason PCA outperforms t-SNE and UMAP, is that PCA is able to implicitly denoise the dataset, by extracting and isolating the principal components, which can likely be attributed entirely to noise in the measurement. An ML model would simply ignore (or provide low weightings) these principal components, which are without signal and just noise.

- However, t-SNE and UMAP, due to their methodology, preserve the noise and incorporate it into the reduced dimensions of their projections. Unlike PCA, these dimensionality reduction techniques are unable to denoise the dataset.

- Denoising the dataset had a significant effect on the classification performance. This suggests it may be an important step in pre-processing, where PCA can be used in combination with classification models. Or, that a model with implicit denoising, such as a denoising auto-encoder [42] with a fully connected network for each sub-task, may yield noteworthy results.

- Furthermore, Generative Adversarial Networks (GAN)s have shown promise in anomaly detection [27], which is a closely related field to contamination detection and identification presented here.

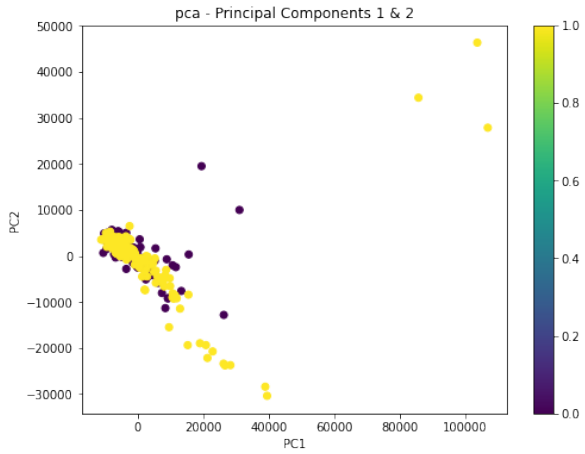
3.3.4 Ablation Studies

We can verify the PFR's conjecture made above, both visually and empirically, with an evaluation of the species identification task. To verify visually the ablation study gives a plot for class distribution for features 1 & 2, versus features 3 & 4, for each dimensionality reduction technique, the plot whose clusters are more visually distinct has less noise and more signal. To verify empirically, the ablation study can measure the prediction accuracy of a classification model trained solely on 1 & 2, versus features 3 & 4, the better performance indicates less noise and more signal in the extracted features.

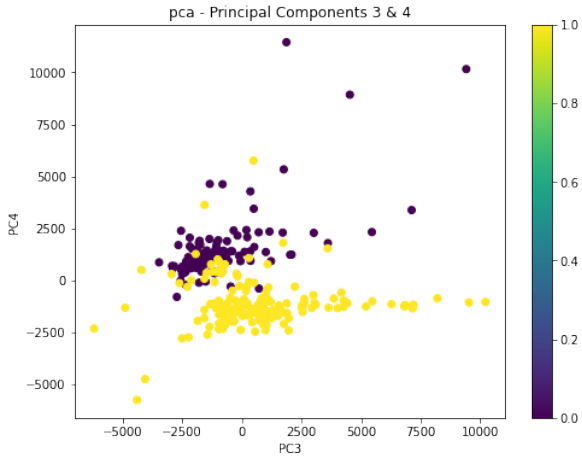
Table 3.5: Visual intuition for dimensionality reduction techniques and their respective feature subsets

PCA [25]

Features 1 & 2

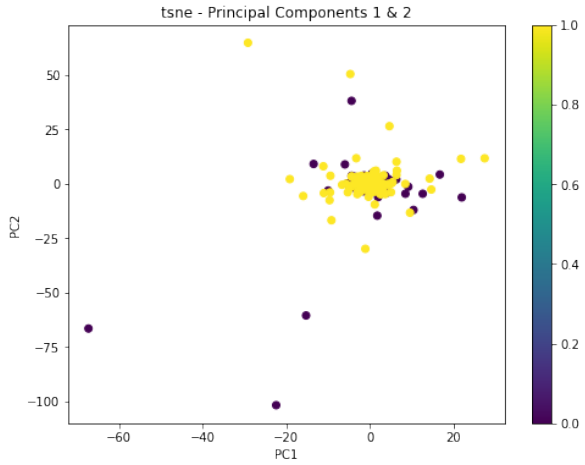


Features 3 & 4

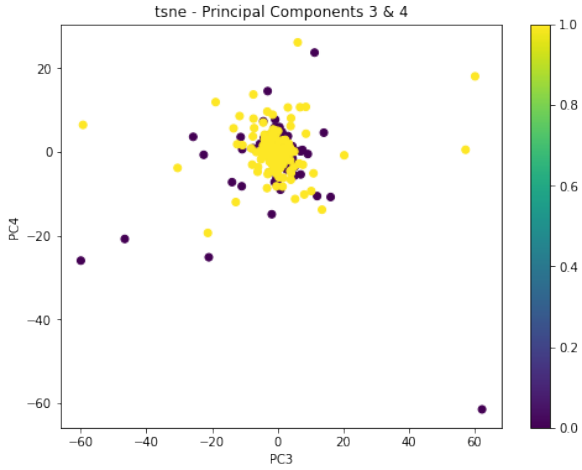


t-SNE [76]

Features 1 & 2

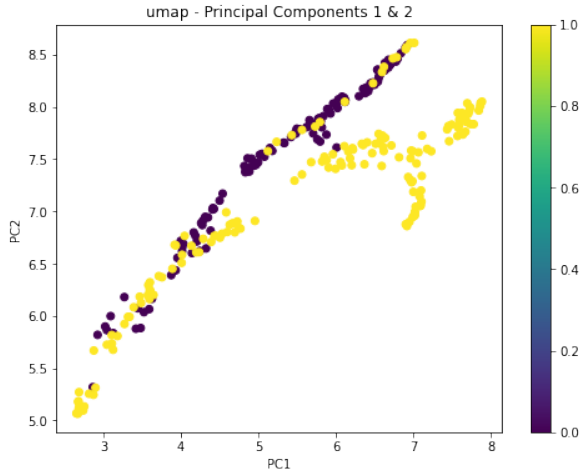


Features 3 & 4



UMAP [77]

Features 1 & 2



Features 3 & 4

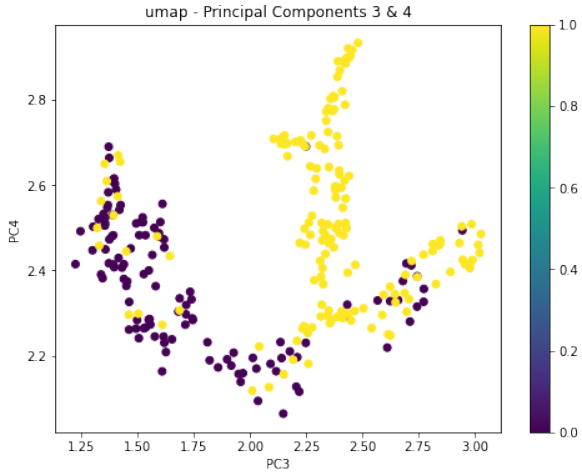


Table 3.5 gives the class distribution for features 1 & 2, versus features 3 & 4, for each dimensionality reduction method, PCA [25], t-SNE [76] and UMAP [77]. This gives intuitive and visual proof of the ability of each technique to tolerate noise in the dataset. The results, agree with the conjecture proposed by PFR, which suggests that the first two principal components are mostly noise, and principal components 3 & 4, offer more signal than the noise of principal components 1 & 2, for the species identification task. This table shows that other dimensionality reduction techniques, t-SNE [76] and UMAP [77], struggle to extract and isolate this noise, as the class distribution remains muddles for both features 1 & 2, and features 3 & 4.

Table 3.6: Empirical evaluation of dimensionality reduction techniques and their respective feature subsets

Method	Features 1 & 2	**Features 3 & 4**
PCA [25]	55.47$\pm$ 6.68	**86.40 $\pm$ 16.25**
t-SNE [76]	57.24 $\pm$ 2.03	55.80 $\pm$ 3.69
UMAP [77]	85.27 $\pm$ 15.17	81.23 $\pm$ 17.15

Table 3.6 gives the cross-validation score for each dimensionality reduction method, PCA [25], t-SNE [76] and UMAP [77], trained exclusively on features 1 & 2, versus features 3 & 4. The table gives the mean and standard deviation classification accuracy, with Support Vector Machine (SVM), on the test set over 10-fold cross-validation. The best-performing dimensionality reduction technique and feature subset, are given in bold. Results show with PCA [25] that features 1 & 2 have the lowest predictive accuracy, suggesting these are mostly noise. Conversely, features 3 & 4 have the highest predictive accuracy, exceeding that of all feature subsets for both t-SNE [76] and UMAP [77], suggesting that these provide an excellent signal for the species identification task.

We have demonstrated visually through intuition, and empirically through classification performance, that the conjecture that principal components 1 & 2 are mostly noise, and principal components 3 & 4 are provide signal, for REIMS data on the task of species identification. Furthermore, PCA [25] provides a pre-processing technique step for denoising REIMS data, it is able to isolate and extract noise, which leads to significant improvements in classification performance.

Chapter 4

Contributions and Project Plan

The remainder of this proposal focuses on execution, the goals of the research, and how to ensure the thesis meets those goals. This chapter presents the contributions this thesis will address, and gives a plan for how they will be delivered, and what is needed in order to achieve them. Specifically, this chapter covers contributions, milestones, thesis outline and resources. Specifically, the sections of this chapter are:

1. **Contributions**
2. **Milestones**
3. **Thesis**
4. **Resources**

The rest of this chapter will delve into each of these sections with greater detail.

4.1 Contributions

This research aims to evaluate two state-of-the-art Mass-Spectrometry techniques on their ability to determine bulk composition and quality of marine biomass rapidly. Both mass spectrometry techniques are used to analyze the same tissue samples. The composition and quality of marine biomass are evaluated by a series of sub-tasks. The contributions are ordered contributions as three tasks, each in ascending order of increasing difficulty. These are all related directly to domain-specific problems in fish processing. AIML techniques of increasing complexity will likely be required to solve these problems as their difficulty increases. In this section, those techniques and sub-tasks are defined, and then each explored in further detail.

4.1.1 Mass Spectrometry

Ultimately, chemists are interested in a technique that can provide rapid, interpretable and accurate analysis of marine biomass in a factory setting. To do so chemists employ state-of-the-art Mass-Spectrometry techniques, one known for its rapid speed, the other its high-resolution granularity. In particular, the two state-of-the-art Mass-Spectrometry techniques are:

1. Rapid Evaporative Ionisation Mass Spectrometry (REIMS) [57]
2. Direct Infusion Mass Spectrometry (DIMS)

There exists an age-old trade-off between speed and quality, told in the fable of the Tortoise and the Hare. These two datasets demonstrate this trade-off - REIMS is fast but low-resolution, DIMS is slow but high-resolution, online versus offline. Work from [18] shows near-instantaneous results (≈ 2 s) for the REIMS (hence the name). On the other hand, DIMS is much less rapid, because oils must first be extracted. Instead, this technique produces high-resolution data [102]. For deployment in a factory setting, speed is a must. Cybermarine want rapid results that match the pace of the production line. However, chemists don't want to sacrifice an acceptable standard of quality for speed. The DIMS dataset provides a benchmark for comparison to REIMS to ensure it meets this acceptable standard.

The analytical chemistry techniques need to work on fresh marine biomass, as cooking the fish produces a chemical change that destroys valuable information, for example, proteins, collagen and active enzymes. Cooking also requires time and energy, which adds expenses to the production line. In [18], REIMS results were worse on cooked biomass. Studies [18, 19] show that Mass-Spectrometry works on raw biomass products. A difference between the REIMS and GC dataset from [12], the GC data was subject to instrumental drift, and required processing to align timestamps. However, the new REIMS dataset has no instrumental drift! The technique will get the same measurements for the same QC sample, even if years apart (only day-to-day drift!).

There are two datasets that describe marine biomass, each with trade-offs - inherent strengths and weaknesses. Now, sub-tasks related to fish processing are needed to evaluate their feasibility for use in a factory setting. In particular, the sub-tasks used to determine the composition and quality of marine biomass are:

1. Identification
   (a) Species
   (b) Part
2. Qualitative Contaminant Anaylsis
   - Contaminants: Cross-species, Mineral Oil

   (a) Detection
   (b) Analysis
   (c) Quantification
3. Traceability
   (a) Detection
   (b) Sample Attribution

For the remainder of this section, each sub-task is defined, concerning biology/chemistry/fish processing, and their relation to machine learning.

4.1.2 Identification

Species identification [66] - can REIMS / DIMS data be used to classify different species tissues? What variables are responsible?

- Same task as [12], but instead of GC-MS, this is REIMS and DIMS
- Classification

- Feature Importance - Interpretable,
  - similar to significant markers from [18, 19]
  - and interpretability from [12, 85].

Fish tissue describes a particular part of the body of a fish. For example, these could include the head, guts, liver, frame, gonads or tail. In [12], one task addressed in that paper, is to predict the fish tissue a sample belongs to from gas chromatography datasets.

- The task of tissue prediction identifies which tissue the sample was taken from, i.e. body part of the fish, e.g. head, liver, gut, fin, gonad, etc...
- Classification
- Feature Importance - Interpretable,
  - similar to significant markers from [18, 19]
  - and interpretability from [12, 85].

4.1.3 Quantitative Contaminant Analysis

Cross-species contamination - can REIMS / DIMS data detect mixed-species contamination in fish tissues? At what concentration? What variables are responsible?

- Quantitative contaminant analysis - (ChatGPT REWORD!!!) The method you described appears to be a quantitative contaminant analysis method, as it is able to determine not only the presence of a contaminant but also the percentage of the sample that is contaminated. This information can be used to evaluate the severity of the contamination and to determine whether the sample meets the required standards for safety and quality.
- Similar to [18], but instead of beef-horse, this is for fish contamination.
- few-shot learning (very few training instances)
  - transfer learning, active learning or zero-shot inference may be needed.
- Detection $\approx$ Multi-label classification
- Identification $\approx$ multi-output regression
  - find anomalous instances!
  - Identify the percentage of cross-species contamination.
  - Potentially, even those outside of annotated labels.
- Feature importance (again) - significant markers
  - profile - how much contamination? confidence?

Mineral oil contamination Can REIMS / DIMS data detect mineral oil contamination in fish? At what concentration? What variables are responsible?

- Marine biomass can be contaminated with several things, for example, plastics and mineral oil - which are carcinogenic (it kills). This research seeks to develop tools that can identify contamination in marine biomass.

- Detection $\approx$ classification
- Identification $\approx$ multi-output regression/classification, i.e. identify true/false oil contaminated, and what percentage is oil?
- Feature importance (again x2) - significant markers
  - profile - how much engine oil? dangerous? confidence?

Black Swans - the unknown unknowns, outliers - PFR emphasized the robustness required, our AIML models need to handle out-of-distribution data, that is to identify classes that are possibly not even in the training data.

- "Black Swans are events or pieces of knowledge that sit outside our regular expectations and therefore cannot be predicted." [103]
- Popularized by Nassim Taleb, a risk analyst, in his books [104, 105].
- real-world examples: Pearl Habour, the internet, COVID-19, ChatGPT
- epistemology:
  - because no one had ever seen a black swan.
  - Until 1679, it was common to refer to impossible things as black swans.
  - Dutch explorer Willem de Vlamingh went to western Australia in 1697 and saw a black swan.
- for fish:
  - Out-of-distribution classes in a supervised learning problem are black swans.
  - For fish processing, these things going into the factory, they have yet to previously encounter.
  - It is not expected that AIML models willy correctly classify out-of-distribution data, but AIML models can try to detect these anomalies.
  - See [27] for anomoly detection using GANs, similar to [19] where thresholds are established for unknown outliers.
  - Detected anomalies found via REIMS, can be sent away for offline high-resolution processing, to identify/profile outliers, and then annotate labels for these classes in future datasets.

4.1.4 Traceability

Instance recognition - can REIMS / DIMS data be used to distinguish between different fish individuals? What variables are responsible?

- Identify the unique chemical signatures of individual fish
- Useful for Quality Assurance purposes, allows identification and isolation of any samples that may be contaminated or otherwise problematic.
- Identification
- Feature importance (again x3) - significant markers
  - profile - species? part? confidence?
- Seasonal variation, i.e. [8]

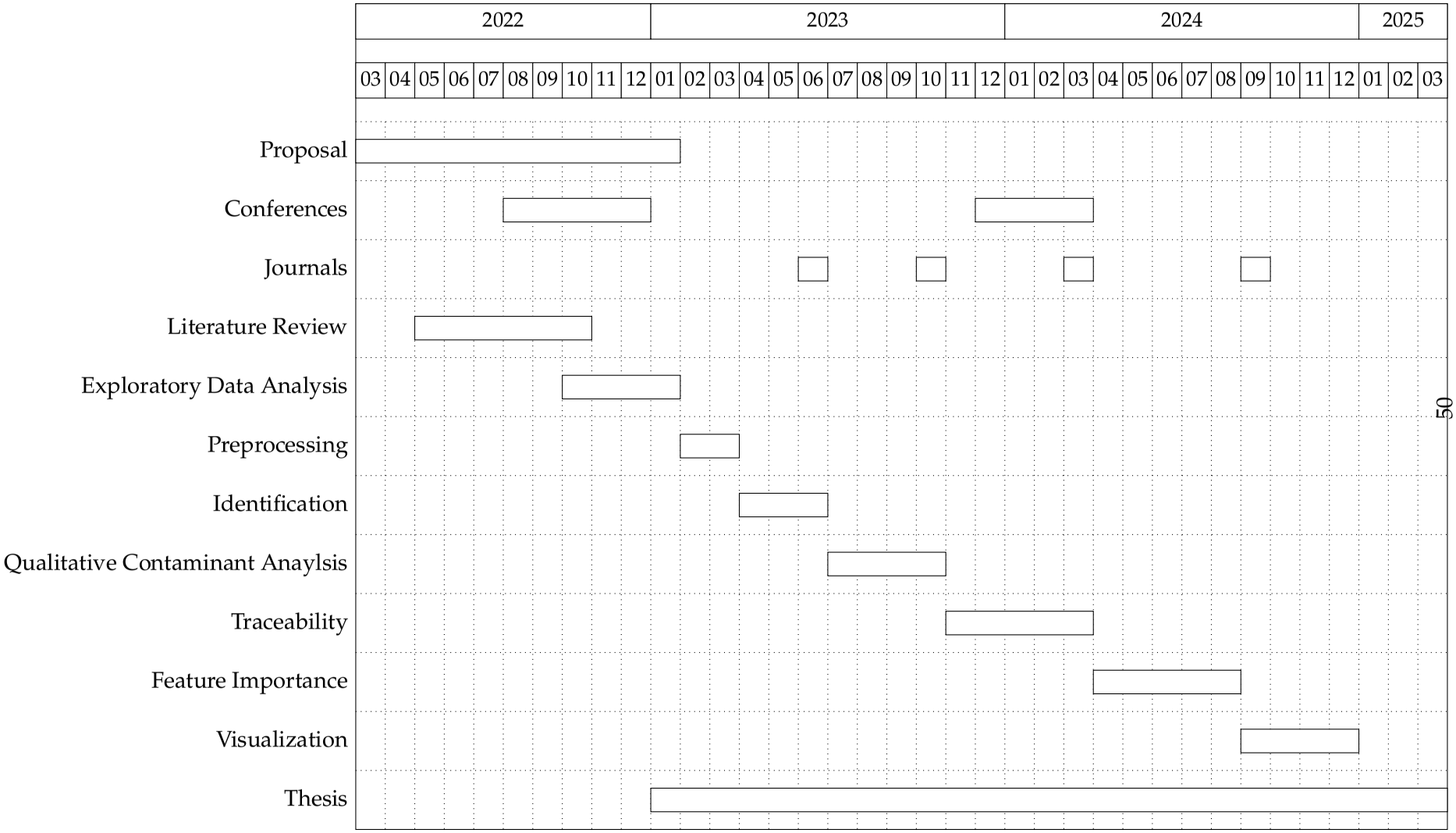
4.2 Milestones

This research project has several key milestones it aims to achieve in the course of our work. In particular, the milestones for this proposal are:

1. Proposal
2. Conferences (x2)
3. Journals (x4)
4. Literature Review
5. Exploratory Data Analysis
6. Preprocessing
7. Identification
8. Contamination - Detection
9. Contamination - Analysis
10. Contamination - Quantification
11. Traceability - Detection
12. Traceability - Sample attribution
13. Feature Importance
14. Visualization
15. Thesis

The work of this thesis will be submitted to relevant peer-reviewed journals and conferences. The aim is for the work to be accepted into (at least) two academic conferences, and four journals. For a 3 - 3.5 year PhD, these publication milestones are ambitious, but they will increase credibility, quality and public awareness of the work completed during the project.

These milestones include completing a literature review, conducting exploratory data analysis (EDA) and preprocessing, implementing classification algorithms for species and part identification, developing methods for contamination detection and identification, identifying significant markers and feature importance, creating visualizations to aid in data interpretation, and completing the final thesis. The milestones are crucial in reaching the overall goal of developing a rapid and accurate method for determining the bulk composition and quality of marine biomass using mass spectrometry.


2022
2023
2024
2025
03 04 05 06 07 08 09 10 11 12 01 02 03 04 05 06 07 08 09 10 11 12 01 02 03 04 05 06 07 08 09 10 11 12 01 02 03
Proposal
Conferences
Journals
Literature Review
Exploratory Data Analysis
Preprocessing
Identification
Qualitative Contaminant Anaylsis
Traceability
Feature Importance
Visualization
Thesis
50

4.3 Thesis Outline

The goal of this research is to develop a rapid and accurate method for determining the bulk composition and quality of marine biomass using Mass-Spectrometry. Specifically, the thesis outline has the following structure:

1. Introduction
2. Background
   (a) Mass-Spectrometry
   (b) REIMS / DIMS
   (c) Detection / identification
   (d) Interpretable ML
3. Preparations
   (a) Exploratory Data Analysis
   (b) Preprocessing
4. Applications
   (a) Identification
   (b) Qualitative Contaminant Anaylsis
   (c) Traceability
5. Demo - (AI Stunt department)
6. Discussion
7. Conclusion
8. Appendix
   (a) Taxonomy
   (b) Glossary

In this thesis, the various steps and techniques that will be employed in this process are given, including the use of Mass-Spectrometry techniques such as REIMS/DIMS and the application of interpretable machine learning for detection and identification. The thesis also describe the necessary preparations, including Exploratory Data Analysis and preprocessing, and the specific applications of this method, including fish species and part identification, cross-species contamination detection, and individual identification. Finally, results of these research are given, and conclusions are drawn.

The appendix includes a taxonomy and glossary to bridge the multi-disciplinary gap in knowledge. The majority of readers will only know one of those disciplines. The glossary provides a quick point of reference for jargon, to reduce the cognitive load for the read. A taxonomy - this will break down the terminology from a chemistry/biology, fish processing and machine learning perspective. This addresses an important gap in the existing literature, where current papers, [18, 19] rely heavily on jargon from chemistry and statistics, where synonyms or equivalent terms in machine learning exist. Removing the barrier of jargon between disciplines will make it easier for multi-disciplinary future work, making the field more accessible to machine learning researchers.

4.4 Resources

Table 4.1: Resources

Software	Hardware	Human
Python	ECS Grid	Plant & Food Research
C++	Rapoi	Callaghan Innovation
Open-source	Niwa HPC	
Documentation		
Project management		

To effectively conduct this research, a variety of resources are utilized. This section breaks those resources down into hardware, software, human resources and financial. Table 4.1 gives a high-level view of those resources. In the remainder of this section, each of those resources is covered in further detail.

4.4.1 Software

This research project will use Python and potentially C++ for programming and will make all source code open-source. Project management practices including agile methodology will be employed, and documentation will be hosted on Read the Docs. In particular, these are the software chosen and their justifications:

- **Project management** - project management practises such as: agile methodology, kanban boards, minutes of the meeting, milestones, sprints, and meeting with the client (industry partner Daniel Killeen), will be adopted to ensure research objectives are met.
- **Python** - is the primary programming language, it is free, versatile, and the most popular programming language worldwide. There is a large developer community, and there exists extensive support for machine learning applications.
- **C++** - while Python is suitable for rapid prototyping and ease of use. Should there be any algorithmic bottlenecks for computations that make the research intractable, I will consider refactoring those algorithms into C++.
- **Documentation** - Read the Docs to host and maintain a documentation website for the software outputs for the research.
- **Open-source** - any source code written for this research will be open-source, released under an MIT license, and openly available on GitHub. An example Google Colab notebook for the preliminary experiments is available here: `https://bit.ly/3iJNaZe`. This increases reproducibility, transparency, and dissemination of knowledge. (Note: the datasets remain the property of Plant and Food Research and Callaghan Innovation)

4.4.2 Hardware

Distributed cloud computing is a powerful resource for running machine learning algorithms, particularly population-based genetic programming. There are several reasons why distributed cloud computing is useful for these types of algorithms:

1. **Scalability** - Distributed cloud computing allows for the parallelization of machine learning algorithms, allowing them to scale up as needed to process large amounts of data. This is particularly useful for population-based genetic programming, which can involve the simultaneous evaluation of many different solutions.

2. **Cost effectiveness** - Distributed cloud computing can be more cost-effective than running machine learning algorithms on local hardware, as it allows for the use of resources on an as-needed basis without the need to invest in expensive hardware.

3. **Flexibility** - Distributed cloud computing allows for the use of a wide range of resources and configurations, allowing users to tailor their setup to the specific needs of their machine-learning algorithms. This can be particularly useful for population-based genetic programming, which may require different configurations depending on the problem being solved.

Overall, the use of distributed cloud computing can greatly improve the efficiency and effectiveness of machine learning algorithms, particularly population-based genetic programming. This is why for hardware, this research will be using the ECS Grid Compute and Rapoi systems, as well as the Niwa HPC through Auckland University.

4.4.3 Human Resources

In addition to these resources, I have also gained valuable experience through previous field trips to NZ Plant and Food Research, where I saw GC-MS first-hand for my previous publication [12]. This trip gave insights into steps in the ocean-to-plate supply chain, as their research laboratory processed whole fish into fish oil tissue samples suitable for Mass-Spectrometry techniques. With another trip to the Nelson-based Plant and Food Research, I could see DIMS in person. Lastly, it would be invaluable to plan a trip to the Wellington-based Callaghan Innovation, to see the REIMS in person.

4.4.4 Financial

Publications to conferences are to be expected, following on from [12], further publications at future AJCAI and the international IJCAI, and other conferences for evolutionary computation, e.g. CEC, GECCO, EvoStar, are to be expected. Therefore a travel grant would be expected to support these endeavours.

Glossary

adulteration Food adulteration is the act of intentionally debasing the quality of food offered for sale either by the admixture or substitution of inferior substances or by the removal of some valuable ingredient [20] . 2, 22, 23

AI Artificial Intelligence. 3, 10

AIML Artificial Intelligence Machine Learning. 4–7, 22, 30, 45, 48

analysis Analysis is concerned with identifying which contaminants are present. Not to be confused with **detection**, which simply tells us if sample is contaminated. Analysis takes this one step further, and gives predictions for which contaminants are present in the fish tissue. Take for example, **cross-species contamination**, contaminant analysis predicts which species are present in a contaminated sample, e.g. detection: contaminated, analysis: Hoki and Mackerel both present . 17, 18, 49

anomalies Anomalies refer to out-of-distibution data that the model could not possibly expect. It is unrealistic for the model to correctly classify these instances, but a model can be built to detect such anomolies, as seen in [27]. In fish processing, an example of an anomoly would be a new species of fish, or marine biomass, that is not a labelled class or present in the training or validation data. . 29, 30

charge characteristic of a unit of matter that expresses the extent to which it has more or fewer electrons than protons. Electric charge is the physical property of matter that causes it to experience a force when placed in an electromagnetic field. In the context of mass spectrometry, particularly REIMS which uses a Time-of-Flight (TOF), this uses an electric field to accelerate generated ions through the same electrical potential, and then measures the time each ion takes to reach the detector. Depending on the charge of each particle, that time will vary, because the electric field applies different amounts of force to particles with different charges . 38

CNN Convolutional Neural Networks. 13, 27, 29, 31

concept drift See **conceptual drift** . 6

conceptual drift A term from data stream mining, [61, 22], that refers to a change in the underlying distribution of the data. In fish processing, conceptual drift occurs in **seasonal variation** where the composition of fish changes between different seasons . 22, 29, 30

contamination Food contamination is generally defined as foods that are spoiled or tainted because they either contain microorganisms, such as bacteria or parasites, or toxic substances that make them unfit for consumption. A food contaminant can be biological,

chemical or physical in nature, with the former being more common. These contaminants have several routes throughout the supply chain (farm to fork) to enter and make a food product unfit for consumption [24] . 2, 3, 6, 7, 9, 14, 18, 19, 22, 23, 41, 47, 51

cross-species Cross-species refers to a form of contamination, where two species are mixed together, e.g. a sample with both Hoki and Mackerel. In the mass spectrometry datasets, these species are mixed thoroughly in a blender to give a homogeneous sample with maximum blend of the two species. . 5, 6, 18, 19

cross-validation For *k*-fold cross-validation, the method divides the data into *k* folds such that the proportions of the classes in each fold are representative of the proportions in the whole dataset. Each fold plays the testing role, while the remaining (*k*-1) folds are combined to form a training set. . 35, 40, 43

Cyber-marine Cyber Physical Seafood Systems (Cyber-Marine) is a new multi-million dollar research programme aimed at achieving 100% utilisation and maximised value for all harvested wild and aquacultured seafood. Making use of all raw material will allow the industry to achieve growth targets without increasing catch volume from wild-capture fisheries as well as maximise value from increasing aquaculture. Once established for the seafood industry, the technology could be adapted for any bio-industrial process [4] . 1

DDIM Denoising Diffusion Implicit Models. 27

DDPM Denoising Diffusion Probabilistic Models. 27

detection Detection finds if something is hidden in a sample. It does not have to specify what was hidden, only that sample had something hiding. E.g., it can detect some form of **adulteration**, **cross-species** contamination, or mineral oil in a fish sample . 6, 7, 9, 17–19, 22, 29, 41, 47–49, 51

DIMS Direct Infusion Mass Spectrometry. 14, 22, 45–48, 51, 53

DL Deep Learning. 5

domain knowledge Knowledge related to the application domain. For example, bio-chemistry and fish processing. . 29

EC Evolutionary Computation. 24, 25

EDA Exploratory Data Analysis. 36, 49–51

FC Feature Construction. 34

FS Feature Selection. 35, 36

GAN Generative Adversarial Networks. 6, 11, 12, 41

gas chromatogram Gas Chromatography for fatty acid analysis in [12]. The gas chromatogram is the artefact of the Gas Chromatography method. The x-axis represents the time required to separate the individual fatty acids (or a packet), and the y-axis represents peak intensity (or the packet intensity), which is proportional to the concentration of each fatty acid. Chemists integrate the area under each peak to measure how much of each fatty acid is present, and use this information to understand the best use of the oil. This process can be slow, labour-intensive and expensive . 33

GC Gas-Chromatography. 46

GC-MS Gas-Chromatography Mass-Spectrometry. i, 23, 31, 35, 46, 53

genotype In biology, the genetic material (i.e. DNA), e.g. the recessive trait for ginger hair colour. In Evolutionary Computation, the representation or encoding for an individual candidate solution. . 25

GP Genetic Programming. 22, 31, 32, 34, 35

heterogeneous The antonym of **homogeneous**. Consisting of many different elements. In the context of fish processing, New Zealand's marine biomass, the incoming catch from trawling vessels, is hetergeneous, as it consists of many different species - a wide range of marine biomass. . 5, 17

homogeneous This term is used heavily in chemistry. In the context of chemistry homogeneous means the same, or having a similar structure. In fish processing, the fish tissue samples are taken from a homogeneous blend of marine biomass. Also, in Hoki season, the input to the flex-factory is predominantly one species, this may also be referred to as homogeneous. The marine biomass of Canada or the United States, the incoming catch from trawling vessels, is homogeneous, as it consists of mostly one (or few) species - a narrow range of marine biomass. . 5, 11, 16

hyperparameter Hyperparameter (machine learning) In machine learning, a hyperparameter is a parameter whose value is used to control the learning process. These are often manually set by the user, and are comparable to nuisance parameters from statistics, as they require tuning for models to perform well. . 25, 26, 29

identification Different to detection, identification involves detecting the presence of phenomena in a sample and then specifying what the phenomena were. E.g., an identification system can find **cross-species** contamination and identify both species in the contamination . 4, 15, 16, 22, 29, 31, 38, 40, 41, 46–51

instance identification In computer vision, this is referred to as instance identification [31], not to be confused with **instance segmentation** [30]. Instance identification is the task of identifying unique instances in a photograph. Take for example a photograph with 5 sheep. Instance identification would identify each of the individual sheep, as a unique individual. In the context of fish processing, instance identification correctly assigns the origin of a sample, which unique individual fish it originated from. In chemistry, we refer to this as **sample attribution**, for the purposes of this proposal, treat the terms as interchangeable. . 8, 9, 57

instance identification In computer vision, this is referred to as instance segmentation [30], not to be confused with **instance identification** [31]. Instance segmentation is the task of drawing identifying the area of a picture, that belongs to a single instance. Note that a picture may include multiple instances that belong to the same class. Take for example a photograph with 5 sheep. Instance segmentation would correctly identify the region that belongs to each individual sheep. . 8, 9

instance recognition The machine learning term for recognizing individuals that may belong to the same class is "instance recognition" [29], or "individual recognition". For fish processing, instance recognition would involve recognizing each individual fish in the samples and assigning a unique identifier or label to each of them. This would

allow the model to differentiate between individual fish even if they belong to the same species. Instance recognition is a type of object recognition task that goes beyond simply recognizing object classes and aims to identify each individual instance of an object class. It is commonly used in various fields such as wildlife monitoring, security surveillance, and biometrics. . 7–9

intensity The intensity on the y-axis refers to the relative abundance of ions in a mass spectrum, the intensity peak in a mass spectrum represents the number of ions with a particular mass-to-charge ratio that are detected by the mass spectrometer . 39

intraclass Intraclass variation, a term from computer vision, refers to the variation that exists within instances of the same class. Computer vision would observe intraclass variation in what defines a chair, they can look very different! For fish processing, their is significant intraclass variation in fish parts, compared to fish species, a finding supported by [12]. . 16

KL Kullback-Leibler. 26

KL K-Nearest Neighbours. 25

LLM Large Language Model. 11

LNBNN Local Naive Bayes Nearest Neighbours. 13

marine biomass A fancy term for fish. To get super technical, marine biomass is a super-set, which includes fish, whales, plankton, crustaceans, marine animals and plants. A fish processing plant will deal with marine biomass from many forms of organic matter. So marine biomass is a catch-all term to refer to the incoming biological materials that enter the factory . 1–3, 21, 22, 45–47, 49, 51

mass The amount of matter in an object . 38

mass charge ratio The mass charge ratio m/z is useful, as it allows us to differentiate between molecules of the same mass, but different charges, or the same charge but different masses. . 38

mass spectrum The mass spectrum, is the artefact of the mass spectrometry technique. A mass spectrum measures mass charge versus intensity, where the **charge ratio** or m/z ratio is on the x-axis, where m is the **mass** - the amount of matter in an object, z is the **charge** of the ion. The mass charge ratio m/z is useful, as it allows us to differentiate between molecules of the same mass, but different charges, or the same charge but different masses. The **intensity** on the y-axis refers to the relative abundance of ions in a mass spectrum, the intensity peak in a **mass spectrum** represents the number of ions with a particular mass-to-charge ratio that are detected by the mass spectrometer . 36, 39

MCIFC Multiple Class-independent Feature Construction Method. 32–36

ML Machine Learning. 29, 30, 51

MO Mineral Oil. 6, 18, 19, 38

MS Mass-Spectrometry. 16, 18, 22, 26, 27, 30, 31, 38, 39, 45, 46, 51, 53

MT-GP Multi-Tree Genetic Programming. 25, 32, 34–36

NBNN Naive Bayes Nearest Neighbours. 13

NSP Next Sentence Prediction. 19

offline see **online** . 46, 48

online In a factory setting, the terms online and **offline** have distinct meanings. For a factory where efficiency and continuous flow of the production line are vital, there exists a tradeoff between online and **offline**. Online describes processes that are instantaneous and inexpensive, these are often low resolution but can be done at scale and at speed, so they don't slow down the production line. Conversely, **offline** means it will take days, take for example a tissue sample that has to be sent away for analysis, where results won't return for several days. We want to avoid offline, unless strictly necessary, or provide a significant benefit. Not to be confused with **online learning** . 14, 46

online learning Online learning refers to a model that can be updated and adapt to new instances after its initial training. Take for example the Tesla FSD training programme. The FSD edge cases are referred to as the long tail of computer vision. These edge cases are where the car demonstrates undesirable behaviour, e.g. a crash, swerve, unsafe/irregular driving, are sent back to the DOJO computing facility, and the model is retrained via Monte-Carlo simulation of that edge case, to perform the desired behaviour. This human-in-the-loop online learning, is a powerful method to bootstrap algorithms for robustness. Not to be confused with **online** . 25, 30

part A fish part refers to which tissue of the fish body the sample was taken from. The fish parts considered in this research include fillet, frames, gonads, head, liver & skin. . 16, 22, 31, 38, 46, 49, 51

PCA Principal Component Analysis. 12, 26, 40, 41, 43

PCA-LDA Principal Component Analysis - Linear Discriminant Analysis. 12, 23, 40

PFR Plant and Food Research New Zealand Ltd.. 41

phenotype In biology, the expression of a gene, e.g. hair colour. In Evolutionary Computation, the output of an encoded representation, e.g. a classification label, regression output, one-hot encoded vector. . 25

PSO Particle Swarm Optimisation. 25, 31, 36

QA Quality Assurance. 17, 48

QC Quality Control. 2, 3, 22, 27, 29, 38

QCA Qualitative Contaminant Anaylsis. 15, 17, 46, 50, 51

quantification Quantification assesses how much a sample is contaminated. Take for example **cross-species contamination**, **quantification** is interested in the percentage of contaminant from each species, e.g. 70% Hoki, and 30% Mackerel. . 17, 18, 49

recall Recall is a metric for classification accuracy. It measures the proportion of actual positives that were correctly identified. It can be thought of as $\frac{TP}{TP+FN}$, where TP is true positives, and FN is false negatives. **contamination detection** requires a high precision. There is lenience for false positives, a flex-factory should catch all samples that are truly contaminated. Closely related, and not to be confused with **precision**. . 18

REIMS Rapid Evaporative Ionisation Mass Spectrometry. 4–6, 8–11, 13, 14, 17, 18, 22, 23, 26, 27, 31, 32, 36, 38–41, 43, 45–48, 51, 53

RSD Relative Standard Deviation. 26, 29, 38

sample attribution Sample attribution is a chemistry term, that refers to identifying which individual sample a measurement was taken from. In **mass spectrometry**, several measurements are taken from the same fish **tissue** sample. Being able to identify measurements as from the a common origin, i.e. the same sample, is important for **traceability**. This can be used to isolate contaminated samples, or deducing a samples path through the factory. . 8, 18, 19, 49

sample complexity A term from deep learning [42], which refers to the number of training samples available. A dataset with many training instances (i.e. thousand or more) has high sample complexity. Conversely, a dataset with few training instances has low sample complexity. In general, complex methods such as deep learning, require large volumes of data for a model to be trained properly. In relation to fish processing, the sample complexity is low, due to the time-consuming and manually intensive process of collecting said samples. However, as these methods are deployed in real-time, the sample complexity will increase. . 28

seasonal variation The composition of **marine biomass** varies by season, a reoccurring **conceptual drift**. The temperature of the ocean, diets of fish, changes from Winter to Summer, oceans heat up, migration/spawning. For example, while spawning, Hoki changes composition, extracting their lipids, and putting them all into their eggs, after spawning adult Hoki is a mess [8]. . 19, 21, 22, 30

significant markers Significant Markers (or important variables) are ions that are unique to a specific offal cut, and present in all samples [19] . 27, 29, 47–49

SOTA state-of-the-art. 29

spawning Spawning is the reproductive process in which marine biomass release their eggs and sperm into the water. This is important for producing new offspring. The spawning of [8] is of particular interest, as it causes **seasonal variation**. . 21

species This refers to the species of fish that the tissue sample belongs to. The fish species in this research are Hoki and Mackerel. The species considered in previous work [12] were Bluecod, Gurnard, Snapper & Tarakihi. For differentiating between distinct species in fish fraud detection see [18]. See [66] for biological definition from Darwin. . 16, 22, 31, 38, 40, 46, 49, 51

spoilage Spoilage in a fish processing context refers to the decay or deterioration of fish or seafood products, resulting in a loss of quality and edibility. Fish spoilage can occur due to various factors such as bacterial growth, enzymatic activity, oxidation, and physical damage during handling, transportation, or storage . 2

ST-GP Single-Tree Genetic Programming. 25, 32, 34

stochastic Stochastic is the opposite of deterministic. A deterministic algorithm will produce the same results each run. A stochastic algorithm does not, it has a degree of randomness to it, in which the results will vary with each run. The stochastic nature of genetic programming is their strength, which allows for global search . 35

SVM Support Vector Machine. 33, 36, 43

t-SNE T-distributed stochastic neighbor embedding. 26, 40–43

taxonomy A taxonomy is a hierarchical classification system that organizes a set of concepts or subjects into categories and subcategories based on shared characteristics. Taxonomies are often used in fields such as biology, where they are used to classify and organize living organisms into a systematic hierarchy based on their characteristics and evolutionary relationships. They are also used in other fields, such as information science and library science, to classify and organize knowledge in a way that is easy to understand and navigate . 25, 29, 30, 51

tissue See part . 16, 33, 38, 45, 47, 53

traceability Traceability is a term from quality assurance, which is important in a factory setting. Should a problem arise, a factory needs to be able to isolate, and determine the origin and potential causes for that problem. Take for example fish tissue contaminated by mineral oil. After detecting said **contamination**, traceability would be concerned in identifying other tissue samples from the same fish that are likely contaminated. . 7, 15, 46, 50, 51

transfer learning Transfer learning is a machine learning technique where shared knowledge is transferred between related tasks. Take for example, the source task of riding a bike, and the target task of riding a motorcycle. Although the tasks are different, their is shared knowledge from the source task, that will be useful when performing the target task. In layman's terms, if you already can ride a bike, it will be easier to ride a motorcycle. . 25, 29, 30

UMAP Uniform Manifold Approximation and Projection for Dimension Reduction. 26, 40, 42, 43

XAI Explainable AI. 3

Bibliography

[1] W. E. Forum, "A conversation with satya nadella, ceo of microsoft — davos 2023." https://www.youtube.com/watch?v=TSLcA66QgMY, 2023.

[2] Plant and F. Research, "A smart green future together plant & food research." https://www.plantandfood.com/en-nz/, 2023.

[3] "Callaghan innovation." https://www.callaghaninnovation.govt.nz/, Feb 2023.

[4] Plant and F. Research, "New research to maximise value from seafood resources - plant & food research." https://www.plantandfood.com/en-nz/article/new-research-to-maximise-value-from-seafood-resources, 2020.

[5] FAO, *The State of World Fisheries and Aquaculture, 2020*. FAO, 2020.

[6] M. Á. Pardo, E. Jiménez, and B. Pérez-Villarreal, "Misdescription incidents in seafood sector," *Food Control*, vol. 62, pp. 277–283, 2016.

[7] K. Lock and S. Leslie, "New zealand's quota management system: a history of the first 20 years," *Social Science Research Network (SSRN)*, 2007.

[8] "Hoki macruronus novazelandiae." https://openseas.org.nz/fish/hoki/, Oct 2021.

[9] "Fisheries and aquaculture in norway." https://www.oecd.org/agriculture/topics/fisheries-and-aquaculture/documents/report_cn_fish_nor.pdf, Jan 2021.

[10] StrictlyVC, "Strictlyvc in conversation with sam altman, part two (openai)." https://www.youtube.com/watch?v=ebjkD1Om4uw, 2023.

[11] A. Cooper, R. Reimann, and D. Cronin, *About face 3: the essentials of interaction design*. John Wiley & Sons, 2007.

[12] J. Wood, B. H. Nguyen, B. Xue, M. Zhang, and D. Killeen, "Automated fish classification using unprocessed fatty acid chromatographic data: A machine learning approach," in *Australasian Joint Conference on Artificial Intelligence*, pp. 516–529, Springer, 2022.

[13] D. P. Killeen, O. C. Watkins, C. E. Sansom, D. H. Andersen, K. C. Gordon, and N. B. Perry, "Fast sampling, analyses and chemometrics for plant breeding: bitter acids, xanthohumol and terpenes in lupulin glands of hops (humulus lupulus)," *Phytochemical Analysis*, vol. 28, no. 1, pp. 50–57, 2017.

[14] T. Miller, "Explanation in artificial intelligence: Insights from the social sciences," *Artificial intelligence*, vol. 267, pp. 1–38, 2019.

[15] T. Miller, P. Howe, and L. Sonenberg, "Explainable ai: Beware of inmates running the asylum or: How i learnt to stop worrying and love the social and behavioural sciences," *arXiv preprint arXiv:1712.00547*, 2017.

[16] T. Miller, "Contrastive explanation: A structural-model approach," *The Knowledge Engineering Review*, vol. 36, p. e14, 2021.

[17] A. C. Clarke, "Hazards of prophecy: The failure of imagination," *Profiles of the Future*, vol. 6, no. 36, p. 1, 1962.

[18] C. Black, O. P. Chevallier, S. A. Haughey, J. Balog, S. Stead, S. D. Pringle, M. V. Riina, F. Martucci, P. L. Acutis, M. Morris, *et al.*, "A real time metabolomic profiling approach to detecting fish fraud using rapid evaporative ionisation mass spectrometry," *Metabolomics*, vol. 13, no. 12, pp. 1–13, 2017.

[19] C. Black, O. P. Chevallier, K. M. Cooper, S. A. Haughey, J. Balog, Z. Takats, C. T. Elliott, and C. Cavin, "Rapid detection and specific identification of offals within minced beef samples utilising ambient mass spectrometry," *Scientific reports*, vol. 9, no. 1, pp. 1–9, 2019.

[20] S. N. Jha, *Rapid detection of food adulterants and contaminants: theory and practice*. Academic Press, 2015.

[21] J. Kaminski, "Diffusion of innovation theory," *Canadian Journal of Nursing Informatics*, vol. 6, no. 2, pp. 1–6, 2011.

[22] Y. Sun, B. Pfahringer, H. M. Gomes, and A. Bifet, "Soknl: A novel way of integrating k-nearest neighbours with adaptive random forest regression for data streams," *Data Mining and Knowledge Discovery*, vol. 36, no. 5, pp. 2006–2032, 2022.

[23] A. Bifet and R. Gavalda, "Learning from time-changing data with adaptive windowing," in *Proceedings of the 2007 SIAM international conference on data mining*, pp. 443–448, SIAM, 2007.

[24] M. A. Hussain, "Food contamination: major challenges of the future," 2016.

[25] H. Abdi and L. J. Williams, "Principal component analysis," *Wiley interdisciplinary reviews: computational statistics*, vol. 2, no. 4, pp. 433–459, 2010.

[26] J. Boccard and D. N. Rutledge, "A consensus orthogonal partial least squares discriminant analysis (opls-da) strategy for multiblock omics data fusion," *Analytica chimica acta*, vol. 769, pp. 30–39, 2013.

[27] F. Di Mattia, P. Galeone, M. De Simoni, and E. Ghelfi, "A survey on gans for anomaly detection," *arXiv preprint arXiv:1906.11632*, 2019.

[28] J. Bridle, "Training stochastic model recognition algorithms as networks can lead to maximum mutual information estimation of parameters," *Advances in neural information processing systems*, vol. 2, 1989.

[29] D. Held, S. Thrun, and S. Savarese, "Deep learning for single-view instance recognition," *arXiv preprint arXiv:1507.08286*, 2015.

[30] A. M. Hafiz and G. M. Bhat, "A survey on instance segmentation: state of the art," *International journal of multimedia information retrieval*, vol. 9, no. 3, pp. 171–189, 2020.

[31] M. Portaz, M. Kohl, J.-P. Chevallet, *et al.*, "Fully convolutional network and region proposal for instance identification with egocentric vision," in *Proceedings of the IEEE International Conference on Computer Vision Workshops*, pp. 2383–2391, 2017.

[32] J. Bromley, I. Guyon, Y. LeCun, E. Säckinger, and R. Shah, "Signature verification using a" siamese" time delay neural network," *Advances in neural information processing systems*, vol. 6, 1993.

[33] J. Zhu, J. Jang-Jaccard, A. Singh, I. Welch, A.-S. Harith, and S. Camtepe, "A few-shot meta-learning based siamese neural network using entropy features for ransomware classification," *Computers & Security*, vol. 117, p. 102691, 2022.

[34] L. Jing, J. Zhu, and Y. LeCun, "Masked siamese convnets," *arXiv preprint arXiv:2206.07700*, 2022.

[35] J. P. Crall, C. V. Stewart, T. Y. Berger-Wolf, D. I. Rubenstein, and S. R. Sundaresan, "Hotspotter—patterned species instance recognition," in *2013 IEEE workshop on applications of computer vision (WACV)*, pp. 230–237, IEEE, 2013.

[36] Z. Xie, A. Singh, J. Uang, K. S. Narayan, and P. Abbeel, "Multimodal blending for high-accuracy instance recognition," in *2013 IEEE/RSJ International Conference on Intelligent Robots and Systems*, pp. 2214–2221, IEEE, 2013.

[37] S. Agarwal, Y. Furukawa, N. Snavely, I. Simon, B. Curless, S. M. Seitz, and R. Szeliski, "Building rome in a day," *Communications of the ACM*, vol. 54, no. 10, pp. 105–112, 2011.

[38] B. S. Bucknall and S. Dori-Hacohen, "Current and near-term ai as a potential existential risk factor," in *Proceedings of the 2022 AAAI/ACM Conference on AI, Ethics, and Society*, pp. 119–129, 2022.

[39] ABC, "Openai ceo sam altman says ai will reshape society, acknowledges risks: 'a little bit scared of this'." `https://abcnews.go.com/Technology/openai-ceo-sam-altman-ai-reshape-society-acknowledges/story?id=97897122`, 2023.

[40] M. Tegmark, Y. Bengio, R. Russell, E. Musk, and S. Wozniak, "Pause giant ai experiments: An open letter." `https://futureoflife.org/open-letter/pause-giant-ai-experiments/`, 2023.

[41] A. Einstein, *The ultimate quotable Einstein*. Princeton University Press, 2011.

[42] I. Goodfellow, Y. Bengio, and A. Courville, *Deep Learning*. MIT Press, 2016. `http://www.deeplearningbook.org`.

[43] S. J. Russell, *Artificial intelligence a modern approach*. Pearson Education, Inc., 2010.

[44] C. Metz, "The godfather of a.i.' leaves google and warns of danger ahead," *The New York Times*, 2023.

[45] O. Gencoglu, M. van Gils, E. Guldogan, C. Morikawa, M. Süzen, M. Gruber, J. Leinonen, and H. Huttunen, "Hark side of deep learning–from grad student descent to automated machine learning," *arXiv preprint arXiv:1904.07633*, 2019.

[46] S. McCann and D. G. Lowe, "Local naive bayes nearest neighbor for image classification," in *2012 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 3650–3656, IEEE, 2012.

[47] R. Behmo, P. Marcombes, A. Dalalyan, and V. Prinet, "Towards optimal naive bayes nearest neighbor," in *Computer Vision–ECCV 2010: 11th European Conference on Computer Vision, Heraklion, Crete, Greece, September 5-11, 2010, Proceedings, Part IV 11*, pp. 171–184, Springer, 2010.

[48] T.-Y. Lin, P. Dollár, R. Girshick, K. He, B. Hariharan, and S. Belongie, "Feature pyramid networks for object detection," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 2117–2125, 2017.

[49] Y. LeCun, B. Boser, J. S. Denker, D. Henderson, R. E. Howard, W. Hubbard, and L. D. Jackel, "Backpropagation applied to handwritten zip code recognition," *Neural computation*, vol. 1, no. 4, pp. 541–551, 1989.

[50] Y. LeCun *et al.*, "Generalization and network design strategies," *Connectionism in perspective*, vol. 19, no. 143-155, p. 18, 1989.

[51] Y. LeCun, B. Boser, J. Denker, D. Henderson, R. Howard, W. Hubbard, and L. Jackel, "Handwritten digit recognition with a back-propagation network," *Advances in neural information processing systems*, vol. 2, 1989.

[52] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," *Proceedings of the IEEE*, vol. 86, no. 11, pp. 2278–2324, 1998.

[53] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," *Communications of the ACM*, vol. 60, no. 6, pp. 84–90, 2017.

[54] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014.

[55] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, "Going deeper with convolutions," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 1–9, 2015.

[56] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778, 2016.

[57] J. Balog, T. Szaniszlo, K.-C. Schaefer, J. Denes, A. Lopata, L. Godorhazy, D. Szalay, L. Balogh, L. Sasi-Szabo, M. Toth, *et al.*, "Identification of biological tissues by rapid evaporative ionization mass spectrometry," *Analytical chemistry*, vol. 82, no. 17, pp. 7343–7350, 2010.

[58] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," *arXiv preprint arXiv:1810.04805*, 2018.

[59] D. E. Goldberg, "Technical writing for fun & profit," 1999.

[60] I. Asimov, "The sun shines bright," *Garden City*, 1981.

[61] H. M. Gomes, J. Montiel, S. M. Mastelini, B. Pfahringer, and A. Bifet, "On ensemble techniques for data stream regression," in *2020 International Joint Conference on Neural Networks (IJCNN)*, pp. 1–8, IEEE, 2020.

[62] H. Mouss, D. Mouss, N. Mouss, and L. Sefouhi, "Test of page-hinckley, an approach for fault detection in an agro-alimentary production system," in *2004 5th Asian control conference (IEEE Cat. No. 04EX904)*, vol. 2, pp. 815–818, IEEE, 2004.

[63] D. Robinson, Q. Chen, B. Xue, D. Killeen, S. Fraser-Miller, K. C. Gordon, I. Oey, and M. Zhang, "Genetic algorithm for feature and latent variable selection for nutrient assessment in horticultural products," in *2021 IEEE Congress on Evolutionary Computation (CEC)*, pp. 272–279, IEEE, 2021.

[64] D. Robinson, Q. Chen, B. Xue, D. Killeen, K. C. Gordon, and M. Zhang, "A new genetic algorithm for automated spectral pre-processing in nutrient assessment," in *International Conference on the Applications of Evolutionary Computation (Part of EvoStar)*, pp. 283–298, Springer, Cham, 2022.

[65] R. Dawkins, "The evolved imagination: Animals as models of their world," *Richard Dawkins Foundation for Reason & Science*, 1995.

[66] C. Darwin and V. J. Wyhe, *On the origin of species: The science classic*. Capstone, 2020.

[67] R. Dawkins, "The selfish gene new york: Oxford university press," *DawkinsThe Selfish Gene1976*, 1976.

[68] D. Ha and J. Schmidhuber, "World models," *arXiv preprint arXiv:1803.10122*, 2018.

[69] Y. Matsuo, Y. LeCun, M. Sahani, D. Precup, D. Silver, M. Sugiyama, E. Uchibe, and J. Morimoto, "Deep learning, reinforcement learning, and world models," *Neural Networks*, 2022.

[70] J. F. Allen and J. A. Koomen, "Planning using a temporal world model," in *Proceedings of the Eighth international joint conference on Artificial intelligence-Volume 2*, pp. 741–747, 1983.

[71] J. Kennedy and R. Eberhart, "Particle swarm optimization," in *Proceedings of ICNN'95-international conference on neural networks*, vol. 4, pp. 1942–1948, IEEE, 1995.

[72] J. R. Koza *et al.*, *Genetic programming II*, vol. 17. MIT press Cambridge, 1994.

[73] B. Tran, B. Xue, and M. Zhang, "Genetic programming for feature construction and selection in classification on high-dimensional data," *Memetic Computing*, vol. 8, no. 1, pp. 3–15, 2016.

[74] B. Tran, B. Xue, and M. Zhang, "Genetic programming for multiple-feature construction on high-dimensional classification," *Pattern Recognition*, vol. 93, pp. 404–417, 2019.

[75] I. Kononenko *et al.*, "Estimating attributes: Analysis and extensions of relief," in *ECML*, vol. 94, pp. 171–182, 1994.

[76] L. Van der Maaten and G. Hinton, "Visualizing data using t-sne.," *Journal of machine learning research*, vol. 9, no. 11, 2008.

[77] L. McInnes, J. Healy, and J. Melville, "Umap: Uniform manifold approximation and projection for dimension reduction," *arXiv preprint arXiv:1802.03426*, 2018.

[78] S. Kullback and R. A. Leibler, "On information and sufficiency," *The annals of mathematical statistics*, vol. 22, no. 1, pp. 79–86, 1951.

[79] K. Bi, D. Zhang, T. Qiu, and Y. Huang, "Gc-ms fingerprints profiling using machine learning models for food flavor prediction," *Processes*, vol. 8, no. 1, p. 23, 2019.

[80] D. D. Matyushin and A. K. Buryak, "Gas chromatographic retention index prediction using multimodal machine learning," *Ieee Access*, vol. 8, pp. 223140–223155, 2020.

[81] J. Ho, A. Jain, and P. Abbeel, "Denoising diffusion probabilistic models," *Advances in Neural Information Processing Systems*, vol. 33, pp. 6840–6851, 2020.

[82] J. Song, C. Meng, and S. Ermon, "Denoising diffusion implicit models," *arXiv preprint arXiv:2010.02502*, 2020.

[83] P. G. Espejo, S. Ventura, and F. Herrera, "A survey on the application of genetic programming to classification," *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, vol. 40, no. 2, pp. 121–144, 2009.

[84] N. Zemmal, N. Azizi, N. Dey, and M. Sellami, "Adaptative s3vm semi supervised learning with features cooperation for breast cancer classification," *Journal of Medical Imaging and Health Informatics*, vol. 6, no. 4, pp. 957–967, 2016.

[85] C. Chen, O. Li, D. Tao, A. Barnett, C. Rudin, and J. K. Su, "This looks like that: deep learning for interpretable image recognition," *Advances in neural information processing systems*, vol. 32, 2019.

[86] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," *Advances in neural information processing systems*, vol. 30, 2017.

[87] A. Radford, K. Narasimhan, T. Salimans, I. Sutskever, *et al.*, "Improving language understanding by generative pre-training," *Arxiv*, 2018.

[88] A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, I. Sutskever, *et al.*, "Language models are unsupervised multitask learners," *OpenAI blog*, vol. 1, no. 8, p. 9, 2019.

[89] T. Brown, B. Mann, N. Ryder, M. Subbiah, J. D. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, *et al.*, "Language models are few-shot learners," *Advances in neural information processing systems*, vol. 33, pp. 1877–1901, 2020.

[90] L. Ouyang, J. Wu, X. Jiang, D. Almeida, C. L. Wainwright, P. Mishkin, C. Zhang, S. Agarwal, K. Slama, A. Ray, *et al.*, "Training language models to follow instructions with human feedback," *arXiv preprint arXiv:2203.02155*, 2022.

[91] Z. Zhao, R. Anand, and M. Wang, "Maximum relevance and minimum redundancy feature selection methods for a marketing machine learning platform," in *2019 IEEE international conference on data science and advanced analytics (DSAA)*, pp. 442–452, IEEE, 2019.

[92] H. Liu and R. Setiono, "Chi2: Feature selection and discretization of numeric attributes," in *Proceedings of 7th IEEE international conference on tools with artificial intelligence*, pp. 388–391, IEEE, 1995.

[93] E. Fix and J. L. Hodges, "Discriminatory analysis. nonparametric discrimination: Consistency properties," *International Statistical Review/Revue Internationale de Statistique*, vol. 57, no. 3, pp. 238–247, 1989.

[94] T. K. Ho, "Random decision forests," in *Proceedings of 3rd international conference on document analysis and recognition*, vol. 1, pp. 278–282, IEEE, 1995.

[95] W.-Y. Loh, "Classification and regression trees," *Wiley interdisciplinary reviews: data mining and knowledge discovery*, vol. 1, no. 1, pp. 14–23, 2011.

[96] D. J. Hand and K. Yu, "Idiot's bayes—not so stupid after all?," *International statistical review*, vol. 69, no. 3, pp. 385–398, 2001.

[97] C. Cortes and V. Vapnik, "Support-vector networks," *Machine learning*, vol. 20, no. 3, pp. 273–297, 1995.

[98] M. Robnik-Šikonja and I. Kononenko, "Theoretical and empirical analysis of relieff and rrelieff," *Machine learning*, vol. 53, no. 1, pp. 23–69, 2003.

[99] C. Ding and H. Peng, "Minimum redundancy feature selection from microarray gene expression data," *Journal of bioinformatics and computational biology*, vol. 3, no. 02, pp. 185–205, 2005.

[100] T. Chen, T. He, M. Benesty, V. Khotilovich, Y. Tang, H. Cho, K. Chen, *et al.*, "Xgboost: extreme gradient boosting," *R package version 0.4-2*, vol. 1, no. 4, pp. 1–4, 2015.

[101] R. A. Fisher, "The use of multiple measurements in taxonomic problems," *Annals of eugenics*, vol. 7, no. 2, pp. 179–188, 1936.

[102] R. González-Domínguez, T. García-Barrera, and J. Gómez-Ariza, "Using direct infusion mass spectrometry for serum metabolomics in alzheimer's disease," *Analytical and bioanalytical chemistry*, vol. 406, no. 28, pp. 7137–7148, 2014.

[103] C. Voss and T. Raz, *Never split the difference: Negotiating as if your life depended on it*. Random House, 2016.

[104] N. N. Taleb, *Fooled by randomness: The hidden role of chance in life and in the markets*, vol. 1. Random House Trade Paperbacks, 2005.

[105] N. N. Taleb, *The black swan: The impact of the highly improbable*, vol. 2. Random house, 2007.