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, motivations, limitations, research goals, and the organisation for the remainder of this proposal.

1.1 Problem Statement

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.

1.2 Motivations

This section introduces the motivations 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.

1.2.1 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.

1.2.2 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.

1.2.3 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.

1.2.4 Application-based AI

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 [10], 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.

1.2.5 Industry

Callaghan innovation hosted an industry workshop for the Cyber-Marine project [4]. The work in [11] was presented, as well the research several chemists like [12]. 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 [13], and move towards tools that can be understood by their users [14, 15]. 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.

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.

1.3 Limitations

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.

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

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) [16] 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:

1. Identification
    - Tasks:
        (a) Species
        (b) Part
    - Perform
        - Binary Classification
        - Multi-class Classification
2. Contamination
    - Tasks:
        (a) Cross-species
        (b) Mineral Oil
    - Perform:
        - Detection
        - Multi-label classification
        - Mutli-output

    * Multi-label classification - contaminants present

    * Regression - percentage of present contaminants

3. Individual Identification

    - Perform

        – Detection

        – Instance identification

These research goals are summarized in fig. 1.1.

This chart shows the three tasks, their subtasks respective and machine learning objectives, and desired output in the application domain.

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.

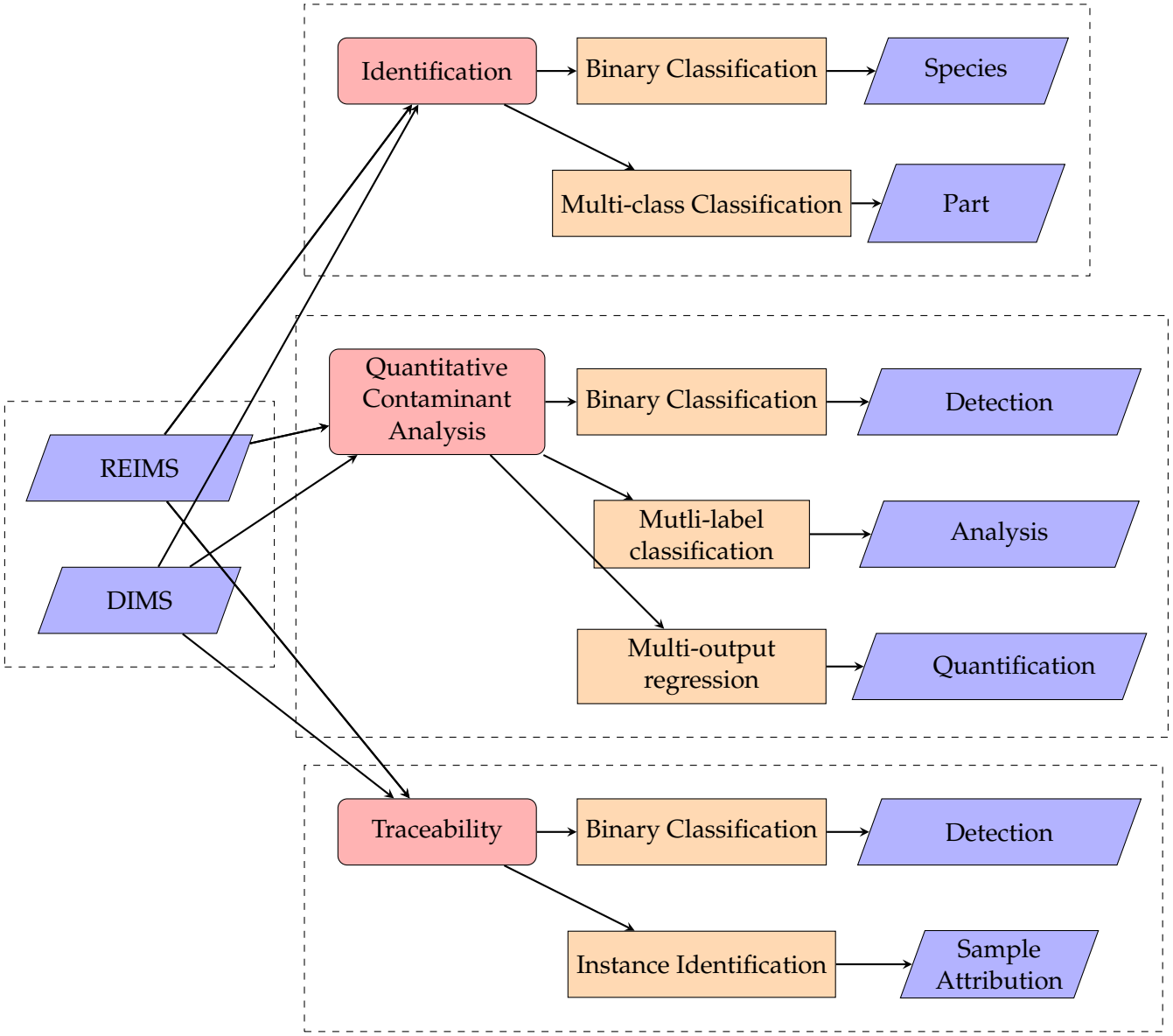


Figure 1.1: Research Goals

1.5 Organisation of the Proposal

The proposal is organised into four chapters. This introduction gives the scope of the problem and how this work intends to solve it. Second, A literature review which outlines existing work in the field and its limitations. The third chapter covers preliminary work on automated fish oil analysis. The final chapter gives the proposed contributions and project plan for this thesis. Specifically, this proposal has four chapters, introduction (this), literature review, preliminary work, contributions and project plan. Each chapter and section provides an brief description of its contents (like this one here) for clarity. Please see the table of contents for a more detailed breakdown of the contents of this proposal.

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.

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 [17], 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 individual identification, 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 [18, 19]. 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 [20, 21] 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)** [16]

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 [16] has shown promise in beef processing, where it was able to detect horse meat contamination in beef [22]. 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 [23, 24].

- **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) [25].

- **Gas-Chromatography Mass-Spectrometry** - Previous work [11] 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 [25], REIMS data modelled with PCA-LDA was able to detect species and catch method. Cross-species contamination is a more complex variation of this problem. In [25], 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. [22] performed detection and identification beef adulteration. It can identify samples that are adultered with offal, and specify which offal was present.

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 [26]
>
> *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 [27], 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 [28]. In later work from 1996, Dawkins proposed the evolved imagination, where every organism is a microcosm of it's environment [26]. 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." [26]

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 [29] and LeCun [30] have argued strong AI require an explicit world model [31].

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 [11], the EC technique of Particle Swarm Optimisation (PSO) [32] 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) [33] and Multi-Tree Genetic Programming (MT-GP) [34, 35] 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) [36]. 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 [37] 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 [25, 22] 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) [38]) were used.
  - PCA [38] 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) [39], Uniform Manifold Approximation and Projection for Dimension Reduction (UMAP) [40]
    - t-SNE [39]
      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 [41] between the two distributions.
- Basic supervised statistical models (e.g. LDA, OPLS-DA) was used for classification. Future work should consider CNNs [42, 43], GANs [44], Diffusion [45, 46]

- Denoising Diffusion Probabilistic Models (DDPM) [45], the original diffusion paper, behind diffusion-based image generation models.
- Denoising Diffusion Implicit Models (DDIM) [46], a generalized DDPM that is faster and deterministic.
- Genetic Programming for classification [47], feature construction [34, 35], 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. [48] 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 [22].
  - This looks like that (R-CNN) [49], 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 [50]
  - Attention is all you need! [51]
  - 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 [52].
  - Task-specific conditional learning [53], 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 [54], in-context learning, GPT-3 can be taught a new task in the input text passed to trained model.
  - Attention [51], and pre-training [50], allows for semantically meaningful features, trained on related tasks, that can the be applied to down-stream applications.
  - [53] and [54], 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) [55], 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 [56], 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 [57], 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 [36], 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 [44]
- Significant markers [25, 22]
- Outliers [25, 22]
- Relative Standard Deviation threshold [25, 22]
- Quality Control (QC) [25, 22]

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 [42, 43], GANs [44], Diffusion [45, 46]

- **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 [44], 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.

3.1 Automated Fish Classification on GC-MS data

The preliminary work starts by introducing previous research [11], 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, [11], 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, [11], that was addressed in the last subsec-

tion. 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) [35]. 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.

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 [11]. Balanced accuracy avoids

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
```

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 [11], for an example gas chromatogram and a more thorough description of the measurement technique.

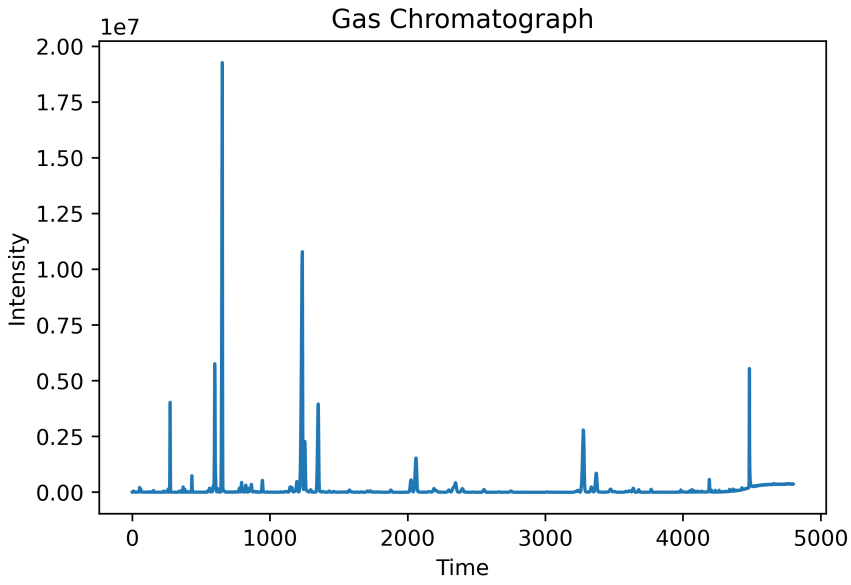


Figure 3.1: Gas Chromatogram of Fatty Acid Methyl Esters from Snapper Skin.

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 [11] 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 [11], 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 [58]	83.57	74.88
	RF [59]	100.0	85.65
	DT [60]	100.0	76.98
	NB [61]	79.54	75.27
	SVM [62]	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 [11]:
  - for fish species identification.

    * MCIFC exceeds performance of all FS methods, [57, 32, 63, 64], with SVM [32]

- for fish part identification.

    * MCIFC is better than χ^2 [57] and the full dataset.
    * MCIFC offers same performance as PSO [32]
    * MCIFC is worse than ReliefF [63] and MRMR [64]
    * MT-GP offers competitive performance to MRMR [64], 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.

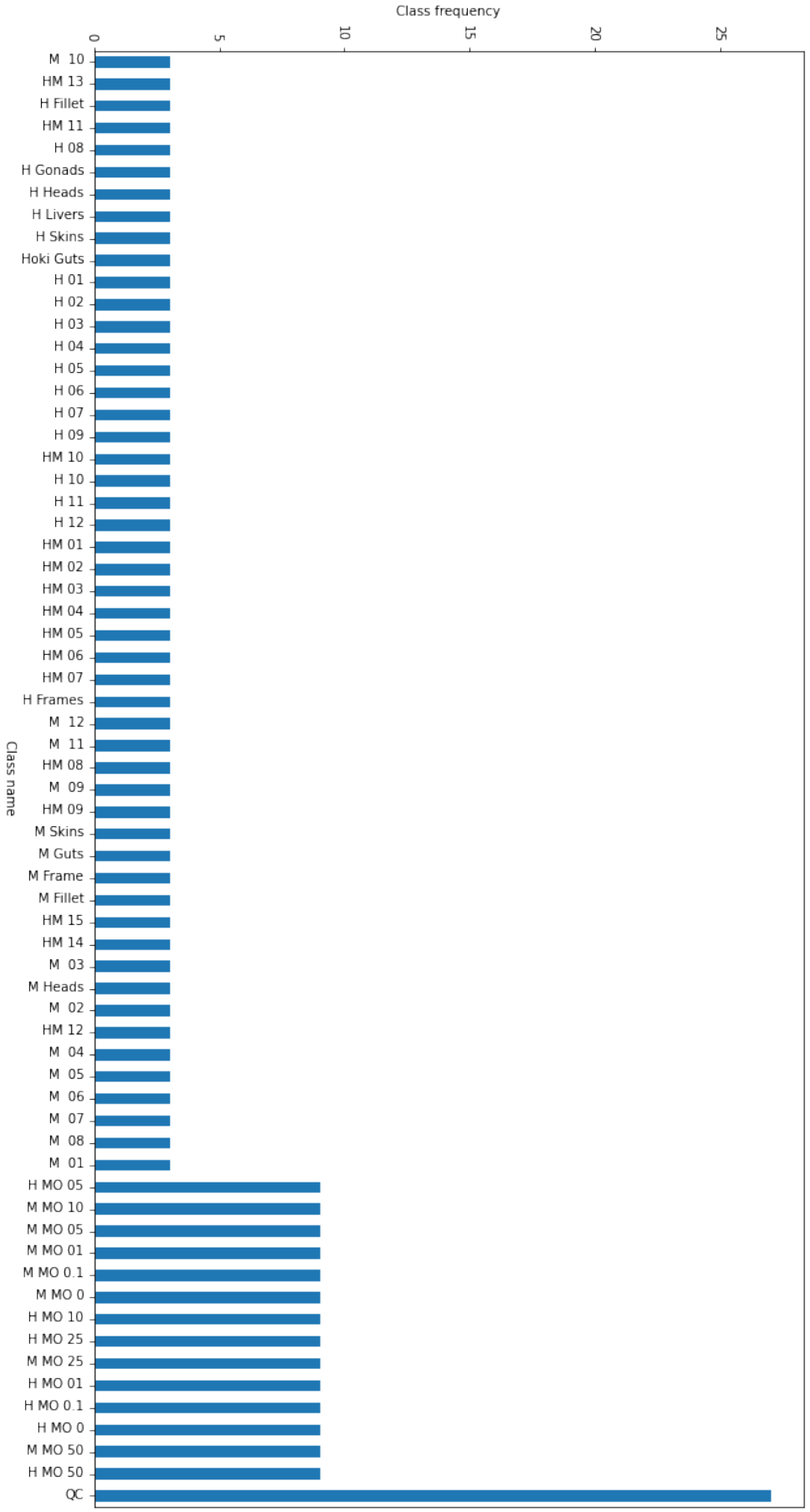


Figure 3.2: Class Distribution

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 3.4. 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.

Figure 3.3: Hoki *Macruronus novaezelandiae*

Figure 3.4: Mackerel *Trachurus symmetricus*

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 3.5 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.

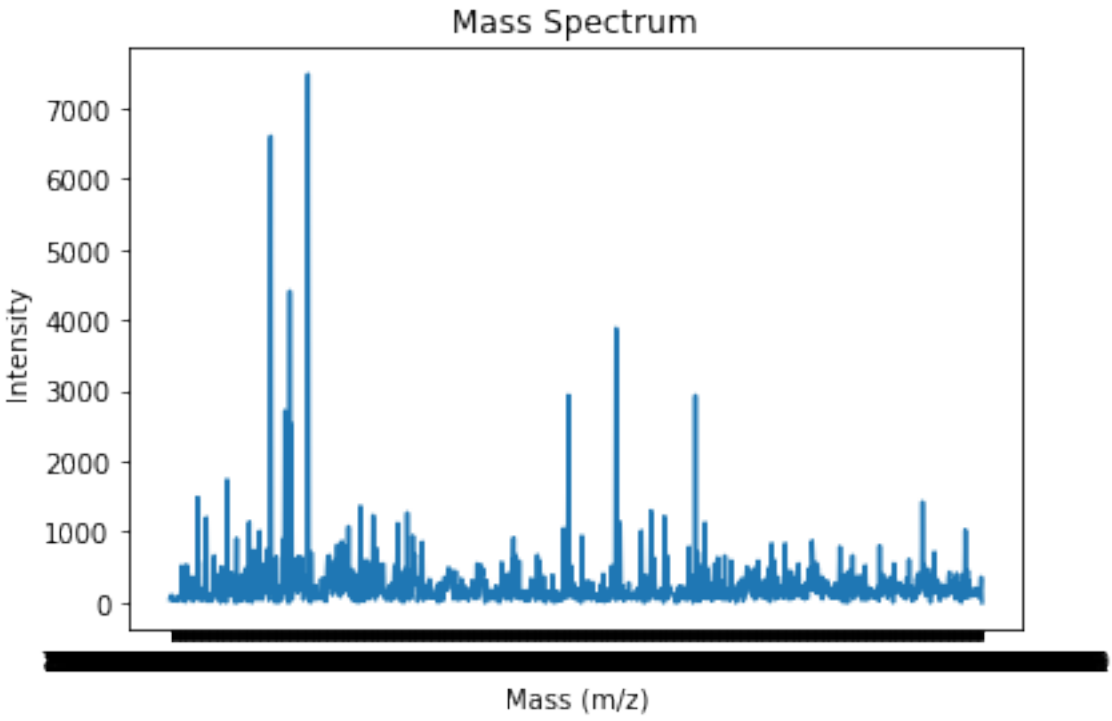


Figure 3.5: Mass Spectrum for a Hoki Fillet

Figure 3.6 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.

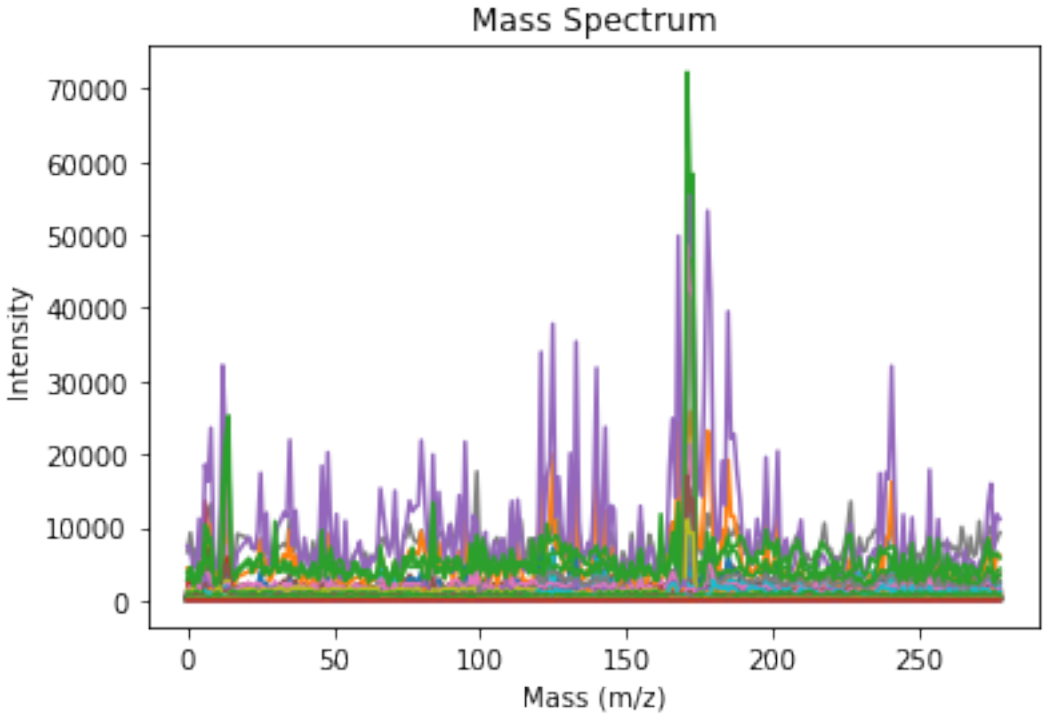


Figure 3.6: Mass Spectrums for Entire REIMS dataset

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 [62]	KNN [58]	DT [61]	RF [59]	XGBoost [65]	**LDA** [66]
PCA [38]	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 [39]	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 [40]	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 [38, 66] (**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, [25, 22] 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 [38] 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 [67] 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 [44], which is a closely related field to contamination detection and identification presented here.

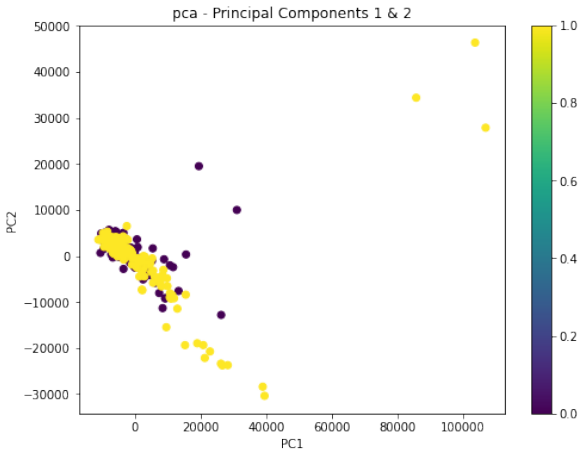
3.3.4 Ablation Studies

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.

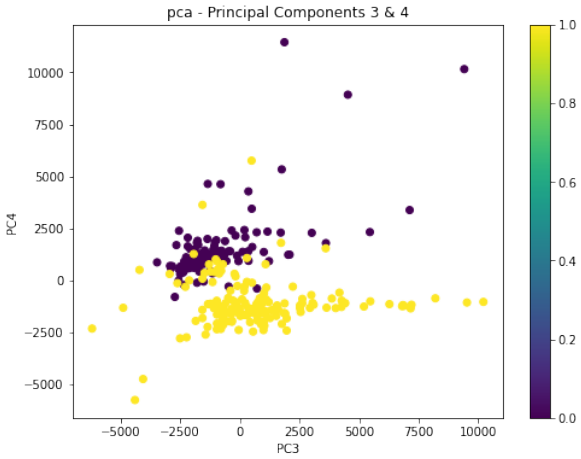
Table 3.5: Visual intuition for dimensionality reduction techniques and their respective feature subsets

PCA [38]

Features 1 & 2

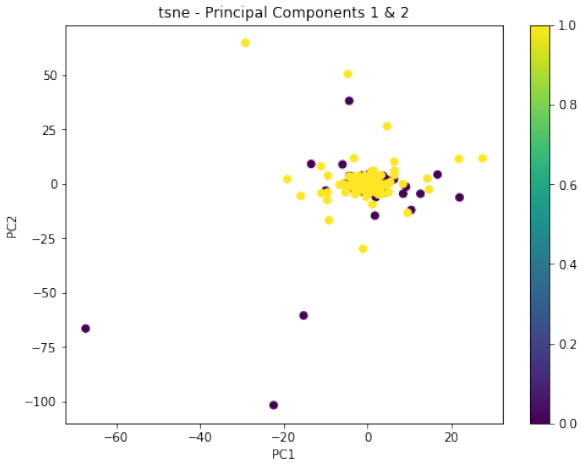


Features 3 & 4

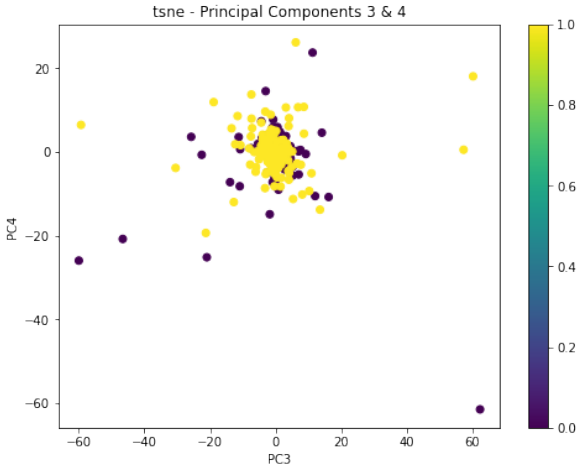


t-SNE [39]

Features 1 & 2

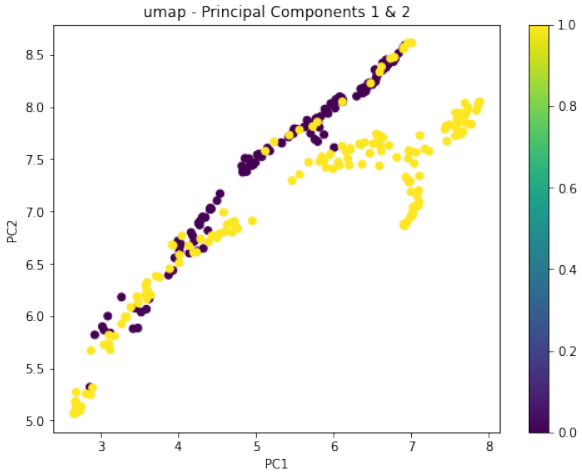


Features 3 & 4



UMAP [40]

Features 1 & 2



Features 3 & 4

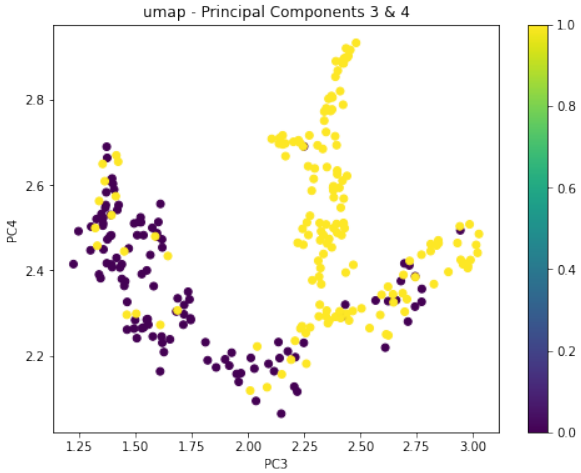


Table 3.5 gives the class distribution for features 1 & 2, versus features 3 & 4, for each dimensionality reduction method, PCA [38], t-SNE [39] and UMAP [40]. This gives intuitive and visual proof of the ability of each technique to tolerate noise in the dataset. The results, agree with the conjecture 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 [39] and UMAP [40], 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 [38]	55.47$\pm$ 6.68	**86.40 $\pm$ 16.25**
t-SNE [39]	57.24 $\pm$ 2.03	55.80 $\pm$ 3.69
UMAP [40]	85.27 $\pm$ 15.17	81.23 $\pm$ 17.15

Table 3.6 gives the cross-validation score for each dimensionality reduction method, PCA [38], t-SNE [39] and UMAP [40], 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 [38] 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 [39] and UMAP [40], 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 [38] 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.

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) [16]
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 [25] 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 [68]. 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 [25], REIMS results were worse on cooked biomass. Studies [25, 22] show that Mass-Spectrometry works on raw biomass products. A difference between the REIMS and GC dataset from [11], 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. Contamination
   (a) Cross-species
   (b) Mineral Oil
3. Individual identification

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 & tissue

Species identification [27] - can REIMS / DIMS data be used to classify different species tissues? What variables are responsible?

- Same task as [11], but instead of GC-MS, this is REIMS and DIMS
- Classification
- Feature Importance - Interpretable,
  - similar to significant markers from [25, 22]
  - and interpretability from [11, 49].

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 [11], 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 [25, 22]
  - and interpretability from [11, 49].

4.1.3 Contamination: Cross-species & Mineral Oil

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 [25], 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." [69]
- Popularized by Nassim Taleb, a risk analyst, in his books [70, 71].
- 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 [44] for anomoly detection using GANs, similar to [22] 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 Individual Identification

Individual identification - 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

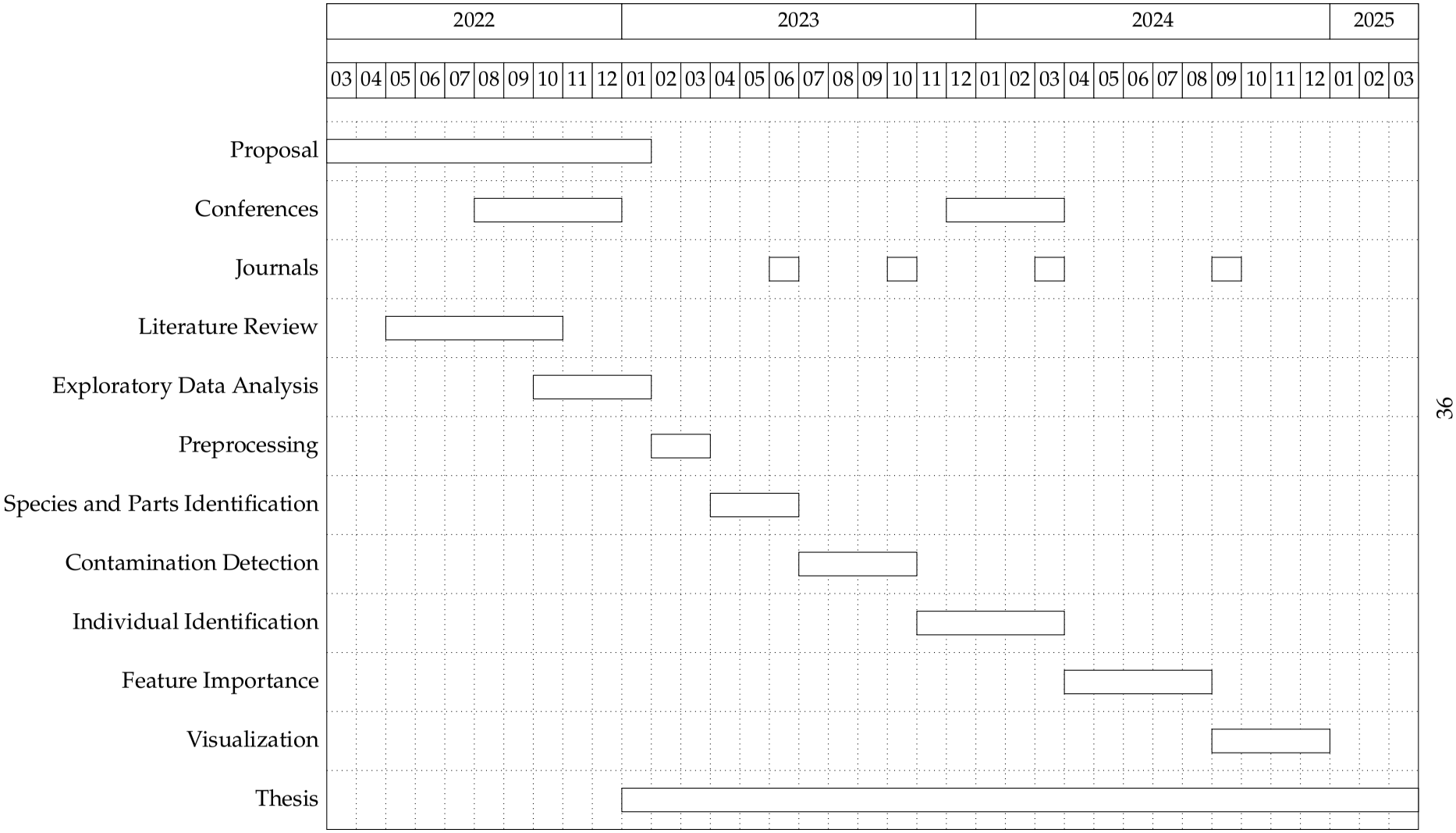
This research project has several key milestones it aims to achieve in the course of our work. In particular, the milestones for this proposal are:

1. Proposal
2. Conferences (x2)

3. Journals (x4)
4. Literature Review
5. Exploratory Data Analysis
6. Preprocessing
7. Species and Parts Identification
8. Contamination Detection
9. Contaminiaiton Indentification
10. Feature Importance
11. Visualization
12. 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
Species and Parts Identification
Contamination Detection
Individual Identification
Feature Importance
Visualization
Thesis

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) Species and Parts Identification
   (b) Cross-species Contamination
   (c) Mineral oil Contamination
   (d) Individual identification
5. Discussion
6. Conclusion
7. 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, [25, 22] 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 [11]. 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 [11], 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 [72] . 2, 8, 9

AI Artificial Intelligence. 3

AIML Artificial Intelligence Machine Learning. 7, 15, 31, 34

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 [44]. 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. . 13, 15

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 . 25

CNN Convolutional Neural Networks. 11, 14, 17

conceptual drift A term from data stream mining, [18, 19], 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 . 7, 8, 13, 15

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 [73] . 2–4, 8, 28, 32, 33, 36, 37

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. . 21, 27, 30

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, 4

DDIM Denoising Diffusion Implicit Models. 12

DDPM Denoising Diffusion Probabilistic Models. 12

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 . 8, 14, 28, 33, 37

DIMS Direct Infusion Mass Spectrometry. 5, 8, 31–34, 37, 39

domain knowledge Knowledge related to the application domain. For example, bio-chemistry and fish processing. . 14

EC Evolutionary Computation. 9, 10

EDA Exploratory Data Analysis. 22, 35–37

FC Feature Construction. 20

FS Feature Selection. 21, 22

GAN Generative Adversarial Networks. 28

gas chromatogram Gas Chromatography for fatty acid analysis in [11]. 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 . 19

GC Gas-Chromatography. 32

GC-MS Gas-Chromatography Mass-Spectrometry. i, 8, 17, 21, 32, 39

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. . 10

GP Genetic Programming. 8, 17, 18, 20, 21

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. . 10, 11, 14

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 . 8, 14, 17, 24, 27, 28, 32–37

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 . 25

KL Kullback-Leibler. 11

KL K-Nearest Neighbours. 10

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, 7, 8, 31–33, 35, 37

mass The amount of matter in an object . 25

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. . 25

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 . 22, 26

MCIFC Multiple Class-independent Feature Construction Method. 18, 20–22

ML Machine Learning. 14, 15, 37

MO Mineral Oil. 24

MS Mass-Spectrometry. 8, 11, 12, 15, 17, 24, 25, 31, 32, 37, 39

MT-GP Multi-Tree Genetic Programming. 10, 18, 20–22

offline see **online** . 31, 34

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** . 4, 31

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** . 10, 15

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. . 8, 17, 24, 32, 35, 37

PCA Principal Component Analysis. 11, 27, 30

PCA-LDA Principal Component Analysis - Linear Discriminant Analysis. 9, 27

PFR Plant and Food Research New Zealand Ltd.. 27, 28

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. . 10

PSO Particle Swarm Optimisation. 10, 17, 22

QA Quality Assurance. 34

QC Quality Control. 2, 3, 8, 12, 14, 24

REIMS Rapid Evaporative Ionisation Mass Spectrometry. 4, 5, 8, 9, 11, 12, 17, 22, 24–27, 30–34, 37, 39

RSD Relative Standard Deviation. 11, 14, 24

sample complexity A term from deep learning, [67], which refers to volume of data. In general, complex methods such as deep learning, require large volumes of data for a model to be trained propoerly. In relation to fish processing, the sample complexity is low, due to the time-consuming and manually intentsive process of collecting said samples. However, as these methods are deployed in real-time, the sample compelixty will increase. . 13

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]. . 7, 15

significant markers Significant Markers (or important variables) are ions that are unique to a specific offal cut, and present in all samples [22] . 12, 14, 32–35

SOTA state-of-the-art. 14

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**. . 7

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 [11] were Bluecod, Gurnard, Snapper & Tarakihi. For differentiating between distinct species in fish fraud detection see [25]. Darwin [27] . 8, 17, 24, 27, 32, 35, 37

spoilage TODO . 2

ST-GP Single-Tree Genetic Programming. 10, 18, 20

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 . 21

SVM Support Vector Machine. 18, 22, 30

t-SNE T-distributed stochastic neighbor embedding. 11, 27, 29, 30

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 . 4, 10, 14, 15, 37

tissue See part . ii, 19, 24, 31, 32, 39

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. . 10, 14, 15

UMAP Uniform Manifold Approximation and Projection for Dimension Reduction. 11, 27, 29, 30

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

[11] 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.

[12] 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.

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

[14] 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.

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

[16] 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.

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

[18] 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.

[19] 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.

[20] 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.

[21] 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.

[22] 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.

[23] 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.

[24] 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.

[25] 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.

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

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

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

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

[30] 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.

[31] 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.

[32] 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.

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

[34] 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.

[35] 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.

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

[37] 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.

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

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

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

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

[42] 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.

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

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

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

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

[47] 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.

[48] 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.

[49] 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.

[50] 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.

[51] 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.

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

[53] 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.

[54] 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.

[55] 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.

[56] 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.

[57] 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.

[58] 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.

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

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

[61] 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.

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

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

[64] 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.

[65] 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.

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

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

[68] 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.

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

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

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

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

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